Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Gestionale
Tesi di Laurea Magistrale

Strategie di campionamento per la valutazione della "Data Quality" di Data Base relazionali: un'applicazione nel settore della Bibliometria

Relatori:
Prof. Fiorenzo Franceschini
Prof. Domenico Augusto Francesco Maisano

Candidato:
Silvia Damiani

Anno Accademico 2021/2022

INDICE

LISTA ACRONIMI

- ANVUR: Agenzia nazionale di valutazione del sistema universitario e della ricerca
- API: Application Programming Interface
- AQL: Acceptance Quality Level
- CA: Controllo di accettazione
- DB: Database
- DBMS: Database Management System
- DLC: Data life cycle
- DQR: Data Quality Requirements
- DOI: Digital Object Identifier
- IA: Intelligenza Artificiale
- ISSN: International Standard Serial Number
- LTPD: Lot Tolerance Percentage Defectives
- ML: Machine Learning
- QM: Quality Measures
- RPA: Robotic Process Automation
- WoS: Web of Science

INDICE DELLE TABELLE

INDICE DELLE FIGURE

PREMESSA

Oggigiorno tutte le organizzazioni, a vario titolo, ruotano intorno ai "dati". La centralità dei dati è ormai comunemente riconosciuta: tutti sono consapevoli, in contesti aziendali e non, di come non si possa prescindere dal loro utilizzo per portare avanti una qualsiasi attività.

I dati sono strumento tramite il quale conoscere i consumatori, analizzare trend, fare ragionamenti, prendere decisioni, elaborare strategie, trarre conclusioni e così via; proprio in virtù della vasta gamma di applicazioni che gli stessi possono avere, non è più possibile trattare in maniera sommaria il concetto di qualità dei dati, ma è imperativo definire metodologie che consentano tanto di progettarla, mantenerla e, laddove possibile, incrementarla, nell'ottica del fornitore, quanto di valutarla, nell'ottica del cliente.

La qualità è d'altronde un aspetto che ha assunto in anni recenti maggiore rilevanza nelle organizzazioni che dovrebbero prestare una costante attenzione a come questo aspetto possa essere declinato ed applicato tanto ai processi quanto ai prodotti/servizi che vengono presi in input e ottenuti in output.

In quest'ottica, i dati, e i database che li contengono, rappresentano il prodotto di interesse e tutte le molteplici attività che su di loro e con loro vengono eseguite i processi da considerare.

Assumendo dunque il punto di vista di un ipotetico acquirente di un database, l'obiettivo sarà quello di definire delle prove che consentano di testare la qualità dei dati, acquisiti gratuitamente o a pagamento, dal momento che da ciò dipenderà non solo il suo grado di soddisfazione, ma anche la qualità degli output che produrrà sulla base delle "informazioni in input".

Più nello specifico, si andrà ad applicare la metodologia al caso di studio dei database bibliometrici con l'obiettivo di prendere decisioni, oggettive e giustificate, circa la loro "accettazione" o il loro "rifiuto". A tal proposito, si farà ricorso all'implementazione di metodi di controllo statistico, quali i piani di campionamento.

Obiettivo che questa tesi si pone è allora in primis quello di aiutare gli acquirenti dei dati a definire procedure rigorose ed oggettive tramite le quali valutarne la qualità, per poi avere delle garanzie circa il fatto che tutte le elaborazioni ed analisi che saranno condotte siano fondate su dati affetti da bassa difettosità.

Questo è ciò che potrebbe desiderare l'ANVUR (Agenzia nazionale di valutazione del sistema universitario e della ricerca) in fase di selezione ed acquisto di un database bibliometrico, da utilizzare, per esempio, nelle valutazioni delle attività di ricerca in Italia. In aggiunta, si vuole

anche brevemente mostrare come strumenti quantitativi, quali i piani di campionamento, solitamente applicati a contesti produttivi classici, possano essere utilizzati e sfruttati anche in relazione a software e prodotti informatici in generale.

Il lavoro è organizzato in quattro capitoli:

- **Capitolo I:** si sono introdotti i pilastri portanti sui quali si baserà la trattazione, ovvero la qualità e i database. Dunque, si è declinato il concetto di qualità in relazione al prodotto considerato e ai dati in essi contenuti, richiamando lo standard SQuaRE e le dimensioni che la normativa introduce per valutare la data quality.
- **Capitolo II:** tale capitolo vuole definire la metodologia che verrà utilizzata per effettuare la valutazione della qualità dei dati. A tal proposito, si sono richiamati gli aspetti teorici relativi a controlli di accettazione e piani di campionamento e la letteratura disponibile in materia di data quality assessment.
  Dunque, si è presentata in termini generali la metodologia di valutazione proposta, sottolineandone punti di forza e punti di debolezza.
- **Capitolo III:** si è presentato il caso di studio trattato, introducendo i due principali database bibliometrici oggi disponibili. Dunque, si è applicata la nuova metodologia in tutte le sue fasi, corredando la descrizione con esemplificazioni ad hoc. Infine, si sono illustrati i risultati ottenuti nei diversi scenari valutati e derivanti dalla proposta di diversi schemi di campionamento.
- **Capitolo IV:** si espongono le conclusioni e gli aspetti distintivi del seguente lavoro di tesi, per poi proporre eventuali migliorie o modifiche della metodologia. Queste sono finalizzate ad ampliare il campo di applicazione dell'algoritmo ed eventualmente incrementarne le prestazioni nel garantire i desiderati livelli qualitativi.

1. CAPITOLO I
CONCETTI INTRODUTTIVI

1.1 Il concetto di qualità

La *ISO 8402:1994* definiva la qualità come "l'insieme delle proprietà e delle caratteristiche di un prodotto o di un servizio che conferiscono ad esso la capacità di soddisfare esigenze espresse o implicite".
Questa definizione è stata poi rivista dalla *UNI EN ISO 9000:2015* che, invece, afferma che la qualità è "il grado con cui un insieme di caratteristiche intrinseche di un oggetto soddisfano i requisiti".
È doveroso entrare nel dettaglio della definizione data dalla norma e cercare di puntualizzare meglio i concetti in essa esplicitati:

- *Grado*: ogni qual volta si valuta la qualità di un bene, occorre effettuare misurazioni quantitative e dunque conciliare e far convivere gli aspetti di qualità e quantità che in altri contesti potevano sembrare in contrasto.
- *Insieme di caratteristiche intrinseche*: la qualità è un insieme di proprietà e caratteristiche, come indicato nella ISO 8402, e non di singoli elementi. Viene allora chiaramente espressa la multidimensionalità di tale concetto.
  Si noti poi che le caratteristiche dell'insieme dovranno essere proprie e distintive dell'oggetto che si sta analizzando.
- *Oggetto*: rappresenta ciò di cui si valuta la qualità e può essere un prodotto, un servizio, un processo, una persona, una organizzazione, un sistema o una risorsa.
- *Requisito*: tale termine sostituisce il concetto di esigenza, espressa o implicita, menzionato nella ISO 8402 e, più nel dettaglio, amplia quanto detto dalla precedente norma. Quando si fa riferimento ai requisiti, infatti, si sta parlando sia di esigenze che di aspettative, che a loro volta potranno essere espresse, implicite oppure obbligatorie.
  Non viene in ogni caso menzionato chi debba essere a definire gli stessi requisiti: a seconda dell'oggetto considerato, gli stakeholders saranno diversi e costituiranno una rete di portatori di interessi, distinti e spesso contrastanti, dalla quale non si potrà prescindere per valutare la qualità.

Non a caso si è deciso di inserire la qualità come primo concetto introduttivo: come già specificato nella premessa, obiettivo della seguente trattazione è quello di definire ed implementare un metodo di verifica della qualità dei database.
Per raggiungere tale scopo, si dovranno prendere accordi sul livello di servizio desiderato, definendo, in virtù dei requisiti espressi da cliente e fornitore, parametri di qualità specifici che il bene in esame dovrà soddisfare prima di essere consegnato.
Il risultato di tale negoziazione sarà la stipula di un contratto di fornitura ad hoc, in virtù del quale si procederà con la messa in esercizio di un piano di campionamento che consentirà non solo di valutare la difettosità del database in questione, ma soprattutto di far prendere al committente una decisione circa l'accettazione o meno della base dati di interesse.

Avere delle garanzie sul fatto che i dati e le informazioni che si stanno acquistando sono di qualità è, ancora di più oggi, fondamentale per una serie motivi e per entrambi i soggetti coinvolti nella transazione:

Lato cliente

1. C'è una sempre maggiore attenzione della committenza verso prodotti e servizi di qualità. In fase di acquisto, infatti, si valuta la loro *buona o cattiva qualità*.
   Si noti però che, per esprimere un giudizio circa la qualità di un prodotto e/o servizio, non si può prescindere dal considerare l'intero ciclo di vita che ha portato alla sua realizzazione e/o erogazione.
2. Ormai diffusa risulta essere la pratica delle certificazioni con cui una parte terza certifica il soddisfacimento dei requisiti da parte di un prodotto/servizio: la committenza avrà in tal caso una garanzia imparziale ed oggettiva circa la qualità del bene in esame. A definire un quadro per la valutazione della qualità dei prodotti software, risulta essere la famiglia di standard di riferimento ISO 25000, denominata SQuaRE "System and software Quality Requirements and Evolution".
   Volendo andare ancora più nel dettaglio, la qualità dei dati contenuti nei database viene definita dagli standard internazionali:
   - ISO/IEC 25012 "Data quality model"
   - ISO/IEC 25024 "Measurement of data quality"

   Si noti che, qualora le organizzazioni scelgano di provvedere autonomamente alla verifica del soddisfacimento dei requisiti, si parlerà di fatto di autocertificazioni,

caratterizzate da una minore affidabilità e fiducia nei risultati ottenuti da un lato e da una maggiore dipendenza dalle caratteristiche idiosincratiche della realtà che sta eseguendo le diverse prove dall'altro.

3. I dati grezzi che vengono raccolti nei database vengono elaborati delle organizzazioni per formulare strategie, prendere decisioni e trarre conclusioni.
   Se a monte ci sono dei problemi di qualità, inevitabilmente questi si ripercuoteranno su tutte le elaborazioni successive.

Lato fornitore

1. L'esecuzione di test volti alla verifica della qualità dei prodotti consente di individuarne i difetti e di rimuoverli prima della consegna, con conseguenti risparmi in termini di successive attività manutentive.
   Qualora le verifiche vengano eseguite dal cliente a prodotto consegnato, la rilevazione di eventuali errori e problematiche è comunque importante in una logica di miglioramento continuo: una volta comunicati al fornitore, sarà suo dovere tenerne conto per realizzare basi di dati sempre più performanti, rispondenti alle esigenze del mercato e di qualità.
   Tale punto sottolinea come dietro al concetto di qualità si celi quello di variabilità nel tempo: l'abilità delle aziende oggi andrà allora necessariamente valutata considerando anche la loro capacità di star dietro alla qualità e ai suoi avanzamenti.
2. L'aver ottenuto una certificazione da un ente esterno consente al fornitore di differenziarsi dalla concorrenza, di ridurre i difetti presenti nel prodotto e dare garanzie in termini di sicurezza e di compatibilità con altri prodotti.
3. Il concetto di qualità è oggi strettamente legato a quello di customer satisfaction, definita dalla ISO 9001 come "la percezione, da parte del cliente, del grado con cui le sue aspettative sono soddisfatte dal prodotto o servizio".
   La qualità diventa infatti deducibile dal grado di soddisfazione espresso da chi utilizza il prodotto o servizio in esame.
   Se il grado di soddisfazione è basso, è perché la qualità è bassa e questo implicherà conseguentemente un basso customer retention rate e ridotte performance economiche per i fornitori.

1.2 Il ruolo della qualità

Adottando un punto di vista più operativo e ricordando la centralità della qualità lungo tutto il ciclo di vita che porta alla realizzazione di un prodotto e/o erogazione di un servizio, occorre specificare come il concetto appena introdotto possa essere applicato sia a livello di progettazione che di produzione.
Se nel secondo caso la qualità implica il controllo della produzione, quando si parla di progettazione di prodotti/servizi, si farà riferimento alla progettazione della qualità con opportuni strumenti.

A seconda dell'uno o dell'altro campo, le metodologie utilizzabili a supporto sono diverse e riassunte nella **Tabella *1.1***.

Tabella 1.1: *Strumenti per la qualità*

STRUMENTI PER LA QUALITA' NELLA PROGETTAZIONE	STRUMENTI PER LA QUALITA' NELLA PRODUZIONE
QFD	Metodi di controllo statistico
DFx	Controllo a tappeto
CAx	
Metodi affidabilistici	

Brevemente, a supporto della progettazione, possiamo considerare:

- **QFD (*Quality Function Deployment*):** strumento che consente di introdurre, nella progettazione di un nuovo prodotto e/o servizio, le esigenze del cliente e di definire caratteristiche tecniche che siano ad esse rispondenti. Nella metodologia introdotta da tale trattazione, verrà utilizzato come strumento di supporto per la definizione delle caratteristiche dei dati da valutare, poiché consentirà di selezionare quelle ritenute rilevanti sulla base dei requisiti espressi dalla committenza.
- **DFx (*Design for X* o *for Excellence*):** insieme di linee guida per la progettazione di prodotti o servizi. X è una variabile che assume valori solitamente legati agli attributi, alle fasi del ciclo di vita, altri aspetti del prodotto o agli obiettivi della stessa progettazione.
- **CAx (*Computer Aided X*):** noto anche come Computer Aided Technologies, si tratta di un approccio che prevede l'utilizzo di software a supporto dell'intero ciclo di vita di un prodotto o servizio, dalla sua ideazione fino alla produzione.

- **Metodi affidabilistici:** si fa qui riferimento alla branca della Reliability Engineering (Ingegneria dell'affidabilità). Tali metodi vanno ad impostare la progettazione per massimizzare l'affidabilità, intesa come probabilità che un sistema compia correttamente la funzione prevista nelle condizioni ambientali e per il periodo tempo per cui è stato specificatamente progettato.

Per quanto riguarda invece la produzione, in relazione ai metodi di controllo statistico, si prevedono tre diverse aree di intervento:

1. **SPC (*Statistical process control*):** anche noto come controllo statistico di processo, implica un controllo della produzione in itinere, man mano che il processo stesso evolve.
   Classico strumento utilizzato in questo campo sono le carte di controllo.
2. **DOE *(Design of Experiment):*** è più un metodo di progettazione e gestione efficace degli esperimenti che un metodo di controllo della produzione.
   Consente di definire le relazioni esistenti tra una serie di variabili in input e una o più variabili in uscita, al fine di migliorare le prestazioni del processo produttivo in esame.
   Ad esempio, si faccia riferimento alle decisioni di progettazione di un nuovo prodotto, miglioramento di prodotti esistenti o ottimizzazione del processo di produzione attuale.
3. **Controlli di accettazione:** si tratta di tutto quell'insieme di controlli che devono essere eseguiti qualora, come nel nostro caso, sia previsto un passaggio di consegne da un fornitore ad un cliente.
   L'obiettivo sarà quello di verificare che quanto fornito soddisfi i requisiti definiti preliminarmente dalle parti coinvolte nello scambio.
   Tipici strumenti tramite i quali si effettuano i controlli di accettazione sono i piani di campionamento.

Sempre in ambito produttivo, è possibile procedere con un controllo a tappeto dove, come deducibile dal nome dello strumento, si andrà a monitorare l'intera produzione.

Comprendere come il concetto di qualità venga concretamente declinato nelle organizzazioni è funzionale alla trattazione dal momento che uno degli obiettivi che tale tesi si pone è quello di mostrare come i tool sopra descritti, storicamente applicati a contesti produttivi e manifatturieri classici, possano essere estesi anche al software e al mondo informatico.

Dunque, nei seguenti capitoli, si procederà all'implementazione di tutta una serie di controlli volti a decretare l'accettazione o meno dei dati e delle informazioni fornite, sulla base di un bundle di caratteristiche definite in seguito all'applicazione del QFD.

1.3 Il concetto di database

Dal momento che la qualità dipende strettamente dell'oggetto al quale si riferisce, si ritiene ora necessario procedere con una breve digressione sul concetto di database, al fine di metterne in evidenze le caratteristiche distintive, sia da un punto di vista funzionale che tecnico/informatico. Ciò consentirà di definire più precisamente i requisiti e le specifiche del prodotto in esame, sia in un'ottica di sua eventuale progettazione che di controllo.

Nella società contemporanea, tutto è dato: dati sono quelli che produciamo quotidianamente con le attività che eseguiamo dai nostri smartphone o da un qualsiasi sistema connesso; dati sono quelli che vengono utilizzati per fornirci servizi sempre più rispondenti alle nostre esigenze e ovviamente dati sono quelli su cui le organizzazioni si basano per prendere decisioni ed elaborare strategie.
Quelli indicati sono solo alcuni dei molteplici esempi di dato che ci si può trovare a dover maneggiare, in realtà aziendali e non: la gestione del dato è allora un processo lungo e articolato che può prevedere diversi step prima di consentire il passaggio dal dato grezzo all'informazione compiuta, che altro non è che un dato contestualizzato e messo a confronto con altri.
In questo scenario, assume un ruolo fondamentale la raccolta, l'archiviazione e l'elaborazione dei dati, con modalità qualitativamente soddisfacenti, tramite opportuni sistemi informativi per basi di dati.
Un database (o base di dati) è uno strumento che consente di trattare grandi quantità di dati ed informazioni in maniera ordinata e funzionale agli obiettivi che vengono perseguiti da organizzazioni di ogni tipo e dimensione.

Volendo dare una definizione più tecnica, un DB viene definito come "una raccolta di dati permanenti, gestiti da un elaboratore elettronico e suddivisibili in due categorie":

1. Metadati: rappresentano lo schema della base di dati, definito ex ante ed indipendentemente dalle applicazioni che verranno previste. Implica la definizione di:
    - Struttura dei dati
    - Relazioni esistenti tra gli insiemi di dati
    - Operazioni eseguibili sui dati
    - Restrizioni sui valori ammissibili (vincoli d'integrità)
2. Dati veri e propri, caratterizzati da:

- Organizzazione in insiemi omogenei, fra i quali sono definite relazioni;
- Maggiore numerosità, in assoluto e rispetto ai metadati;
- Stabilità: i dati sono permanenti e, una volta creati, continuano ad esistere finché non sono esplicitamente rimossi. La loro vita, quindi, non è limitata dalla durata delle applicazioni che ne fanno uso;
- Accessibilità mediante transazioni, sequenze di azioni di lettura e scrittura della base di dati e di elaborazioni di quest'ultimi;
- Protezione dall'accesso di utenti non autorizzati;
- Affidabilità, in presenza di corruzioni dovute a malfunzionamenti hardware e software;
- Condivisibilità tra applicazioni ed utenti. (Albano et al., 2019)

Per comprendere in maniera più agevole i concetti sopra esposti, si considerino i database bibliometrici, caso di studio della suddetta trattazione: in breve, tali database hanno l'obiettivo di memorizzare e presentare informazioni relative alle pubblicazioni, ai ricercatori, ma anche alle riviste sulle quali i diversi articoli vengono pubblicati al fine di consentire agli utenti l'esecuzione di ricerche di tipo bibliometrico ed analisi citazionali.
Semplificando, ognuna delle classi indicate (i.e. pubblicazioni, ricercatori, riviste) può essere vista, in un'ottica di database relazionale, come una tabella con tante colonne quanti saranno gli attributi di interesse.
Si considerino le seguenti tabelle esemplificative (**Tabella 1.2** e **Tabella 1.3**)

Tabella 1.2: *Esempio di database relativo alle Pubblicazioni*

Pubblicazioni					
Authors	*Article Title*	*DOI*	***Source title***	*Pages*	*...*
A1	AT1	D1	ST1	P1	…
A2	AT2	D2	ST2	P2	…
A3	AT3	D3	ST3	P3	…
…	…	…	…	…	…

Tabella 1.3: *Esempio di database relativo alle Riviste*

Riviste					
ID	*Source Title*	*ISSN*	*Category*	*Citations*	*...*
ID1	ST1	ISSN1	CA1	CI1	…
ID2	ST2	ISSN2	CA2	CI2	…
ID3	ST3	ISSN3	CA3	CI3	…
…	…	…	…	…	…

I metadati riguardano le seguenti informazioni:

1. il fatto che esistono diverse collezioni di interesse, tra le quali Pubblicazioni e Riviste
2. la struttura degli elementi di queste collezioni, in termini di intestazione delle tabelle: ad esempio, ogni pubblicazione ha degli autori, di tipo stringa, un titolo, sempre di tipo stringa, un certo numero di pagine, di tipo intero e così via;
3. il fatto che ad ogni rivista corrispondano una, nessuna o più pubblicazioni e che ad ogni pubblicazione corrisponda una sola rivista;
4. alcuni vincoli sui valori ammissibili, quali ad esempio che il numero di pagine sia un valore maggiore o uguale a 1 oppure che tutte le stringe relative al DOI siano tra loro diverse, etc.

I dati invece sono le righe delle tabelle che contengono le informazioni relative alle singole pubblicazioni, riviste, autori, etc: per essere precisi, i valori effettivi assunti dai dati vanno sotto il nome di istanze.

A dare garanzie circa il corretto funzionamento di un DB è il sistema per la gestione di una base di dati (Data Base Management System), software che consente di:

a) Definire gli schemi di basi di dati;
b) Scegliere la struttura dati per la memorizzazione e l'accesso ai dati;
c) Memorizzare, recuperare e modificare i dati, interattivamente o da appositi programmi, da parte degli utenti autorizzati (Albano et al., 2019).

Per comprendere meglio il legame tra DB e DBMS, si faccia riferimento alla **Figura 1**: grazie alla gestione dei dati e quindi degli schemi, autorizzazioni ed operazioni, il Database Management System consente agli utenti di accedere alle informazioni contenute nel database, o direttamente o tramite quei programmi applicativi che ricevono le medesime informazioni in input.

Si noti tuttavia che molto spesso si è soliti far riferimento ai dati, al sistema DBMS e a tutte le applicazioni associate come sistema di database, ulteriormente abbreviato in database.

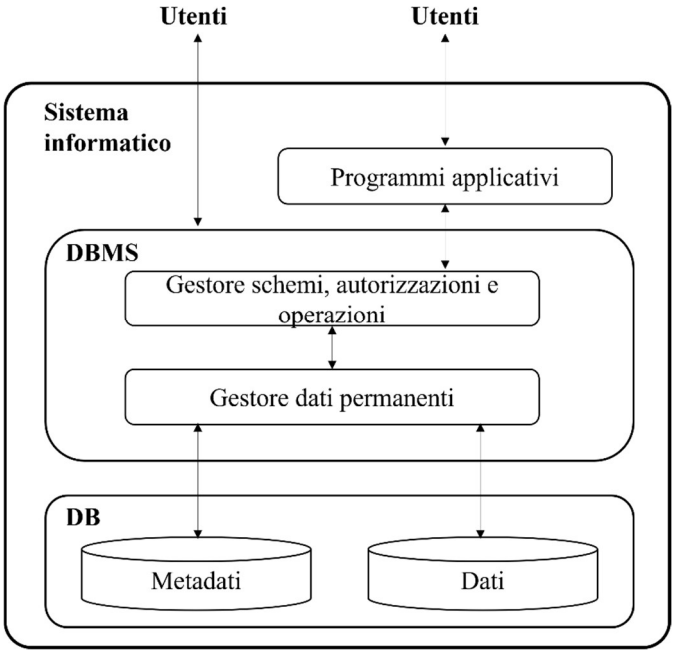


Figura 1: *Il sistema informatico e il sistema per la gestione delle basi di dati (Albano et al., 2019)*

Il concetto di DBMS appena introdotto è stato citato più per una questione di completezza e rigore descrittivo che per un effettivo legame con la seguente trattazione: in fase di definizione dei controlli volti a decretare l'accettazione o il rifiuto dei DB, il focus risulterà essere infatti sulla valutazione della qualità della base di dati in sé, in termini di dati e metadati, più che sull'architettura informatica sovrastante che consente l'accesso agli stessi.
Tuttavia, si noti che ciò con cui ci si interfaccerà nell'esecuzione dei diversi controlli è proprio il sistema informatico e quindi il DBMS dal momento che non si ha un diretto accesso alle tabelle costituenti i database in analisi.

Prima di valutare, anche tramite degli esempi, quali sono le funzionalità delle basi di dati, occorre specificare che possono esserci diversi modelli di dati, ovvero diverse modalità per la loro organizzazione e strutturazione. Tra le principali si annoverano:

- FOGLI DI CALCOLO: rappresentano la forma più semplice di database e sono costruiti in modo tale da avere tante righe quante sono le informazioni che si vogliono storicizzare e tante colonne quanti sono i campi che vorremmo conoscere

in relazione ai diversi record. Questo tipo di archiviazione non è ovviamente molto performante in termini di tempi di ricerca, ridondanza delle informazioni memorizzate e manutenzione delle stesse, ma è pur sempre molto facile da utilizzare in presenza di ridotte quantità di dati da gestire. Un suo altro limite è quello di consentire l'accesso ai dati ad un unico utente o al più ad un numero ridotto di loro, con conseguenti problematiche nell'organizzazione del lavoro.

- DATABASE GERARCHICI (**Figura 2**): derivanti dai fogli di calcolo, utilizzano più file per memorizzare le informazioni, introducendo una relazione gerarchica di tipo padre-figlio (ad albero) fra gli stessi. Si tratta di una tipologia di database non più utilizzata.

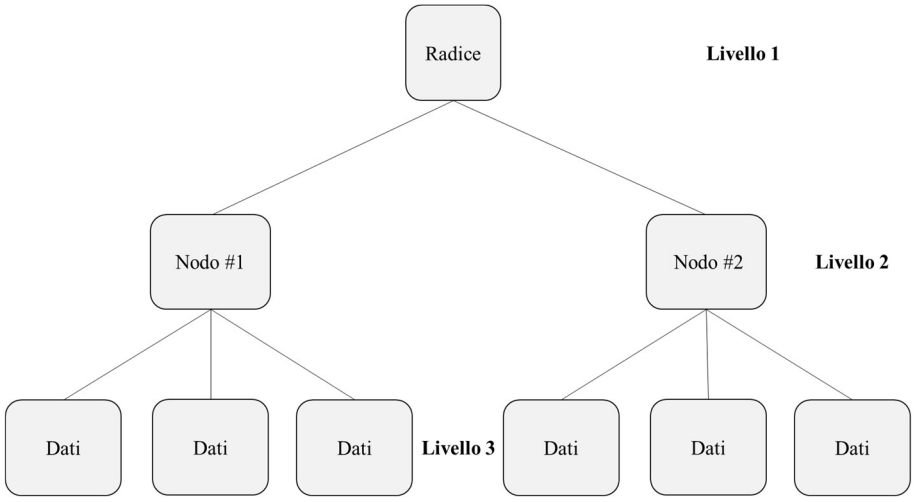


Figura 2: *Esemplificazione di un modello gerarchico di database*

- DATABASE RETICOLARI (**Figura 3**): espandono il concetto di relazione gerarchica, rendendo possibile il raggiungimento di una tabella da più percorsi. A differenza dei database gerarchici, la struttura è a grafo non orientato, i rami sono percorribili in entrambi i sensi e le tabelle sono tutte equamente importanti.
  Nonostante sia possibile definire relazioni multiple tra le varie informazioni, la gestione di tali tipi di database diviene sempre più complessa all'aumentare del numero di relazioni, tant'è che non vengono più utilizzati.

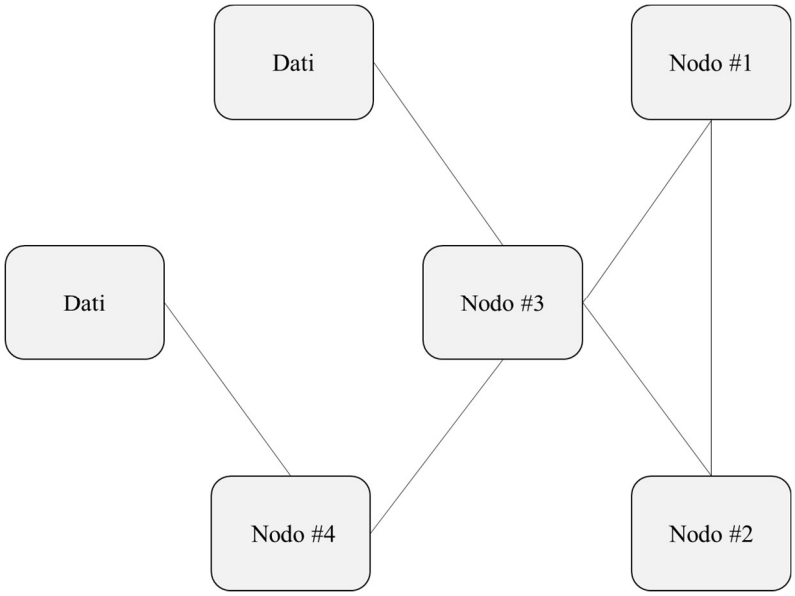


Figura 3: *Esemplificazione di modello reticolare di database*

- DATABASE RELAZIONALI: un database relazionale è organizzato come un set di tabelle, composte da colonne e righe, tra loro correlate.
  Le informazioni da memorizzare nel database vengono inserite nelle righe, costituendo i record, mentre sulle colonne troviamo i campi o attributi.
  Riprendendo l'esempio fatto in **Tabella 1.2**, possiamo individuare gli elementi distintivi di un database relazionale (**Figura 4**).
  Consentono di raggiungere sia elevate prestazioni che flessibilità nella gestione dei dati e per questo rappresentano una delle tipologie di database ad oggi più diffusa.
  Di contro, nei database NoSQL (o non relazionali), è possibile archiviare e manipolare anche dati semi-strutturati e non strutturati e, proprio per questo, la loro diffusione è aumentata notevolmente con la crescente complessità delle applicazioni Web.

- DATABASE AD OGGETTI: i dati vengono archiviati in un oggetto insieme con le loro funzioni, i metodi di accesso all'oggetto e le definizioni degli attributi dello stesso. I singoli oggetti vengono poi raggruppati in classi, definendo anche in tal caso una gerarchia, che non è però rigida come quella del modello gerarchico.

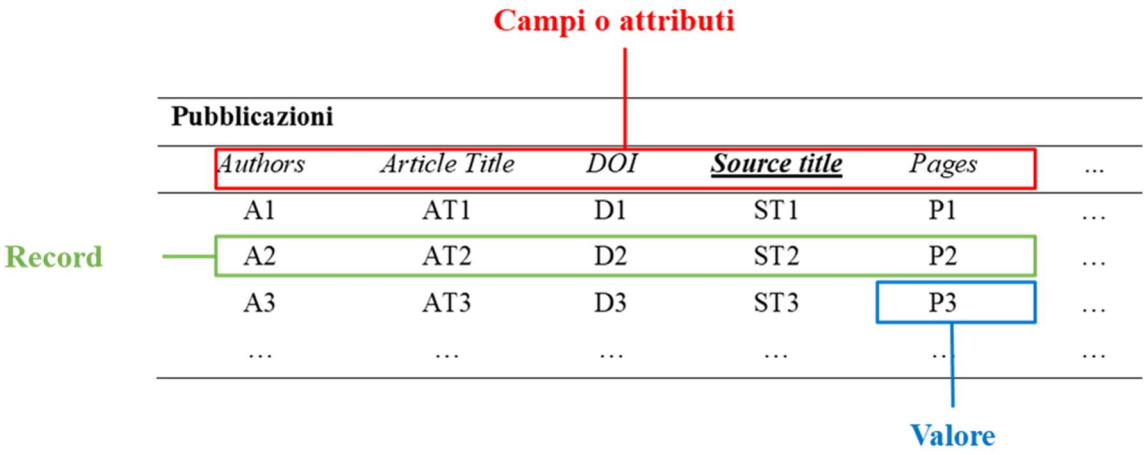


Figura 4: *Elementi costituenti le tabelle di un DB relazionale*

Le tipologie menzionate rappresentano solo alcune delle molteplici alternative oggi disponibili sul mercato: sempre più frequente è infatti la personalizzazione e customizzazione degli applicativi sulla base delle specifiche definite da cliente e key users. Le diverse tipologie di database sopra indicate verranno allora di volta in volta selezionate in modo tale da soddisfare al meglio tali esigenze e tali requisiti.

Quanto appena descritto consente non solo di avere un quadro completo circa il concetto di database, ma anche di definire con esattezza il campo di applicazione della metodologia di controllo di seguito presentata. Nello specifico, l'algoritmo verrà applicato a database relazionali, ma si procederà anche a valutare come lo stesso potrebbe essere rivisto ed adattato alle altre tipologie di strutture menzionate.

Per avere un'idea ancor più concreta di come i database vengono utilizzati in ambito aziendale e non solo, vengono ora presentati degli esempi che vogliono anche mostrare come tali sistemi possono essere più o meno integrati e più o meno complessi a seconda delle specifiche esigenze di business. Si vuole anche evidenziare come, nel caso di prodotti custom, ci sia la possibilità di costruire applicativi specifici per una certa organizzazione e quindi tali da soddisfare al massimo le esigenze espresse dai diversi stakeholders.

Tali esempi sono stati definiti sulla base di applicativi utilizzati e gestiti in esperienze personali e lavorative precedenti.

Esempio 1. Database di gestione di una piccola-media impresa

Si consideri una piccola-media impresa operante nel settore secondario che produce macchinari per la saldatura: viste e considerate le esigenze specifiche che tale realtà potrebbe avere, nonché le dimensioni del business, un applicativo realizzato su misura è la miglior alternativa possibile.

Si noti a tal proposito che la suite Office offre un interfaccia software, Microsoft Access, tramite la quale è possibile costruire DB custom per le organizzazioni interessate ed integranti, al loro interno, tutte le funzionalità desiderate.

Tramite un quadro di comandi generale, allora, sarà possibile accedere alle diverse sezioni, tra le quali si trovano:

- **Relazioni**: consente di accedere all'elenco completo dei soggetti con la quale l'azienda si relaziona, in termini di clienti e fornitori. All'interno della maschera sarà possibile filtrare l'elenco in ordine alfabetico, ma anche visualizzare alternativamente solo i fornitori o solo i clienti. Cliccando sul record corrispondente, sarà poi possibile accedere ad un'ulteriore maschera in cui saranno presenti dati aggiuntivi.
- **Articoli:** sezione nella quale troviamo tutti i componenti che nel corso del tempo l'azienda ha acquistato o realizzato internamente. In relazione ad ognuno di essi, saranno mostrati i campi di interesse, tra i quali un IDArticolo, necessario per identificare univocamente gli stessi. Dal momento che i record memorizzati nel database potrebbero essere particolarmente numerosi, sarà prevista una maschera che consentirà di operare una massiccia filtrazione attraverso apposite queries.
- **Tipologie di macchine**: sono qui elencate le macchine che l'azienda realizza e vende. Selezionando una qualsiasi di queste, si aprirà un elenco di tutti i sottoassiemi che la costituiscono, con la possibilità di definire la sua distinta base.
- Sotto la voce **Macchine ed Impianti** troviamo, invece, tutte i macchinari a disposizione nello stabilimento, classificati sulla base del reparto in cui si trovano.
- **Commesse:** elenco delle commesse che, nel corso del tempo, l'azienda ha ricevuto, con la possibilità di distinguere tra commesse aperte e chiuse.
- **Ordini:** consente di effettuare ordini ai diversi fornitori e di visualizzarne lo storico;
- **Vendite:** sono qui presenti tutti i diversi articoli di vendita, specificando per ognuno di essi il prezzo.

- I pulsanti **Utilità** e **Utilità varie** raccolgono una serie di tabelle, come quelle relative alla classificazione delle relazioni, alle valute, alle causali di documento o ai tipi di pagamento.
- Con **Autorizzazioni** si accede, infine, ad un'area dalla quale si potranno gestire i permessi degli utenti in relazione alle diverse funzionalità del database.

Si può concludere allora come i database possano supportare l'esecuzione e la gestione di una vasta gamma di attività operative, che vanno dalla progettazione agli acquisti, dal pricing alle vendite.

Evidente è come però in tal caso non sia previsto alcun meccanismo automatico di valutazione della qualità dei dati, con tutte le problematiche che ne possono conseguire.

Esempio 2. Database di sintesi dati provenienti da fonti eterogenee

Se nel primo esempio il database assumeva a tutti gli effetti le sembianze di un applicativo gestionale a 360°, obiettivi della base di dati ora presentata sono quelli di raccogliere dati provenienti da fonti eterogenee e definire logiche di archiviazione che siano il più possibile efficienti e funzionali alle successive analisi.

Si consideri allora un'importante società di consulenza che si occupa di analisi previsionale e scenariale: le fasi che segue per completare i suoi task sono di seguito descritte specificando, in relazione ad ognuna, le caratteristiche che il database deve presentare in virtù delle operazioni che su di esso o con esso vengono eseguite.

1. Popolamento del DB: una volta definiti i campi delle tabelle costituenti il database, l'upload dei dati viene effettuato sia tramite caricatori manuali sia tramite opportuni strumenti di *data ingestion* che consentono di automatizzare tali operazioni. A titolo di esempio, si citano:
    - RPA, laddove sia necessario il *download* di file. Per verificare che l'operazione compiuta dal BOT sia conclusa con successo, è previsto un sistema di *quality checking* che consente di visualizzare l'esito del processo e gli eventuali errori che si sono presentati.
    - API, laddove i provider le forniscano.

    Sempre in un'ottica di automatizzazione della fase di popolamento, è possibile schedulare l'aggiornamento dei dati di mercato, predisponendo, in corrispondenza di date ben precise, il download automatico dei dati dai diversi provider e il loro successivo caricamento nel DB.

2. Quality check dei dati: la verifica sulla qualità dei dati, in termini di loro correttezza e consistenza con le altre informazioni contenute nel DB, viene effettuata manualmente e soprattutto sulla base dell'esperienza di coloro che quotidianamente gestiscono ed elaborano le informazioni.

   Tuttavia, nel momento in cui si procede con la modifica manuale di un valore, è previsto un sistema di auditing delle modifiche che non solo tiene traccia dello storico dei valori ma anche del nome dell'utente che ha effettuato la modifica e della data di modifica.
3. Rielaborazioni dei dati: in questa fase, i dati contenuti nel database vengono rielaborati tramite appositi modelli previsionali implementati nell'applicativo.
4. Storicizzazione output: i risultati derivanti dall'applicazione dei modelli previsionali vengono anch'essi memorizzati in apposite sezioni del database, in modo tale che possano poi essere analizzati a vari livelli di dettaglio e diventare input per successive analisi.

Tirando le somme, la trattazione fatta nel presente capitolo e gli esempi appena introdotti consentono di estrapolare alcuni dei vantaggi che un generico database permette di ottenere:

- Archiviazione dati come risorsa comune per l'intera organizzazione, con riduzione di ridondanze ed inconsistenze;
- Maggiore agilità e scalabilità delle organizzazioni, grazie alla possibilità di creazione ed utilizzo diretto dei database;
- Automatizzazione dei processi manuali di storicizzazione, analisi o rielaborazione dati, oggi particolarmente lunghi e costosi. Di conseguenza, gli utenti aziendali potranno dedicarsi ad attività a maggior valore aggiunto;
- Analisi dei dati, anche derivanti da molteplici sistemi;
- Raggiungimento di maggiori performance nell'esecuzione di attività e processi decisionali;
- Raggiungimento di adeguati standard di sicurezza sui dati.

1.4 La qualità dei database

Visto e considerato l'impiego dei database e dei sistemi informatici in generale in un insieme sempre più ampio di aree applicative e la centralità del loro corretto funzionamento ai fini del successo aziendale, sviluppare e selezionare prodotti di qualità è di primaria importanza.
In quest'ottica, la definizione delle specifiche qualitative da un lato e la valutazione della qualità dei sistemi tramite opportune misure dall'altro sono attività chiave che, proprio per tale ragione, costituiscono i pilastri della suddetta trattazione.

A definire come gli sviluppatori, i clienti e i valutatori di prodotti e sistemi software devono condurre questi task è la già citata serie ISO/IEC 25000:2014, anche nota come SQuaRE (*Systems and software Quality Requirements and Evaluation*).
Tale standard definisce la qualità del software come "la capacità di un prodotto software di soddisfare bisogni espliciti ed impliciti in specifiche condizioni d'uso" dove con software, o più precisamente con prodotto software, si intende un "insieme di programmi, procedure e possibilmente documentazione e dati associati".
L'organizzazione della serie è illustrata in **Figura *5*** e di seguito descritta:

- QUALITY MANAGEMENT DIVISION: obiettivo della sezione è quello di definire modelli comuni, riferimenti e vocabolario dell'intera serie SQuaRE, per fornire agli utilizzatori linee guida univoche.

- QUALITY MODEL DIVISION: vengono qui presentati i modelli di qualità validi per sistemi e software, qualità in uso e dati. In particolare, la ISO 25012 definisce il data quality model da applicare ai dati organizzati in formato strutturato e memorizzati in un sistema informatico.
  Lo stesso modello può essere utilizzato per definire i requisiti e le metriche di valutazione della DQ, nonché per pianificare ed eseguire valutazioni a riguardo.

- QUALITY MEASUREMENT DIVISION: focus degli standards che costituiscono tale sezione è quello di definire, anche analiticamente, i modelli di riferimento per la misurazione della qualità di software e sistemi e di fornire chiare linee guida per la loro applicazione.
  Vengono inoltre introdotte metriche per la qualità interna, esterna e in uso.

La ISO 25024 definisce le metriche da adottare per quantificare la qualità dei dati, sulla base delle caratteristiche definite dalla ISO 25012.

- QUALITY REQUIREMENTS DIVISION: i requisiti specificati in tale sezione possono essere utilizzati nel processo di raccolta degli stessi in relazione a nuovi prodotti da sviluppare o come input per le operazioni valutative.

- QUALITY EVALUATION DIVISION: obiettivo della sezione è fornire requisiti, raccomandazioni e linee guida per la valutazione dei prodotti, a prescindere dal soggetto che la esegue.

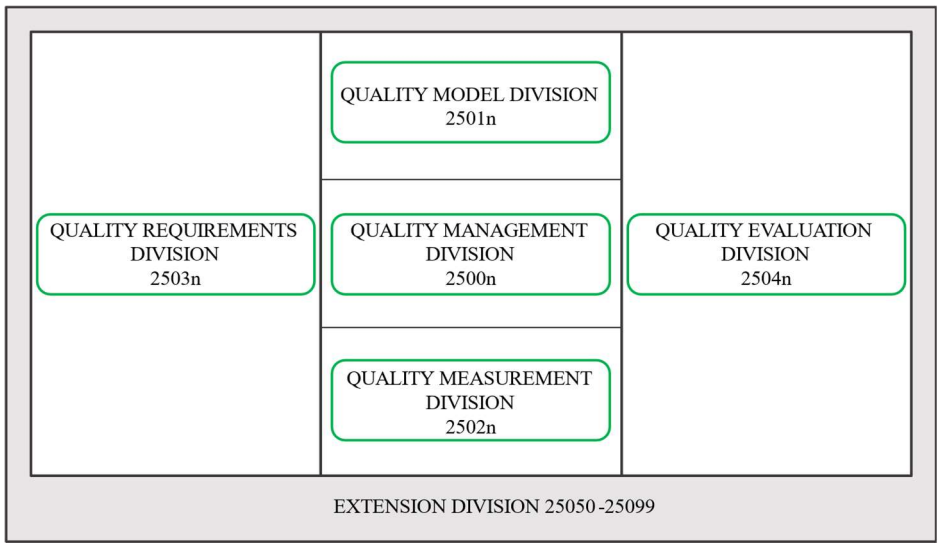


Figura 5: *Organizzazione della serie di standard SQuaRE (ISO 2500:2014. Adattata)*

Si noti poi che il modello di qualità SQuaRE prevede tre distinte caratteristiche:

- Modello per la qualità di software e sistemi
- Modello per la qualità in uso
- Modello per la qualità dei dati

Dal momento che la qualità dei dati scambiati, processati ed usati è un importante prerequisito per garantire adeguati livelli prestazionali complessivi dei database, la nostra attenzione verrà ora rivolta proprio all'ultimo modello, estrapolando trasversalmente dalle diverse sezioni e documenti le nozioni di interesse: se i dati contenuti non sono di qualità, i database non potranno che avere le stesse caratteristiche secondo il principio del *garbage in- garbage out.*

A tal proposito, si introducono le seguenti definizioni: se per dato si intende la "rappresentazione di informazioni in una modalità formalizzata idonea alla comunicazione, interpretazione e elaborazione" (ISO/IEC 12207:2008), la data quality viene definita come il "grado con cui le caratteristiche dei dati soddisfano bisogni espliciti ed impliciti, quando vengono usati in specifiche condizioni" (ISO/IEC 25000:2014).
Tutte le diverse tipologie di dato (i.e. stringhe, testo, date, numeri, immagini, suoni, etc.) vengono considerate dallo standard, che non tralascia neppure la qualità dei valori assegnati ai dati e le relazioni tra essi esistenti.

La necessità di porre il focus sulla qualità dei dati è legata al fatto che l'importanza della sua gestione è comunemente riconosciuta per tutta una serie di motivi:

1. I dati potrebbero essere acquisiti da organizzazioni caratterizzate da un processo di produzione dei dati ignoto o debole qualitativamente parlando;
2. Potrebbe essere necessario dover processare dati ambigui o caratterizzati da mancanza di coerenza con altri dati esistenti;
3. Acquisire ed elaborare i dati in accordo a standard riconosciuti ed accreditati consente di avere benefici in termini di interoperabilità e condivisione degli stessi tra diversi utilizzatori;
4. Se i dati in input sono difettosi, le informazioni derivanti saranno inadeguate, i risultati inutilizzabili e i clienti insoddisfatti;
5. Esistono sistemi informativi (come il world wide web) in cui i dati cambiano frequentemente, per i quali la gestione secondo prassi standard è fondamentale;
6. Potrebbero essere individuati errori ed inefficienze e dunque successive azioni correttive inerenti:
    - Dati (i.e. riprogettazione; analisi; trasformazione, …)
    - Software (i.e. modifica dei programmi per implementare ad esempio controlli automatici; …)
    - Hardware (i.e. modifica dei sistemi per ridurre i tempi di risposta; …)
    - Processi condotti dagli umani (i.e. formazione degli utenti per evitare errori nel processo di immissione dati; …)

Contenuto attorno al quale ruota la ISO 25012:2008 è poi la definizione delle caratteristiche dei dati target usati tanto dai sistemi informatici quanto dagli umani nelle successive

rielaborazioni; con dati target si intendono tutte quelle informazioni che l'organizzazione decide di analizzare e validare attraverso il modello. La **Figura *6*** mette in evidenza la loro collocazione in un generico sistema informatico.

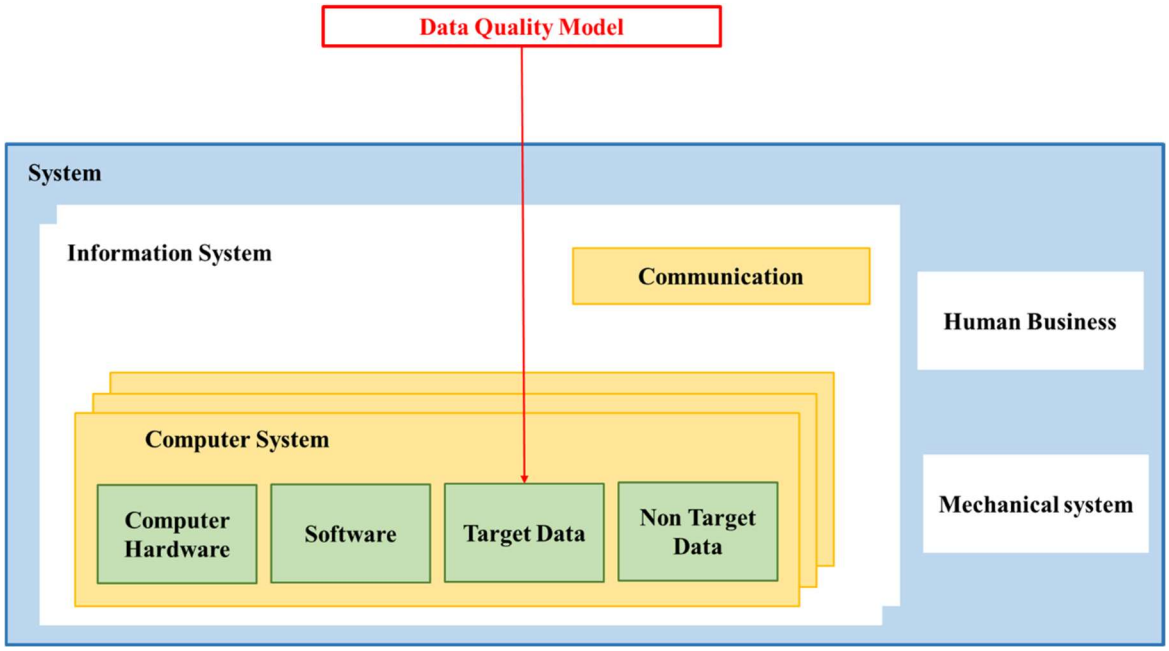


Figura 6: *Ambito di applicazione del modello di data quality (ISO/IEC 25012:2008)*

Di contro, i dati che non costituiscono il target appartengono a due distinte famiglie:

- Dati non persistenti
- Dati sui quali l'organizzazione ha scelto di non applicare il modello

Le caratteristiche che i dati target devono soddisfare sono molteplici perché, come sottolineato anche in letteratura, il concetto di data quality è multidimensionale; a tal proposito, il modello previsto dalla norma prevede quindici caratteristiche, che possono essere inerenti e/o dipendenti dal sistema e le cui priorità ed importanze dovranno essere valutate a seconda dello stakeholder considerato.

Caratteristiche inerenti

Si tratta di caratteristiche che si riferiscono ai dati stessi, ad esempio in termini di metadati, relazioni, dominio e restrizioni dei valori attribuibili ai dati.

Tabella 1.4: *Caratteristiche inerenti e loro definizione*

CARATTERISTICA	DEFINIZIONE
ACCURATEZZA	Grado di accordo tra il valore reale del dato e quello considerato corretto. Può essere sintattica o semantica.
COMPLETEZZA	Grado con cui sono presenti valori dei dati per tutti i diversi campi previsti (completezza di colonna) e per l'intera popolazione osservata (completezza di popolazione)
COERENZA	Grado con cui i dati sono privi di contraddizioni e coerenti con gli altri dati, in uno specifico contesto di utilizzo. Può essere di chiave (non devono esserci record con il medesimo valore in corrispondenza del campo individuato come chiave primaria), interna (non devono esserci contraddizioni tra i valori di un'istanza) ed esterna (deve esserci coerenza tra i dati contenuti in uno stesso campo per record distinti).
CREDIBILITA'	Grado con cui i dati presentano valori considerati come veri e credibili dagli utenti. Viene qui incluso il concetto di autenticità. Il requisito corrispondente viene considerato soddisfatto qualora i dati siano stati certificati da un ente indipendente.
ATTUALITA'	Grado con cui i dati hanno valori che sono temporalmente in linea con lo specifico contesto di utilizzo.

Caratteristiche dipendenti dal sistema

Tali caratteristiche considerano la qualità dei dati come dipendente dal contesto in cui gli stessi vengono utilizzati e quindi dai sistemi informatici di contorno.

Tabella 1.5: *Caratteristiche dipendenti dal sistema e loro definizioni*

CARATTERISTICA	DEFINIZIONE
DISPONIBILITA'	Grado che indica quanto facilmente i dati possono essere recuperati da utenti e/o applicazioni autorizzati.
PORTABILITA'	Grado con cui i dati possono essere caricati, sostituiti o spostati tra vari sistemi, mantenendo la qualità originaria.

RECUPERABILITA’	Grado con cui i dati sono conservati in un’ambiente sicuro che ne consente il recupero, anche e soprattutto in caso di malfunzionamenti.

Ci sono infine caratteristiche che sono sia inerenti che dipendenti dal sistema: in tal caso saranno previste metriche di data quality diverse in relazione ai due filoni.

Tabella 1.6: *Caratteristiche inerenti e dipendenti dal sistema e loro definizioni*

CARATTERISTICA	**DEFINIZIONE**
ACCESSIBILITA’	Grado con cui i dati sono accessibili da tutti in uno specifico contesto di utilizzo.
CONFORMITA’	Grado con cui i dati rispettano standard, convenzioni o normative in vigore o regole simili relative alla qualità dei dati.
RISERVATEZZA	Grado con cui i dati risultano essere strutturati in modo tale da essere accessibili solo dagli utenti autorizzati.
EFFICIENZA	Grado con cui i dati vengono processati utilizzando quantità e tipi di risorse appropriate.
PRECISIONE	Grado con cui i dati hanno attributi esatti o che consentono di operare una discriminazione.
TRACCIABILITA’	Grado con cui sono disponibili campi che consentono di registrare gli accessi ai dati e/o lo storico delle modifiche introdotte.
COMPRENSIBILITA’	Grado con cui i dati sono espressi con linguaggi, simbologia e unità adeguate e tali da rendere la loro lettura ed interpretazione chiara ed immediata

Le dimensioni appena introdotte possono essere definite a diversi livelli di dettaglio:

- Schema: la qualità del dato viene valutata in relazione agli schemi logici impiegati per rappresentare i diversi dati, per verificarne la loro idoneità ed adeguatezza.
- Processo: in tal caso il focus risulta essere sul processo tramite il quale il dato viene raccolto.
- Dato: la qualità viene valutata in relazione al dato puro memorizzato, indipendentemente dalle modalità di rappresentazione e dal processo di raccolta.

Dal momento che il punto di vista adottato nella seguente trattazione prevede una verifica del dato in sé, il livello al quale faremo riferimento sarà l'ultimo citato: tale scelta risulta essere dettata anche dal fatto che il processo di raccolta dei dati da parte dei DB bibliometrici non è noto, così come non si hanno informazioni approfondite sugli schemi utilizzati.

Saranno allora proprio queste caratteristiche a rendere possibile la valutazione della qualità dei database da parte di tutti coloro che dovranno eseguire questo task in virtù delle loro responsabilità: clienti, fornitori, sviluppatori, utenti, valutatori, manutentori, etc.
A seconda dell'ottica adottata, saranno poi introdotte apposite metriche (QM o Quality Measures), definite in accordo alla ISO 25024:2014.
Si noti che il fatto che sia stato l'ente normatore a definire quali siano le QM più opportune da utilizzare per misurare la qualità dei dati ha consentito di superare un limite esistente in passato quando l'esistenza di dimensioni ed indicatori sviluppati ad hoc per risolvere specifici problemi rendeva i risultati ottenuti affetti da un elevato grado di soggettività.

Nonostante l'utilizzo di tali indicatori sia in questa trattazione funzionale alla valutazione della qualità dei DB da parte del committente che sceglie di acquisire questo prodotto, è importante sottolineare come tali metriche possano essere utilizzate anche in altre fasi del ciclo di vita dei dati (DLC o data life cycle), esemplificato in **Figura** *7* .

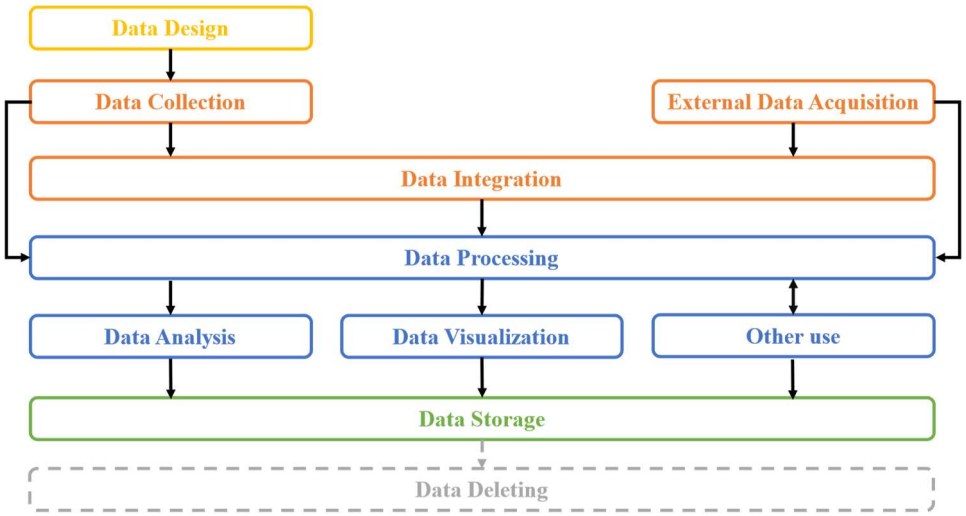


Figura 7: *Esempio di DLC*

Ad esempio, le caratteristiche e le relative metriche possono essere utili per:

- Definire i requisiti nella fase di design della base dati; per definire in modo univoco le specifiche tecniche del prodotto che meglio rispondono a tali requisiti, si potrebbe utilizzare il già citato QFD (Quality Function Deployment). L'adeguata applicazione di tale metodo, congiuntamente agli standard SQuaRE, consentirebbe di raggiungere elevati standard qualitativi.
- Confrontare la qualità dei dati di alternative disponibili, qualora si scelga di acquisire il DB da fornitori esterni
- Implementare e supportare processi di data governance, management e documentation, nonché processi di gestione dei servizi IT;
- Migliorare la data quality e l'efficacia dei processi decisionali;
- Valutare la qualità dei sistemi o dei software che generano dati come output.

2. CAPITOLO II
QUALITA' IN FASE DI ACQUISTO DI UN DATABASE

2.1 Richiami su controlli di accettazione e piani di campionamento

Come già specificato, obiettivo della trattazione è in primis quello di definire una metodologia che consenta agli acquirenti di un ipotetico database di prendere una decisione circa l'accettazione o il rifiuto dello stesso. In virtù dell'ottica adottata, tale decisione è strettamente interconnessa alla qualità dei dati presenti nei DB, da valutare grazie all'applicazione di opportuni metodi di controllo statistico: più nello specifico, i tool che supportano clienti e fornitori nell'esecuzione di tali prove in accettazione risultano essere i piani di campionamento. In teoria, però, quando si parla di controllo qualità di un processo, occorre distinguere tra:

1. Controllo di fabbricazione, eseguito nel corso del processo di produzione del prodotto. Nel caso dei database, si tratterebbe di quei test eseguiti man mano che si procede con l'implementazione dello stesso applicativo. Tali controlli dovrebbero comunque essere effettuati sulla base delle caratteristiche tecniche e delle metriche definite dallo standard SQuaRE.
2. Controllo di accettazione, volto a valutare se il prodotto soddisfa le specifiche di tipo ingegneristico ed è quindi idoneo ad essere utilizzato. È questo il punto di vista che si adotterà, dal momento che l'interesse è proprio sulla qualità del prodotto in esame, indipendentemente dalle modalità con cui risulta essere stato organizzato il processo produttivo in sé.

Per comprendere a pieno gli obiettivi che l'applicazione di tali controlli consente di raggiungere, si procede ora con un esempio relativo al caso di studio considerato: si immagini di essere un ipotetico acquirente di un database bibliometrico, quale ad esempio un'ente pubblico che desidera misurare la produzione annuale dei ricercatori per concedere poi, a coloro che si saranno contraddistinti per le loro attività, dei premi monetari. Ciò è quello che fa, tra le altre cose, l'ANVUR in fase di valutazione delle attività di ricerca in Italia.

Tale processo dovrà concludersi con la redazione di una graduatoria e con la diffusione di documenti ad hoc volti a rendere noti gli esiti: saranno allora previste diverse rielaborazioni in sequenza delle informazioni in input (**Figura *8***).

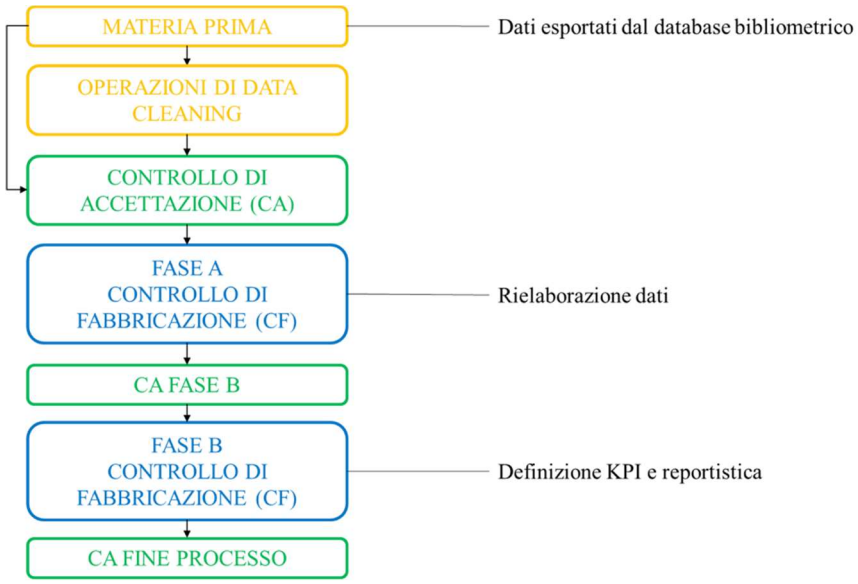


Figura 8: *Esemplificazione dell'applicazione dei controlli di accettazione e fabbricazione*

Nel dettaglio, materia prima di tale processo di rielaborazione delle informazioni sono i dati estratti dal database di tipo bibliometrico, sottoposti ad un primo controllo di accettazione da parte dell'ente pubblico.

Qualora si proceda con l'applicazione del controllo senza aver prima eseguito alcun tipo di operazione di data cleaning, la probabilità di accettazione della base dati risulterà essere generalmente bassa, soprattutto se la progettazione e il testing della stessa non sono stati eseguiti sulla base delle stesse dimensioni caratterizzanti la metodologia di assessment della qualità.

Proprio per tale motivo, possono essere previste, prima di eseguire il controllo, delle operazioni di pulizia dei dati che probabilmente consentiranno di avere una maggiore probabilità di successo: visto che nel caso di studio considerato si procederà con tutta una serie di operazioni di riorganizzazione dei dati prima di applicare la metodologia di verifica, il controllo di accettazione a cui si farà riferimento sarà proprio quello sopra menzionato.

Eseguito tale step, si procederà allora con le rielaborazioni di interesse previste dalla fase A e funzionali alla generazione dell'output richiesto.

Secondo una modalità di lavoro di tipo Stage & Gate, verrà a questo punto eseguito un nuovo CA per verificare l'idoneità delle informazioni prodotte a passare alla fase B.

Con la fase B, il grado di rielaborazione dei dati sarà ancora maggiore e anche qui potrebbe essere previsto un controllo in itinere, eseguito man mano che i dati grezzi vengono elaborati e trasformati in informazioni.
Gli output di tale fase saranno dei KPI, che consentiranno di mettere in graduatoria i diversi ricercatori, nonché tutta quella documentazione volta alla comunicazione delle informazioni di interesse. Ovviamente, prima di essere diffuse, tali informazioni dovranno essere sottoposte ad un ulteriore controllo di accettazione di fine processo.

Alcune precisioni:

- I controlli di accettazione possono essere finalizzati a controllare la materia prima direttamente proveniente dal fornitore, come nel caso in esame (controllo in ingresso), ma anche quanto prodotto tra una fase e l'altra del processo di lavoro (controllo in progress) e gli elementi in uscita.
- Se nelle prime fasi, quando si ha a che fare con dati grezzi, è possibile e ha senso automatizzare i controlli di accettazione, definendo delle regole standard, man mano che le informazioni diventano sempre più aggregate e a valore aggiunto, la valutazione della loro qualità, soprattutto in termini di correttezza, dovrà essere effettuata manualmente dal personale preposto alla loro approvazione e divulgazione.
- Eseguire sia i controlli di accettazione che quelli di fabbricazione implica dei costi e quindi avrà senso implementarli laddove la criticità della fase di applicazione e l'impatto dei piani in termini di benefici sono entrambi particolarmente elevati.

Il controllo che si andrà ad eseguire sui database bibliometrici sarà effettuato per valutarne la difettosità e, in maniera complementare, la qualità degli stessi e quindi prendere conseguenti decisioni circa l'accettazione o meno della fornitura: è dunque classificabile a tutti gli effetti come un controllo di accettazione, che di conseguenza ha i seguenti obiettivi:

- Verificare la conformità alle specifiche funzionali;
- Verificare la percentuale massima e/o media di elementi difettosi, ovvero non dotati di caratteristiche conformi alle specifiche;
- Condurre audit sulla qualità del prodotto;
- Motivare gli addetti nella logica "del miglioramento continuo": se si sa che l'output di una qualche attività verrà controllato, chi esegue la stessa sarà più attento e motivato nel trovare soluzioni che ne incrementino la qualità.

Qualora l'esito del controllo di accettazione eseguito dal committente fosse negativo e il prodotto non soddisfacesse le specifiche:

1. Se non fosse possibile apportare modifiche migliorative, dovrà essere scartato.
2. Se invece fosse possibile apportare modifiche, si dovrà intervenire su tutti quei dati individuati come errati, in modo tale che vengano sistemati, corretti e quindi riverificati tramite un nuovo controllo. Il problema principale è però dato dal fatto che effettuare queste rilavorazioni implica dei costi.

Per entrare più nel dettaglio del piano che si dovrà implementare, occorre specificare che possono esserci diverse tipologie di controllo.

1. Controllo "a tappeto" o al 100%: dovremmo considerare tutti i record presenti nel database in esame e controllare che le informazioni riportate siano o meno corrette.
   Nonostante l'esecuzione di questo tipo di controllo darebbe la massima garanzia sull'esito, a prescindere da errori e superficialità, i costi e i tempi richiesti sarebbero troppo proibitivi. Qualora potesse esser messa in piedi una metodologia di controllo totalmente automatica, la scelta ricadrebbe chiaramente su questa tipologia di ispezione.
2. Controllo campionario: in tal caso si accetta una certa percentuale di difettosità nelle informazioni fornite dal momento che non si controlleranno tutti i record, ma solo un sottoinsieme degli stessi.
   Il controllo campionario è a sua volta distinguibile in:
   - Controllo "percentuale": si decide a priori di controllare una % dell'intera popolazione, indipendentemente da quanto è grande la popolazione, con l'idea, errata, di ottenere un grado di protezione costante.
   - Controllo "statistico": in tal caso si procederà con la determinazione del campione e dei criteri di accettazione sulla base della numerosità e della tipologia qualitativa della popolazione.
   - Controllo a spot: si effettuano dei controlli di tanto in tanto.

In aggiunta, è sempre possibile non eseguire alcun tipo di controllo.

Nel caso in esame, si procederà con un controllo campionario di tipo statistico dal momento che così facendo non solo riusciremo a ridurre i tempi e il consumo di risorse ma anche, in

caso di diffusa applicazione di tale metodologia, a stimolare i fornitori nel concentrarsi sulla qualità dei database.

È allora necessario definire un piano di campionamento ed il suo opportuno dimensionamento; in primis, occorre scegliere il tipo di piano, dal momento che possono essere classificati in due famiglie principali:

1. Piani di campionamento per variabili: ciò che viene controllato è una vera e propria variabile misurabile ed il ragionamento sull'accettazione e/o rifiuto si basa sul controllo dei valori assunti dalla sua distribuzione;
2. Piani di campionamento per attributi: si basano sull'analisi della presenza di elementi difettosi o non difettosi. Possono essere semplici, doppi, multipli, sequenziali.

In relazione a questi ultimi, devono essere introdotti una serie di parametri, per i quali si fornisce nella **Tabella 2.1** una breve descrizione.

Tabella 2.1: *Definizioni relative ai controlli di accettazione*

	DEFINIZIONE
LOTTO	Quantità di un prodotto o di un materiale accumulatosi sotto condizioni che sono valutate come uniformi dal punto di vista del campionamento. La numerosità del lotto viene indicata con N.
DIFETTOSO	Elemento di fornitura non conforme ad una o più specifiche. Il numero di elementi difettosi nel lotto viene indicato con m ≤ N.
DIFETTO	Manifestazione del non raggiungimento di una specifica
PERCENTUALE DI DIFETTOSI (p)	Rapporto tra il numero di elementi di fornitura difettosi e la dimensione del lotto. Questo valore è ottenibile solo nel caso in cui si conosca il valore esatto di elementi difettosi presenti nel lotto. In simboli, $p = \frac{m}{N}$, con $0 \leq p \leq 1$
CAMPIONE	Insieme di elementi estratti casualmente dal lotto. La numerosità del campione viene indicata con n.
NUMERO DI ACCETTAZIONE	Massimo numero di difettosi ammessi nel campione, indicato con c.

Il caso più semplice, nonché quello da noi adottato, è relativo ad un piano singolo dove si procede all'estrazione del campione di numerosità n dal lotto di N elementi. Quindi, dall'esame degli stessi si deciderà se accettare o meno la fornitura; in questo contesto vanno considerate le posizioni di fornitore e committente.

Il primo, definito un certo limite di difettosità p_1, vorrebbe che tutti i lotti con p inferiore a p_1 (o AQL) fossero sempre accettati e che quelli che invece presentano un valore superiore rifiutati; dal canto suo, il committente desidererebbe che tutti i lotti con p superiore a p_2 (LTPD, soglia da lui definita) fossero sempre rifiutati e sarebbe invece d'accordo ad accettare tutti quei lotti con difettosità al di sotto di tale limite.

Affinché ciò sia possibile, è in primis necessario che $p_1 \leq p_2$.

Dal momento che però il controllo non è a tappeto, ma statistico, le posizioni teoriche di cliente e fornitori si modificano ed entrambi accetteranno di correre un certo rischio (**Figura 9**):

- Il fornitore accetta di vedersi rifiutati lotti che in realtà dovrebbero essere accettati (Rischio α o del fornitore)
- Il cliente acconsente ad accettare alcuni lotti che in realtà dovrebbero essere rifiutati (Rischio β o del committente)

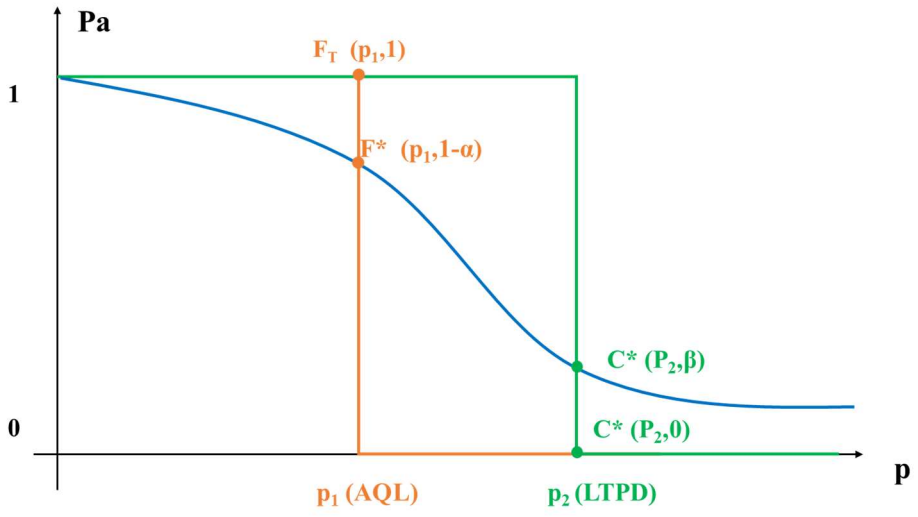


Figura 9: *Rappresentazione grafica delle posizioni teoriche e pratiche di fornitore e committente*

Adottando un punto di vista più operativo, si deve ricordare come la progettazione dei piani di campionamento possa avvenire secondo diverse modalità:

a. Progettazione di piani ad hoc
b. Progettazione secondo norme (i.e. ISO, MIL STD, etc)

Progettazione ad hoc di un piano semplice

È questo il caso in cui fornitore e committente si accordano per definire uno specifico piano di campionamento: sulla base dei valori di AQL, LTPD, α e β, si dovranno determinare la numerosità del campione da estrarre (n) e il numero di accettazione (c).

Ciò può essere fatto risolvendo il seguente sistema:

$$\begin{cases} 1-\alpha = \sum_{i=0}^{c} \binom{n}{i} * AQL^i * (1-AQL)^{n-i} \\ \beta = \sum_{i=0}^{c} \binom{n}{i} * LTPD^i * (1-LTPD)^{n-i} \end{cases}$$

Le due equazioni derivano dal fatto che, per un piano singolo ed emulando un processo di estrazione mediante distribuzione binomiale, la probabilità di accettazione può essere calcolata come:

$$P_a(N,n,c,p) = \sum_{i=0}^{c} \binom{n}{i} * p^i * (1-p)^{n-i}$$

Si è utilizzata la distribuzione binomiale e non la ipergeometrica poiché, in relazione al caso di studio analizzato, risulta che $\frac{n}{N} \leq \frac{1}{10}$.

Tale espressione rappresenta, al variare di p, la curva "Caratteristica Operativa" (curva OC) di un piano di campionamento (curva blu in **Figura 9**).

Dal momento che non è poi così agevole risolvere un sistema come quello sopra indicato, si utilizzano frequentemente appositi normogrammi, quali ad esempio il diagramma di Larson (**Figura *10***).

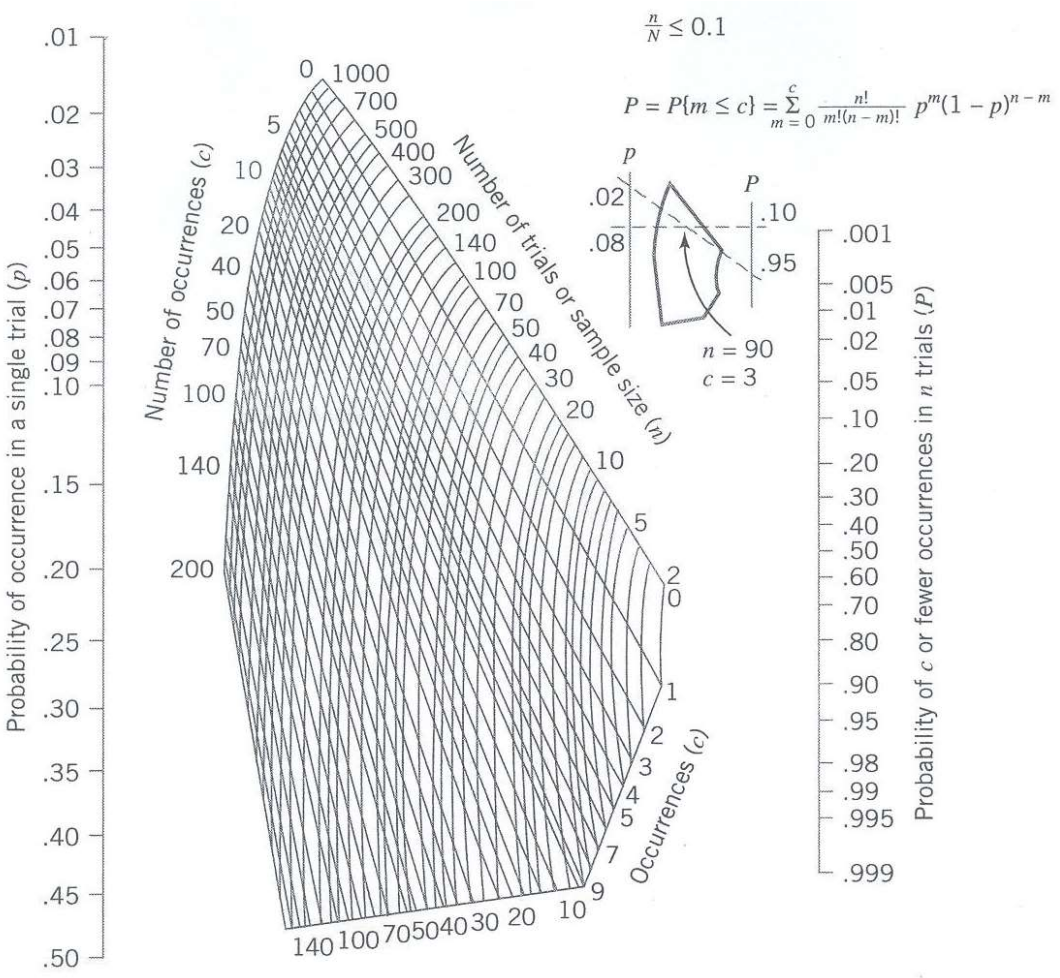


Figura 10: *Esempio di normogramma (diagramma di Larson)*

Progettazione secondo norme (i.e. ISO, MIL STD, etc)

In alternativa, il fornitore e il cliente possono farsi guidare dalle norme nella definizione dei controlli da attuare.

Figura 11: *Utilizzo delle norme nei piani di campionamento semplici per attributi*

A tal proposito, una delle norme più note è la UNI ISO 2859, evoluzione della MIL STD 105E: questa deve essere considerata come una black box che, sulla base delle informazioni in ingresso, consentirà di ottenere dei valori in uscita. Per un piano di campionamento singolo per attributi, gli input e gli output sono mostrati in ***Figura 11***.

Punto interessante da menzionare è il fatto che la stessa norma sottolinea come i piani e gli schemi di campionamento in essa definiti possano essere applicati, tra le altre, anche a dati e record: nonostante non sia comune applicare piani di campionamento ai database, lo stesso standard prevede questa casistica.

In relazione ad un piano singolo, ma eventualmente anche doppio o multiplo, la norma prevede tre diverse tipologie di ispezione (Normale, rinforzata o ridotta).
In linea generale, dal momento che inizialmente il rapporto tra fornitore e cliente è buono, si parte con l'ispezione normale; qualora ci si accorga di evidenti problemi di qualità, si passerà all'ispezione rinforzata. Se nonostante la maggiore severità del piano, non si otterrà alcun tipo di miglioramento, il fornitore non verrà più utilizzato.
Viceversa, qualora si riscontri in un'ispezione normale una qualità della fornitura particolarmente alta, si passerà ad un piano ridotto.
Nel concreto, i passi da eseguire per applicare la norma sono i seguenti:

1. Selezionare il valore di AQL.
2. Selezionare il livello di ispezione, che può essere generale (I, II o III) o speciale (S1, S2, S3 o S4). I livelli speciali vengono utilizzati per campioni piccoli, caratterizzati da un alto livello di rischio.
3. Definire la numerosità del lotto, in modo da estrarre un codice letterale che consentirà di definire i valori di n, sulla base della tabella in ***Figura 12***.
4. Selezionare il tipo di piano desiderato (semplice, doppio, multiplo)
5. Consultare la tabella (**Figura *13***) relativa al tipo di piano e al livello di ispezione desiderati per definire i valori di n, c e r (numero di rifiuto).

Una delle problematiche più grandi connesse all'applicazione di tale norma sta nel fatto che considera, ai fini della definizione di n e c solo la posizione del fornitore, non consentendo di tenere in considerazione anche la posizione del committente; a ciò si può ovviare definendo il

valore di AQL in funzione di LTPD, ad esempio considerando la relazione che dice AQL ≤ LTPD.

Lot or Batch Size	Special Inspection Levels				General Inspection Levels		
	S-1	S-2	S-3	S-4	I	II	III
2 to 8	A	A	A	A	A	A	B
9 to 15	A	A	A	A	A	B	C
16 to 25	A	A	B	B	B	C	D
26 to 50	A	B	B	C	C	D	E
51 to 90	B	B	C	C	C	E	F
91 to 150	B	B	C	D	D	F	G
151 to 280	B	C	D	E	E	G	H
281 to 500	B	C	D	E	F	H	J
501 to 1200	C	C	E	F	G	J	K
1201 to 3200	C	D	E	G	H	K	L
3201 to 10000	C	D	F	G	J	L	M
10001 to 35000	C	D	F	H	K	M	N
35001 to 150000	D	E	G	J	L	N	P
150001 to 500000	D	E	G	J	M	P	Q
500001 and over	D	E	H	K	N	Q	R

Figura 12: *Tavola per definizione numerosità del campione MIL STD 105E*

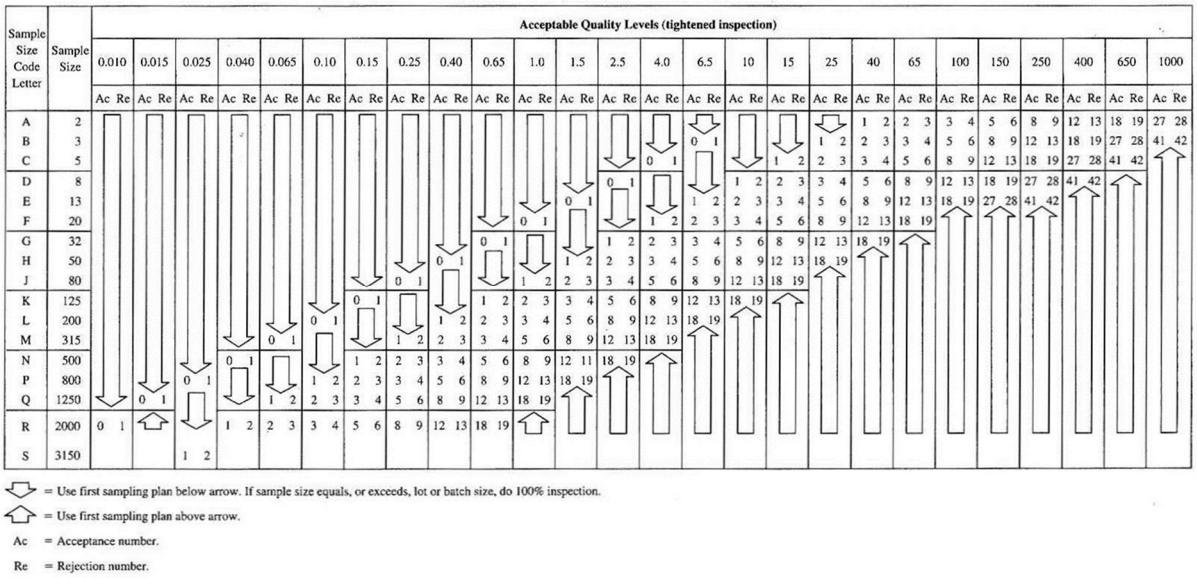


Figura 13: *Tavola per la definizione di numero di accettazione e numero di rifiuto nel caso di ispezione ridotta (MIL STD 105E)*

2.2 Panoramica delle tecniche di data quality assessment

Prima di procedere alla presentazione, in termini generali, della metodologia che consentirà di valutare la qualità dei database bibliometrici, occorre fare una panoramica sulle tecniche disponibili in lettura per il quality assessment dei dati. Con tale termine, si indicano le attività di misurazione e valutazione della qualità delle raccolte dati in relazione a tutta una serie di dimensioni ritenute rilevanti: con le misurazioni andremo allora concretamente a quantificare il valore di queste dimensioni, per poi confrontare i risultati ottenuti con quelli di soglie di riferimento e ottenere un vero e proprio giudizio e/o decisione.

Le fasi normalmente previste in tale processo sono di seguito riportate:

1. Analisi dei dati: obiettivo è quello di ottenere una visione d'insieme sui dati, la loro architettura e le regole di gestione, al fine di raggiungere un certo livello di conoscenza delle informazioni sulle quali si dovrà lavorare.
2. Analisi dei requisiti: vengono raccolte le opinioni di tutti gli stakeholders coinvolti nel DLC con molteplici scopi, quali identificare i requisiti da soddisfare, individuare i problemi di qualità e definire nuovi target;
3. Identificazione aree critiche: vengono identificate le aree e i flussi dati più importanti da esaminare;
4. Modellazione di processo: viene realizzato un modello relativo al processo che produce e/o aggiorna i dati, per essere poi in grado di definire quali sono i passaggi più critici;
5. Misurazione della qualità: si selezionano le dimensioni da valutare, derivanti dalla fase di analisi dei requisiti, e si definiscono le corrispondenti metriche di valutazione.
   Le misurazioni condotte possono essere oggettive o soggettive a seconda che siano basate su metriche quantitative o su valutazioni qualitative effettuate da utenti ed amministratori dei dati e dovranno focalizzarsi sulle aree definite in virtù degli studi condotti ai punti 3 e 4.

Le tecniche di seguito presentate sono quelle che prevedono, tra le altre fasi, la misurazione della qualità: nessuna di loro prevede l'esecuzione di un piano di campionamento o il chiaro riferimento alla normativa SQuaRE, anche a causa del fatto che molte risultano essere antecedenti alla data di pubblicazione dello standard.

Dunque, per ognuno dei metodi, verrà posto l'accento sullo step di error detection e sulle caratteristiche dei dati testate. Ciò consentirà di effettuare dei confronti mirati con l'algoritmo proposto e di conseguenza di metterne in evidenza punti di forza e di debolezza.

TDQM (Total Data Quality Management, Wang, 1998)

Prima metodologia pubblicata nell'ambito della letteratura in materia di data quality, si pone l'obiettivo di estendere i principi del Total Quality Management anche ai dati e di supportare il processo di miglioramento della qualità in tutte le sue fasi.

Va dunque a considerare i dati come un vero e proprio prodotto e si basa sull'idea che per ottenere informazioni di elevata qualità deve essere utilizzato lo stesso framework impiegato nella produzione di componenti fisici. L'applicazione dei piani di campionamento introdotta nella suddetta trattazione rimanda allora perfettamente a questo principio.

Si distinguono quattro ruoli principali:

1. Fornitori di informazioni che creano e raccolgono i dati;
2. Produttori di informazioni che disegnano, sviluppano e si occupano della manutenzione dei dati e dei sistemi collegati;
3. Consumatori di informazioni che usano le informazioni nelle loro attività;
4. Manager del processo informativo che hanno il compito di gestire l'intero ciclo di vita delle informazioni.

Questi soggetti vengono alternativamente coinvolti nelle fasi previste da tale metodologia: step iniziale è quello di definizione dell'insieme dei dati da considerare, sia in termini di funzionalità che gli stessi dovranno avere che in termini di sistema informatico che ne dovrà garantire la corretta fruizione.

Sulla base di queste informazioni, si potranno definire i requisiti in relazione a tutte le diverse figure sopra elencate e dunque ipotizzare una possibile progettazione del sistema nel suo complesso.

Una volta che il sistema di raccolta ed elaborazione dati sarà entrato in funzione, si procederà con la misurazione della qualità che, sulla base delle dimensioni ritenute rilevanti, prevede la costruzione di apposite metriche di valutazione. Confrontando i valori misurati con quelli di soglie predefinite, potranno essere individuate eventuali problematiche, in relazione alle quali si andrà alla ricerca delle cause radice, importanti da conoscere in un'ottica di definizione di opportune azioni correttive e/o di miglioramento.

Si noti che le dimensioni della qualità dei dati qui considerate non fanno riferimento allo standard SQuaRE dal momento che questo non era ancora stato emesso quando è stata pubblicata la metodologia in esame; tuttavia, molte delle caratteristiche previste rimandano alla norma, tanto che si considerano, tra le altre, l'accuratezza, l'accessibilità, la completezza, la riservatezza, la facilità di manipolazione, etc.

DWQ (Data Warehouse Quality, Jeusfeld et al, 1998)

Si tratta di un metodo sviluppato nell'ambito di un progetto europeo che considera la soggettività del concetto di qualità, ovvero la sua dipendenza dagli obiettivi degli stakeholders di volta in volta considerati.

Rispetto al TDQM, nella fase iniziale vengono qui analizzate non solo le dimensioni di data quality ritenute rilevanti, ma anche le relazioni reciproche tra le stesse e con gli elementi costituenti la base di dati. Ciò consente di indirizzare meglio le attività operative su quegli aspetti ritenuti rilevanti dagli stakeholders: proprio a tale scopo, si prevede l'utilizzo del QFD, con impostazione di una casa della qualità, al fine di priorizzare i requisiti espressi dagli utenti, collegarli alle caratteristiche tecniche della data warehouse e trovare eventuali correlazioni tra quest'ultime.

Come si vedrà nel paragrafo successivo, anche la metodologia di seguito proposta prevede l'uso del QFD, al fine di selezionare le dimensioni di DQ ritenute rilevanti e quindi da considerare nelle successive prove.

Output dell'applicazione della casa della qualità è allora una lista delle dimensioni più importanti per gli stakeholders, a sua volta input per la fase di misurazione vera e propria, dove i valori quantificati verranno confrontati con gli intervalli di valori attesi (requisiti di qualità) e si procederà ad una valutazione delle eventuali dipendenze di tipo statistico tra le diverse dimensioni.

A questo punto, sarà possibile definire quali caratteristiche e quali aree sono più critiche e, proprio per questo, oggetto di opportune azioni correttive.

Le dimensioni qualitative qui considerate sono quelle della ISO 9126, poi sostituita dallo standard SQuaRE: in particolare, vale la pena citare la correttezza, la completezza, la tracciabilità, etc.

TIQM (Total Information Quality Management, English 1999)

Metodologia disegnata per supportare l'integrazione di molteplici fonti dati in un database unico, punta a ridurre, già in fase di costruzione del DB, gli eventuali errori ed eterogeneità presenti.

Dal momento che prevede anche approfondite analisi costi-benefici, consente non solo di raggiungere elevati standard di data quality, ma anche di intraprendere azioni di miglioramento se e solo se queste risultano essere convenienti in termini di costi e benefici.

In tal caso, la fase di misurazione della qualità si colloca nello step di assessment, dal momento che obiettivo della metodologia in esame è quello di andare a valutare la qualità delle diverse fonti integrate: si procede allora con l'estrazione di campioni casuali, sui quali si andranno ad effettuare le misurazioni previste ed applicare le metriche opportunamente definite.

Eseguire le diverse analisi su campioni casuali consente di avere maggiori garanzie circa i risultati ottenuti, dal momento che minimizza l'insorgere di effetti sistematici; proprio per tale ragione, anche nella metodologia di seguito introdotta, si raccomanda di effettuare le varie prove su campioni estratti casualmente dai database in esame.

Le valutazioni dei risultati tecnici dovranno essere in tal caso affiancate da opportune analisi dei costi, proprio per quantificare il costo delle problematiche inerenti la qualità dei dati.

Una volta identificate le diverse tipologie di errori presenti, le relative cause e i costi associati, si potrà procedere alla progettazione di soluzioni mirate al miglioramento della DQ, ad esempio in termini di standardizzazione dei dati, loro correzione, pulizia, matching, trasformazione e consolidamento, e al loro successivo monitoraggio.

Le dimensioni qualitative qui considerate sono molteplici ma tutte inerenti e quindi relative alla qualità dei dati in sé: si parla infatti di completezza, accuratezza, precisione, usabilità e ovviamente, visto l'approccio anche economico che tale metodologia adotta, i costi.

AIMQ (A Methodology for Information Quality Assessment, Lee et al. 2002)

Metodologia che utilizza il benchmarking come tecnica valutativa della qualità, parte da una matrice 2x2 che va sotto il nome di modello PSP/IQ (Product and Service Performance Model for Information Quality), illustrato in **Figura *14***: questa consente di classificare le dimensioni qualitative sulla base della loro importanza per gli utenti e per i manager.

Ad esempio, le dimensioni che consentono di valutare l'affidabilità delle informazioni possono essere la correttezza, la completezza, la rappresentazione concisa e consistente, mentre l'utilità può essere valutata in funzione della quantità, rilevanza, comprensibilità,

interpretabilità e oggettività. L'identificazione di queste dimensioni avviene tramite la somministrazione agli utenti di un primo questionario, a cui ne segue un secondo volto a misurare il giudizio dei soggetti coinvolti in relazioni agli aspetti precedentemente definiti.
Il questionario potrebbe anche richiedere agli utenti di definire un ranking degli aspetti di data quality che richiedono un qualche tipo di miglioramento.

	Conforme alle specifiche	***Soddisfa o supera i requisiti espressi dai consumatori***
Qualità del prodotto	Informazioni affidabili	Informazioni utili
Qualità del servizio	Informazioni attendibili	Informazioni utilizzabili

Figura 14: *Modello PSQ/IQ*

L'assessment che viene allora condotto con tale metodo è prettamente soggettivo. A questo punto si procederà con la comparazione dei risultati ottenuti con il benchmark, tramite l'applicazione di opportune tecniche di gap analysis: a titolo informativo, si citano l'Information Quality Benchmark Gaps (confronto con i valori delle diverse dimensioni in organizzazioni considerabili quali best practises) e la Information Quality Role Gaps (confronto tra assessment eseguiti da soggetti con diversi ruoli organizzativi).
Le dimensioni dei dati considerate sono le stesse previste nel TDQM.

CDQ (Complete Data Quality, Batini e Scannapieco, 2006, Batini et al, 2008)
Tale metodologia è stata concepita per essere allo stesso tempo completa, flessibile e facile da applicare:

- È completa dal momento che si sono considerate ed integrate le tecniche disponibili in letteratura in un framework che possa essere applicato in contesti più o meno ampi e a tutti i tipi di dati (strutturati, semi strutturati o non strutturati).
- È flessibile dal momento che consente all'utente di selezionare le tecniche e gli strumenti più adatti alla fase considerata o al contesto di utilizzo.
- È semplice dal momento che è organizzata in fasi e ognuna di queste è caratterizzata da specifici obiettivi da raggiungere e tecniche da applicare.

Uno degli aspetti particolarmente interessanti di questo metodo è che, a differenza degli altri presentati, non considera come dato il processo di definizione del contesto e dei relativi requisiti. La prima fase è allora proprio quella di sua ricostruzione, in termini di identificazione delle relazioni tra unità organizzative, processi, servizi e dati, ma anche di descrizione del suo funzionamento.
Una volta definiti i target qualitativi sulla base del contesto precedentemente individuato, si andranno a valutare i loro costi e benefici e si procederà con la scelta del piano di miglioramento da implementare: questo sarà costituito dalla sequenza di attività che ha il più alto rapporto costi/benefici.
Le dimensioni qualitative qui considerate fanno riferimento sia allo schema dei dati, in termini di correttezza rispetto al modello, correttezza rispetto ai requisiti, completezza, sia ai dati in sé (i.e. accuratezza semantica/sintattica, completezza, consistenza, etc.).

Accanto a queste metodologie più generali e trasversali, sono stati sviluppati anche delle tecniche ad hoc volte ad effettuare valutazioni sulla qualità dei dati in specifici campi:

- Metodologia CIHI (Canadian Institute for Health Information): uno degli aspetti più interessanti relativi a tale caso è il fatto che il database qui considerato risulta essere particolarmente ampio ed eterogeneo. Come conseguenza di ciò, si è lavorato su un subset di dati e si è cercato di definire delle prassi che consentissero di tenere in considerazione l'eterogeneità. Tali aspetti caratterizzeranno, come vedremo, anche la metodologia definita nella seguente trattazione.
- Metodologia ISTAT (Italian National Bureau of Census): obiettivo principale è quello di garantire un adeguato livello di qualità dei dati che vengono integrati dai molteplici database gestiti dai diversi livelli della pubblica amministrazione.
- Metodo AMEQ (Activity-based Measuring and Evaluating of Product Information Quality), appositamente studiato per effettuare valutazioni della qualità dei database di prodotto utilizzati da realtà manifatturiere.
- Metodo DaQuinCIS (Data Quality in Cooperative Information Systems), specificatamente elaborato per gestire la data quality in sistemi informativi cooperativi.
- Metodo QAFD (Quality Assessment of Financial Data), volto a definire misure qualitative standard per i dati finanziari.

Quelli appena presentati rappresentano dei framework a trecentosessanta gradi tramite i quali eseguire un assessment della DQ, tanto che si concludono generalmente anche con una fase di definizione delle azioni correttive da implementare per migliorare la qualità dei dati e il sistema di gestione degli stessi.
Nonostante anche per la metodologia sviluppata si vada a definire un opportuno framework, il punto di vista che si adotta è più focalizzato sulla sola misurazione della qualità della base di dati, anche perchè non si hanno né le competenze né le possibilità di intervenire sui sistemi informatici che ci sono dietro.
Quello che potremo fare è allora mettere in atto delle prove che consentano in primis di attestare la qualità dei dati e che, sulla base dei risultati ottenuti, permettano in tranquillità di operare la scelta sull'utilizzo o meno del DB in esame; nulla vieta di andare poi ad effetuare una qualche classificazione dei difetti riscontrati e proporre soluzioni che consentano di evitare la loro prolificazione.

Spostandoci verso un approccio più operativo, vengono ora presentati alcuni esempi di tecniche di identificazione degli errori nei database: se ne darà una lettura da un punto di vista funzionale, al fine di evidenziare le dimensioni sulla base delle quali vengono eseguite le operazioni di error detection.

Metodi orizzontale e verticale (Sporleder et al., 2006)
Queste due metodologie automatiche di rilevazione degli errori presenti in database testuali consentono di operare anche in assenza di conoscenza di base (i.e. conoscenza del dominio o della struttura del database) e si basano sugli stessi dati contenuti nei DB per individuare incongruenze ed errori.
L'identificazione degli errori orizzontali si fonda sul fatto che i diversi campi di un database non sono solitamente statisticamente indipendenti: avendo allora a disposizione un considerevole insieme di dati, è possibile determinare queste relazioni e utilizzarle al fine di identificare possibili campi anomali.
In Sporleder et al., si sono però utilizzati strumenti di machine learning che risultano essere in grado di prevedere il valore di una cella, dati i valori di altri campi per la medesima istanza. Se il valore previsto differisce dal valore presente nella cella, verrà segnalato come un potenziale errore.
Dunque, mentre il metodo orizzontale mira a identificare i valori che non sono coerenti con i campi rimanenti di un record del database, l'identificazione degli errori verticali vuole

evidenziare stringhe di testo immesse in un campo errato. Errori di questo tipo sono abbastanza frequenti: così come possono essere accidentali, possono anche derivare da modifiche nella struttura del database stesso. In tal caso, si andrà a valutare se, dato il contenuto di una cella, questo risulta essere memorizzato nel campo in cui è più probabile che si trovi. Dunque, quelle stringhe di testo classificate come appartenenti a una colonna diversa da quella in cui si trovano attualmente rappresentano un potenziale errore.
Sebbene il controllo di questo tipo di difetti sia molto più rapido rispetto a quello eseguito nel metodo orizzontale, a volte è difficile stabilire se un errore contrassegnato sia un errore reale, soprattutto quando non è ovvio a quale colonna appartenga una stringa e ci sono campi molto simili.
Infine, obiettivo delle due metodologie non è solo quello di procedere all'identificazione degli errori, ma anche effettuare una loro correzione dal momento che si presuppone l'accesso sia in lettura che in scrittura ai dati contenuti nei database.

Nella **Tabella *2.2***.vengono brevemente presentate altre tecniche operative utilizzate per l'identificazione di errori nei database.

Tabella 2.2: *Tecniche di identificazione degli errori nei database*

METODO	***DESCRIZIONE***
APPROCCIO BASATO SULLE DISTRIBUZIONI (Hawkins, 1980)	I dati vengono modellati con apposite distribuzioni di probabilità e tutte le istanze che si discostano troppo saranno contrassegnate come possibili outlier e quindi errate.
METODI DI CLUSTERING	Si applica un algoritmo di clustering ai dati e le istanze che non possono essere raggruppate in nessun cluster vengono considerate anomale. Altro approccio prevede invece di considerare direttamente come anomali e quindi difettosi quei cluster che contengono un ridotto numero di istanze.

METODI BASATI SULLA DISTANZA (Knorr et al., 1998)	Si utilizza un metodo basato sul k-nearest neighbour approach, dove gli outliers sono definiti come quelle istanze la cui distanza dal loro vicino più prossimo supera una certa soglia.
METODI BASATI SU REGOLE DI ASSOCIAZIONE (Marcus et al., 2000)	Dopo aver definito delle regole di associazione per i dati, vengono considerati errati quei record che non sono conformi a nessuna regola.

2.3 La nuova metodologia di assessment della qualità

È ora arrivato il momento di presentare la nuova metodologia di assessment che questa trattazione si propone di definire: come già accennato precedentemente, il focus è sulle fasi di misurazione delle qualità da parte degli utilizzatori dei dati e di successiva valutazione circa l'accettazione o meno del database che li contiene.
Tuttavia, è evidente come non si possa procedere con questi task se non si è preliminarmente eseguita un'analisi approfondita del DB e non si sono valutate le dimensioni di qualità che occorre verificare, in virtù dei requisiti espressi dai diversi stakeholders coinvolti.
Si procederà ora ad una presentazione del metodo di lavoro nei termini più generali possibili, per facilitare e garantire poi la sua applicazione anche ad altri contesti di utilizzo; tuttavia, si noti che la sua progettazione e definizione è avvenuta tenendo in considerazione i database bibliografici, oggetto di studio della seguente trattazione.
Prima di procedere alla descrizione di dettaglio della metodologia di assessment, occorre fare alcune assunzioni:

- Il focus della metodologia sono i dati che i DB memorizzano e non i sistemi informatici che ne garantiscono la storicizzazione, fruizione, gestione, etc;
- I database analizzabili sono di tipo relazionale;
- I dati memorizzati nei database considerati risultano essere strutturati.

Nello specifico, le fasi in cui si articola il metodo sono le seguenti:

1. **Raccolta dei requisiti**

   Prima di condurre una qualsiasi attività di analisi su un certo prodotto, occorre studiarlo a fondo al fine di approfondirne gli aspetti distintivi e definire, sulla base degli stessi, quali requisiti deve soddisfare nell'ottica dei diversi stakeholders coinvolti.

   Punto di partenza per la definizione dei customer requirements sono i bisogni, ovvero ciò che spinge un certo cliente ad acquistare un determinato prodotto/servizio: generalmente, al cliente non interessa come il bisogno verrà soddisfatto e proprio per questo i bisogni sono generici e completamente indipendenti dalle soluzioni tecniche che verranno selezionate dai progettisti, mentre i requisiti sono una loro trasposizione in termini tecnici e quantitativi.

La fase di definizione dei requisiti, fondamentale per una corretta gestione della qualità, può essere eseguita con l'ausilio di diverse tecniche, più o meno strutturate. Tra queste, vale la pena citare:

- Questionari: una delle forme più semplici per acquisire dei feedback da parte dei soggetti coinvolti, sono di fatto sondaggi composti da una serie di domande predefinite in diversi formati (ad es. caselle di testo o scelta multipla). Sono utili per valutare e monitorare i bisogni secondari e quindi espliciti dei clienti; a partire dagli stessi, saranno poi definiti i bisogni primari, ovvero quelli taciti, latenti, sui quali si gioca realmente la soddisfazione della clientela.
- Intervista personale: tale metodo prevede che vengano eseguiti dei colloqui personali con i singoli soggetti appartenenti al campione di popolazione.
- Focus group: si intendono gruppi, di dimensione variabile, costituiti da soggetti estratti casualmente e considerati quali campioni rappresentativi della popolazione da intervistare.
- Tecniche qualitative strutturate: si basano sul mettere a confronto prodotti/servizi della stessa tipologia, ma realizzati da produttori distinti. Si andrà allora a chiedere ai clienti intervistati di definire un ranking degli stessi.
- Tecniche di analisi di prodotto: al cliente, messo davanti al prodotto, viene chiesto quali sono le caratteristiche per cui sceglierebbe un determinato prodotto e quelle per cui non lo farebbe.
- Esperienza personale del gruppo di lavoro: quest'ultimo si immedesima nel mercato al fine di estrarre i possibili bisogni della clientela. È sicuramente il metodo meno costoso, dove il gruppo di lavoro fa sia da intervistato che da intervistatore.

I requisiti che si andranno ad estrarre in relazione ad un generico database possono essere di vario tipo (**Figura *15***), a seconda che si vada a valutare:

- Qualità in fase di utilizzo (QIURs): vengono qui considerati i bisogni dei diversi stakeholders che vanno ad utilizzare il DB in un dato contesto;
- Qualità del prodotto software (PQRs): tali requisiti sono di natura più tecnica e fanno riferimento ad attributi informatici che il prodotto ICT deve soddisfare, ad esempio in termini di architettura e tipologia del database in esame. Sono particolarmente importanti in fase di sviluppo e manutenzione dei DB e molto

spesso derivano dall'identificazione delle funzionalità che gli utenti desiderano avere nel prodotto in esame (QIURs).

- Qualità dei dati (DQRs): come indica il nome stesso, sono requisiti relativi ai dati associati al prodotto in esame e vanno utilizzati, come nel nostro caso, per effettuare delle valutazioni in termini di qualità dei dati. Sono proprio i DQRs i requisiti sui quali dovremo concentrare la nostra attenzione.

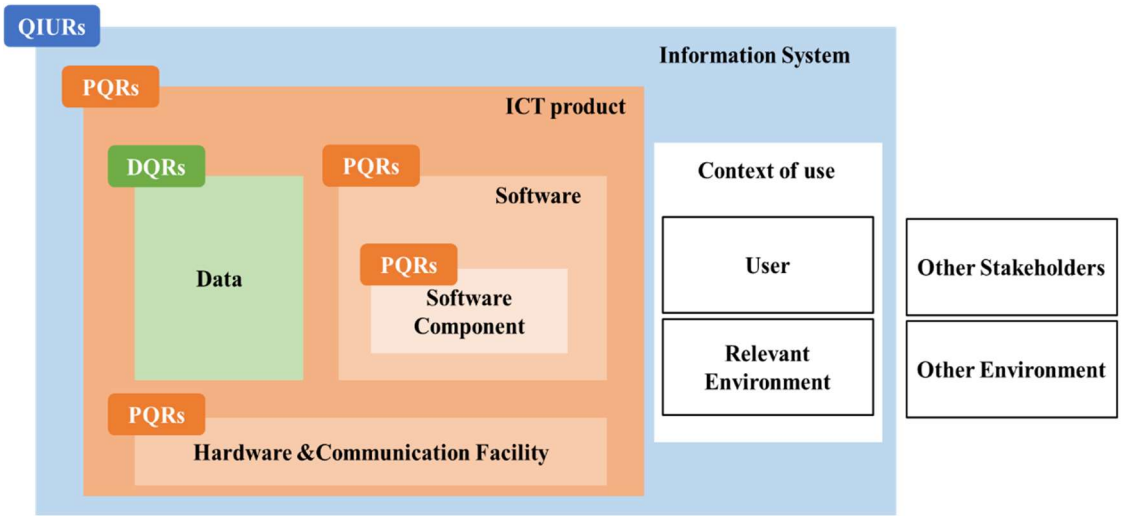


Figura 15: *Scope dei requisiti di qualità (ISO/IEC 25030:2019)*

2. Identificazione delle dimensioni SQuaRE da valutare

Il passaggio da requisiti a caratteristiche tecniche del database in esame avviene con l'utilizzo della già citata tecnica del QFD (Quality Function Deployment).

Strumento a supporto della progettazione di prodotti/servizi, consente di definire in modo univoco quali sono le caratteristiche tecniche del prodotto che risponderanno al meglio alle reali esigenze del cliente.

Nella fase di progettazione, la matrice da utilizzare è la PRODUCT PLANNING MATRIX che consente di tradurre i bisogni della clientela in caratteristiche tecniche. Qualora si vogliano poi dettagliare anche le caratteristiche tecniche dei componenti del prodotto e/o del processo, si potranno implementare anche le altre matrici previste della metodologia.

La struttura della PPM è illustrata in **Figura *16***.

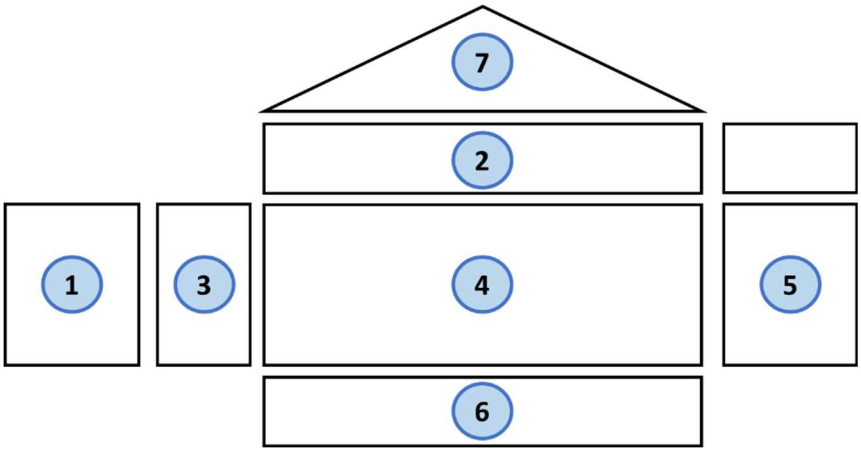


Figura 16: *Struttura della Product Planning Matrix (QFD)*

1. CUSTOMER REQUIREMENTS: il primo blocco sulla sinistra è un vettore colonna che avrà tante righe quanti sono i diversi requisiti.
   Il singolo elemento è indicato con CR_i con i che varia da 1 a n (numero totale di bisogni individuati).
2. PRODUCT ENGINEERING/DESIGN REQUIREMENTS: si tratta di un vettore che memorizza le diverse caratteristiche tecniche, indicate come EC_j con j che varia da 1 a m, numero totale delle stesse specifiche.
   Nel nostro caso, le caratteristiche tecniche da considerare sono le dimensioni previste dallo standard SQuaRE: questo opera una distinzione tra caratteristiche intrinseche, dipendenti dal sistema e caratteristiche che invece possono essere classificate sotto entrambi i punti di vista indicati.
   Volendoci concentrare sulla qualità dei dati in sé e non tanto sul sistema che li contiene e gestisce, il focus risulterà essere sulle caratteristiche intrinseche.
3. REQUIREMENTS DEGREE OF IMPORTANCE: il secondo vettore colonna avrà tante righe quanti sono i bisogni e riporterà le importanze assegnate ad ogni bisogno.
4. RELATIONSHIP MATRIX: la matrice centrale ha n righe e m colonne e definisce le relazioni tra i Customer Requirements e le Engineering Characteristics.
   Il contenuto della generica cella è indicato con il simbolo R_{ij}, coefficiente di intensità di correlazione tra il bisogno i-esimo e la caratteristica tecnica j-esima.

5. COMPETITIVE BENCHMARK ASSESSMENT: tale sezione consente di effettuare la pianificazione e il deployment della qualità attesa, attraverso la gerarchizzazione delle richieste del cliente e l'analisi della concorrenza.
6. TECHNICAL IMPORTANCE RANKING: risultato di elaborazioni eseguite a partire dai dati sopra indicati, permette di gerarchizzare le caratteristiche tecniche e quindi definire quelle che sono più importanti in un'ottica di soddisfacimento dei bisogni della clientela. In aggiunta, potrebbe anche essere sottoposto un questionario ai soggetti coinvolti, nel quale si chieda loro di valutare, su un'opportuna scala, l'importanza delle diverse dimensioni.
7. CORRELATION MATRIX: la matrice triangolare mette in correlazione le caratteristiche tecniche con loro stesse e quindi evidenzia la possibile presenza di legami reciproci.

Dunque, al termine di tale fase, si saranno definite le dimensioni che la metodologia dovrà andare prima a testare e poi a valutare.

I criteri sulla base dei quale selezionare le dimensioni possono essere diversi e dunque, dovranno essere valutati di caso in caso: ad esempio, si potrebbero selezionare le tre caratteristiche con importanze relative maggiori oppure quelle che presentano un valore per la stessa importanza superiore al 10%.

3. Definizione prove da condurre per testare le dimensioni selezionate

La definizione delle metriche e delle relative prove di misura da mettere in atto per quantificarle è un processo che deve prendere in esame lo standard SQuaRE e le indicazioni che lo stesso fornisce.

Dal momento che la metodologia presentata si concentra sui dati in sé, di seguito si elencano le caratteristiche intrinseche, le corrispondenti metriche e le prove da eseguire per arrivare ad una loro quantificazione (**Tabella *2.3***).

Si noti che gli indicatori introdotti sono definiti in modo da mettere in evidenza la difettosità dei dati e quindi il non soddisfacimento dei requisiti di qualità previsti dallo standard.

Tabella 2.3: *Caratteristiche intrinseche, con indicazione di metriche e prove corrispondenti*

DIMENSIONE	***METRICA***	***PROVA***
ACCURATEZZA	#record con contenuto non accurato sintatticamente e semanticamente	Si confronta il DB in esame con un riferimento assoluto appositamente costruito e si contano i record per i quali sono presenti campi caratterizzati da una non corrispondenza nel contenuto.
COMPLETEZZA	#record con celle vuote	Si conta il numero di record che presentano celle vuote, ovvero dove ci si aspettava un valore in virtù del campo di appartenenza.
COERENZA	#record con chiave primaria non univoca #record non coerenti internamente #record non coerenti esternamente	Si valuta la coerenza di chiave, verificando che non ci siano record con il medesimo valore in corrispondenza della chiave primaria, la coerenza interna, verificando che le regole circa i legami esistenti tra i diversi campi siano rispettati, e la coerenza esterna, valutando l'uniformità del tipo di dato all'interno di uno stesso campo.
CREDIBILITA'	#record con valori non credibili	Si quantifica il numero di record che presentano valori considerati non credibili perché non certificati/validati da un ente riconosciuto. Se il database nel suo complesso è stato certificato, il requisito di credibilità è automaticamente soddisfatto e il numero di record non credibili viene posto pari a 0.
ATTUALITA'	#record non aggiornati alla frequenza richiesta	Conoscendo la frequenza con cui devono essere aggiornati i valori memorizzati nel DB e la data di ultima modifica, si

		individuano e contano i record che non soddisfano il requisito di attualità.
ACCESSIBILITA'	#record non accessibili	Tale metrica viene misurata andando a contare il numero di record non considerati accessibili dagli utenti interessati, in un determinato contesto di utilizzo.
CONFORMITA'	#record non conformi	Tale requisito andrà valutato solo qualora i dati debbano essere conformi a standard, convenzioni, normative in vigore o regole simili relative alla qualità dei dati. In tal caso, si andranno a contare il numero di record che presentano valori non conformi.
RISERVATEZZA	#record non correttamente criptati e decriptati	Tale test, da applicare ai soli dati che devono soddisfare requisiti di riservatezza, prevede l'individuazione e il conteggio dei record per i quali le operazioni di criptografia non sono state eseguite correttamente e/o non sono andate a buon fine.
EFFICIENZA	#record memorizzati in un formato non considerato efficiente	La prova mira ad individuare quei record che contengono valori memorizzati in formati non efficienti in relazione all'informazione da storicizzare. La qualificazione di un formato come efficiente può derivare da standard internazionali o nazionali o anche da test appositamente condotti.
PRECISIONE	#record che contengono celle che	Si tratta di individuare i dati che non rispettano i requisiti di precisione

	non hanno la precisione richiesta	richiesta (ad esempio, in termini di numero di cifre decimali previste dai diversi campi) e di segnalare e contare i record che li contengono.
TRACCIABILITA'	#record per i quali non esistono campi che consentano di tracciare accessi e modifiche	Si individuano tutti quei record per i quali non esistono specifici campi che consentano di tener traccia di modifiche ed accessi eseguiti.
COMPRENSIBILITA'	#record che presentano celle con valori rappresentati da simboli non noti	Si tratta di verificare che le informazioni storicizzate nel database siano rappresentate e memorizzate con una simbologia comune; qualora ciò non avvenga, si andranno a conteggiare un non soddisfacimento del requisito relativo alla comprensibilità.

4. Progettazione del piano di campionamento

Una volta che si sono selezionate le dimensioni di interesse, definite le relative metriche e progettate le prove da condurre per quantificarle, si può identificare esattamente cosa dovrà essere o meno considerato come difetto nel database in esame.

Questo consentirà di costruire concretamente il piano di campionamento per attributi, che andrà a considerare come difettosi tutti quei record che contengono uno o più difetti.

Dunque, il numero di accettazione c sarà riferito ai record difettosi, ognuno dei quali potrebbe non rispettare i requisiti relativi a una o più dimensioni qualitative identificate.

Abbiamo visto che i piani possono essere sia progettati ad hoc sia definiti in accordo a norme prestabilite: se nel primo caso si potrà costruire una procedura custom in relazione al caso specifico considerato, nel caso di utilizzo degli standard si avrà la garanzia di operare secondo modalità comunemente riconosciute ed accreditate. Per

mostrare le potenzialità dei due metodi, si procederà, nel **CAPITOLO III**, all'applicazione di entrambi.

A prescindere dalla strada che si sceglierà di seguire, occorre definire una modalità di settaggio dei diversi parametri in input ai piani di campionamento: si potrebbero utilizzare dei questionari nei quali si chiede ai diversi stakeholders coinvolti di esprimere le loro valutazioni circa le variabili e quindi individuare dei valori riassuntivi oppure definire, sulla base dell'esperienza, diverse ipotesi di lavoro che cerchino il più possibile di tenere in considerazione i diversi scenari ed i diversi punti di vista.

Altro aspetto molto importante per la corretta costruzione di un piano di campionamento è estrarre casualmente i record che costituiranno il campione di analisi: ciò consente di evitare l'introduzione di effetti sistematici e fa sì che i record analizzati siano rappresentativi del lotto e portino potenzialmente in loro tutte le caratteristiche e le proprietà dell'insieme da cui derivano.

5. Esecuzione del piano e delle prove

Le modalità di esecuzione possono essere più o meno automatizzate a seconda di come si andranno ad implementare le prove di testing delle dimensioni di interesse ed il piano di campionamento prescelto.

Durante questa fase operativa, importante è documentare le eventuali assunzioni e scelte fatte, al fine di dare un quadro il più possibile esaustivo e comprensibile a chi accederà alle analisi e ai relativi risultati.

Una volta eseguite le prove ed individuati i diversi tipi di difetti, emergeranno quali sono i record difettosi, ovvero quelli che hanno presentato difetti in relazione ad una o più dimensioni.

A questo punto, si arriverà alla scelta circa l'accettazione o meno del database: questa sarà effettuata semplicemente confrontando il numero di record difettosi individuati nel campione di interesse con il numero di accettazione previsto dallo specifico piano di campionamento.

6. Definizione di eventuali azioni correttive e loro implementazione

A prescindere dall'esito del controllo, ma soprattutto qualora si sia rifiutata la fornitura, è importante eseguire un'attività di identificazione degli errori più ricorrenti

e delle possibili azioni correttive/ di miglioramento da implementare per aumentare la qualità della base di dati.

Questo è importante tanto nell'ottica del fornitore quanto in quella del cliente, dal momento che consentirà di avere, sia per il presente che a maggior ragione per il futuro, una maggiore fiducia nelle informazioni storicizzate nel database.

Per concludere ed avere un quadro più chiaro circa la metodologia appena presentata e riassunta in **Figura *17***, si illustrano i principali punti di forza e debolezza.

PUNTI DI FORZA

- Accettazione/rifiuto sulla base di criteri noti e definiti a priori;
- Applicabilità da parte di cliente e fornitore; quest'ultimo può quindi anticipatamente eseguire anche lui la prova, provare a prevedere la decisione che prenderà il cliente e mettere in atto azioni correttive che consentano di eliminare gli errori dal momento che lui, a differenza del cliente, ha accesso ai dati anche in scrittura;
- Metodologia basata su normative standard e comunemente riconosciute, sia per quanto riguarda la serie SQuaRE, che definisce le dimensioni che chiunque vuole valutare la qualità di un software dovrebbe considerare, che per la MIL STD 105E, per l'impostazione dei piani di campionamento.

PUNTI DI DEBOLEZZA/ LIMITI NELL'APPLICABILITÀ

- Applicabile solo a database relazionali e/o dati strutturati;
- Test di valutazione della qualità dei dati difficilmente automatizzabili, con conseguente necessità dell'intervento umano. Ciò implica che la metodologia richieda tempi di esecuzione particolarmente lunghi, soprattutto laddove il lotto di partenza abbia dimensioni particolarmente elevate, oppure perché il settaggio dei parametri richiesti in input al piano di campionamento si traduca nella selezione di campioni particolarmente numerosi;
- Molteplici parametri da settare per la corretta implementazione del piano di campionamento;
- Necessità di interfacciarsi con gli stakeholders per la definizione delle dimensioni qualitative da testare e dei parametri in input al piano di campionamento;

- Elevata influenza dello use case specifico che si sceglie di esaminare, soprattutto qualora si estragga il campione di analisi da una determinata selezione di record del DB e non dalla base dati nella sua interezza.

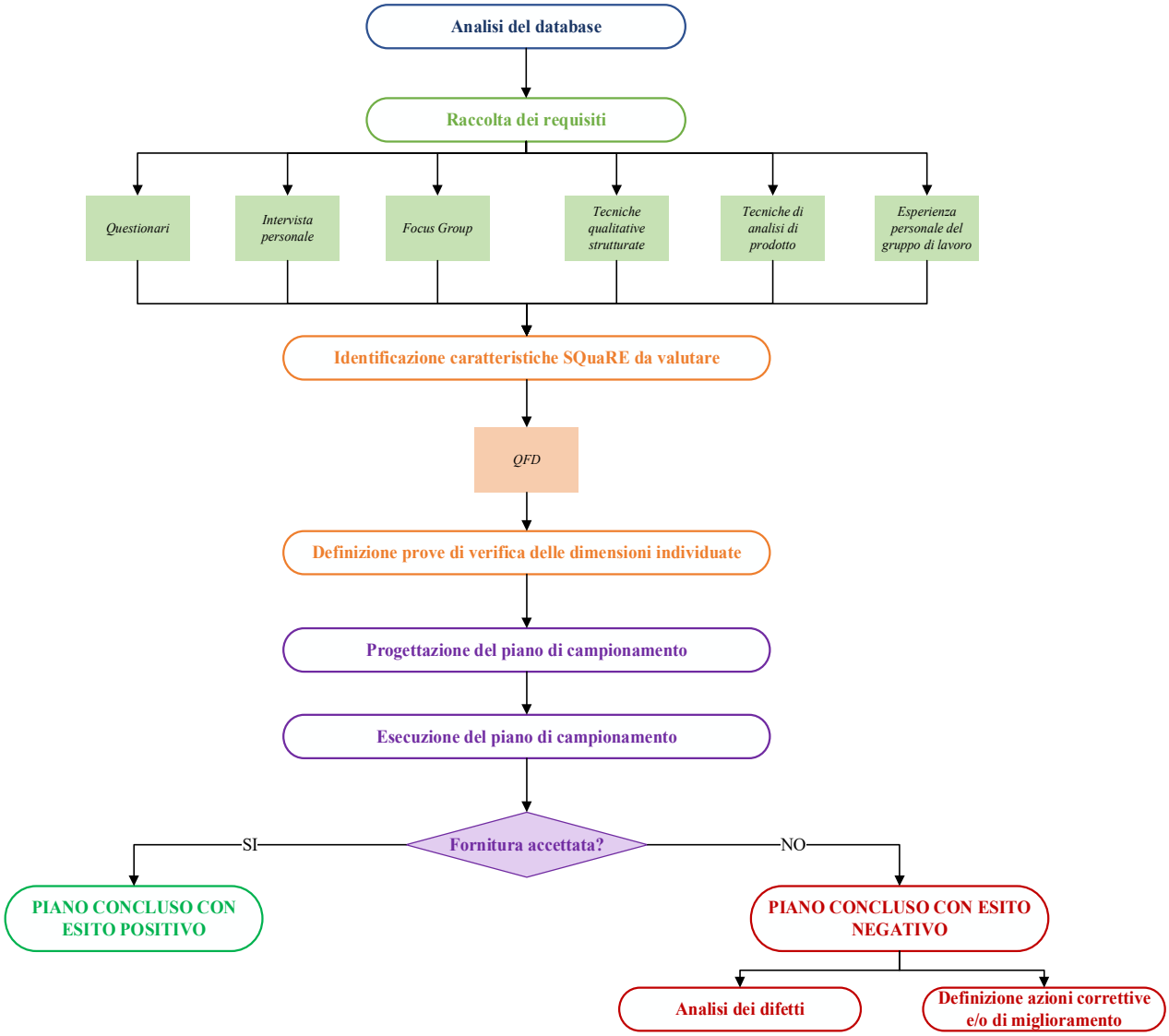


Figura 17: *Flusso della metodologia di assessment della qualità dei database*

3. CAPITOLO III
IL CASO DI STUDIO: I DATABASE BIBLIOMETRICI

3.1 I database di tipo bibliometrico

I database ai quali si applicherà la metodologia appena presentata sono database di tipo bibliometrico, ovvero basi di dati comunemente utilizzate per:

- la ricerca di documenti scientifici;
- l'esecuzione di analisi bibliometriche, grazie agli indicatori messi a disposizione;
- la selezione di giornali scientifici sui quali pubblicare un certo articolo (Franceschini et al., 2016).

Ne consegue che tra i fruitori degli stessi DB compaiono tanto i ricercatori universitari, quanto gli studenti, gli enti pubblici che si occupano delle attività valutative e chiunque voglia ottenere informazioni su una qualche pubblicazione per le sue attività operative: la qualità e affidabilità di tali basi di dati diventano allora due aspetti molto importanti dal momento che da questi scaturirà l'esito delle attività specifiche che ciascun utente vorrà condurre.

I database bibliometrici sui quali andremo a focalizzare la nostra attenzione sono Web of Science e Scopus, nonostante venga spesso classificato come tale anche Google Scholar che tuttavia è più propriamente da considerare come una via di mezzo tra un database bibliometrico e un motore di ricerca.

Nonostante contenga un numero più alto di documenti scientifici (**Tabella 3.1**), la sua inferiorità rispetto a WoS e Scopus è ampiamente riconosciuta dal momento che indicizza le pubblicazioni e le citazioni in modo totalmente automatico attraverso appositi web crawler, con molteplici e particolarmente evidenti errori (Franceschini et al., 2016).

Per tale ragione, si è scelto di non includere anche Google Scholar tra i DB ai quali viene applicata la metodologia presentata.

Si noti poi che tutti e tre i DB sono caratterizzati dall'essere multidisciplinari, dal momento che indicizzano pubblicazioni relative a molteplici discipline: tale aspetto aumenta la complessità degli strumenti visto che dovranno avere la capacità di gestire l'eterogeneità delle fonti di acquisizione dati.

Ciò spiega ad esempio perché esistano pochissimi database multidisciplinari e una vasta gamma di database bibliometrici disciplinari.

Tabella 3.1: *Numero di documenti nei database bibliometrici multidisciplinari (Mongeon et al., 2016; Orduna-Malea et al., 2015)*

DATABASE	*NUMERO DI DOCUMENTI*
WEB OF SCIENCE	10 M
SCOPUS	13 M
GOOGLE SCHOLAR	160 M

A far considerare WoS e Scopus come più affidabili e sicuri è anche il fatto che la loro consultazione avviene dietro apposita sottoscrizione a pagamento, mentre Google Scholar è gratuito: questo non deve però far assolutamente pensare che i primi due database siano esenti da errori ma, al contrario, studi precedenti e anche questa stessa trattazione mostrano come i due DB presentino diversi difetti, con conseguenti record difettosi e informazioni non corrette.

In fase di presentazione della metodologia di assessment della qualità dei dati contenuti nei database, si è specificato come fosse necessario, preliminarmente rispetto a qualsiasi task, procedere con uno studio approfondito della struttura e delle caratteristiche distintive del database in esame: in quest'ottica, si inserisce di seguito una descrizione dei due DB di interesse, Scopus e Web of Science, cercando di metterne in evidenza le funzionalità e le peculiarità in termini di storicizzazione, fruizione e gestione dei dati.

WEB OF SCIENCE

Web of Sciences è, come già specificato, una banca dati bibliografica citazionale e multidisciplinare online attiva dal 2002; di proprietà prima della *Thomson-Reuters*, poi di *Clarivate Analytics*, contiene articoli full-text, recensioni, editoriali, cronologie, abstract, atti (riviste e libri) e relazioni tecniche relativi a scienze esatte ed applicate, scienze sociali e materie umanistiche.

La sua core collection è a sua volta costituita da sei database, le cui dimensioni e coperture variano da caso a caso: questo lo rende sicuramente uno strumento di ricerca unificante che consente all'utente di acquisire, analizzare e diffondere le informazioni in modo tempestivo.

Prima di analizzare nel dettaglio il suo funzionamento, si noti che la selezione dei contenuti accettabili per il portale avviene secondo un processo di valutazione formalizzato e basato su ben precisi criteri (i.e. impatto, influenza, tempestività, revisione tra pari e rappresentazione geografica).

Ciò ovviamente implica che le informazioni rese disponibili siano state, in fase di caricamento, verificate e filtrate, con conseguenti garanzie circa la loro affidabilità.
Sempre in ottica di qualità dei dati storicizzati, si noti che il sito rende possibile segnalare la presenza di eventuali campi errati e quindi richiederne una correzione, secondo un processo di miglioramento continuo.

La schermata iniziale che viene presentata all'utente è mostrata in **Figura *18*** e consente di:

- Selezionare se ricercare documenti o ricercatori
  Nel caso di documenti:
- Selezionare il database in cui si vuole effettuare la ricerca
- Selezionare se ricercare documenti, referenze o formule chimiche
- Selezionare il campo di ricerca desiderato; si noti che è possibile anche cercare in tutti i campi previsti dal database (selezionando "*All Fields*"), nonché aggiungere righe per ricercare contemporaneamente più campi ("*Add Row*") e filtri ai dati ("*Add Data Range*").

Dal pannello di ricerca avanzata, infine, è possibile costruire query ad hoc per eseguire la ricerca desiderata, nonché salvarle per avere degli automatismi da sfruttare in futuro.

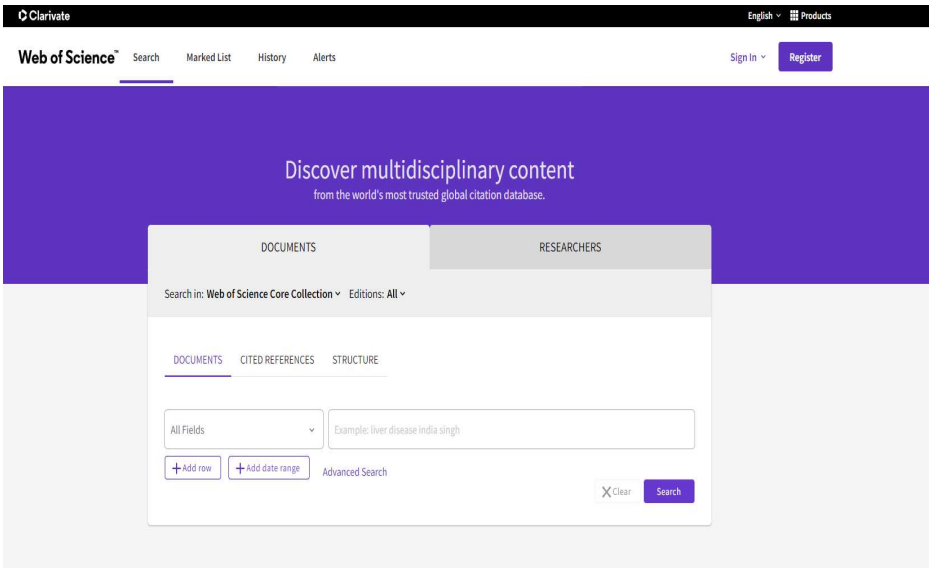


Figura 18: *Schermata iniziale di Web of Science*

Una volta eseguita la ricerca, saranno visualizzati i risultati (**Figura *19***), con la possibilità di analizzarli e creare un alert che avvisi in automatico dell'introduzione di nuovi articoli.

Qualora il numero di risultati sia inferiore a diecimila (10.000), è possibile ottenere un report che mostra la distribuzione delle citazioni e delle pubblicazioni nel corso del tempo.
Non solo è possibile filtrare i risultati in base a rilevanza, data di pubblicazione, numero di citazioni, ordine alfabetico, etc, ma anche raffinare la selezione di record ottenuta con le molteplici opzioni previste nel pannello a sinistra.
Particolarmente di interesse per la nostra applicazione, è il pulsante Export che consente di esportare i risultati e i campi di interesse in diversi formati, tra cui Excel: avremo così a disposizione il dataset in un formato facilmente gestibile ai fini delle attività di analisi da condurre.

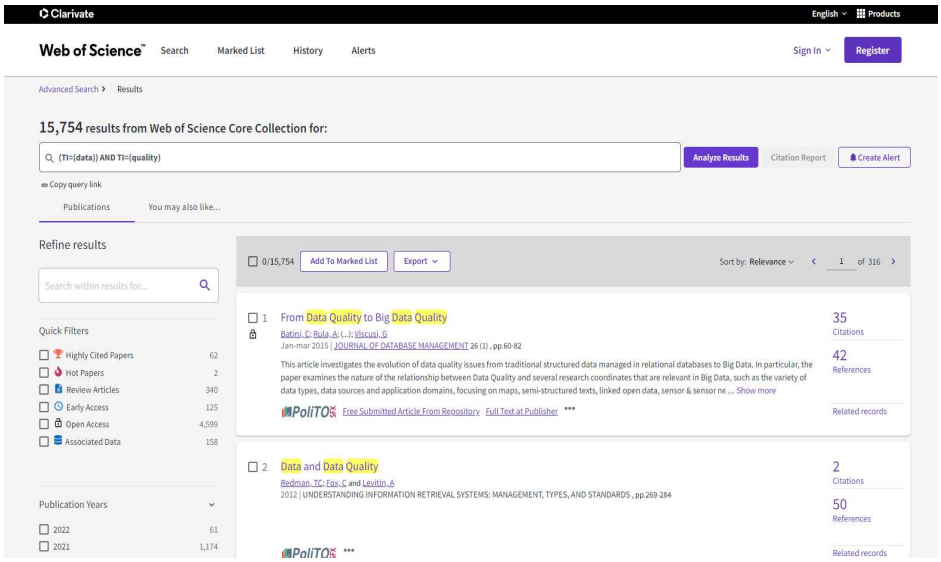


Figura 19: *Schermata di visualizzazione dei risultati della ricerca in Web of Science*

Tra i campi che vengono mostrati nella schermata in ***Figura 19***, abbiamo:

- Titolo: in vista delle future analisi, si noti che in WoS i titoli delle pubblicazioni in lingua straniera sono tradotti in inglese e quindi non possono essere trovati tramite ricerche nella lingua originale;
- Autori, indicati con cognome e iniziali del nome;
- Anno e rivista di pubblicazione, in relazione alla quale possono essere specificati il volume, l'issue e le pagine in cui l'articolo risulta essere pubblicato;
- Abstract;
- Numero di citazioni: numero di volte in cui una certa pubblicazione risulta essere stata citata da altri documenti;
- Numero di referenze: numero di documenti citati nella pubblicazione in questione.

Qualora si vogliano ottenere più informazioni, è sempre possibile cliccare sul titolo del record di interesse e quindi consultare la pagina di dettaglio.

SCOPUS

Sito Web creato dalla casa editrice Elsevier nel 2004, Scopus comprende riviste, riviste di settore, atti di convegni, pubblicazioni commerciali, serie di libri, pagine web e registrazioni di brevetti, relativi alle discipline di scienza, tecnica, medicina, scienze sociali e scienze umane, per un totale di 240 campi.

I dati citazionali partono dal 1996, ma la copertura è dal 1823: l'aggiornamento viene fatto periodicamente al fine di fornire agli utenti dati sempre attuali.

Le fonti da cui Scopus acquisisce i suoi contenuti sono più di settemila e vengono costantemente controllate e selezionate da un consiglio indipendente, il Content Selection and Advisory Board: questo è un gruppo multidisciplinare di scienziati e ricercatori con esperienza in ambito editoriale che, ogni anno, valuta non solo la qualità dei nuovi record da indicizzare, ma anche il mantenimento di un adeguato livello di soddisfazione qualitativa in relazione ai record già esistenti.

Si noti ad esempio che mediamente ogni anno sono tremilacinquecento i record per i quali viene richiesta l'inclusione in Scopus; di questi, solo il 33% soddisfa i criteri tecnici previsti dal DB e del gruppo conforme generalmente solo il 50% viene accettato dal comitato.

Ciò è sicuramente una garanzia del fatto che la qualità dei contenuti indicizzati possa raggiungere standard particolarmente elevati.

Interessante è poi notare come i servizi bibliometrici messi a disposizione dalla piattaforma siano stati utilizzati nella prima valutazione della qualità della ricerca pubblica italiana, riferita al periodo 2004-2010, da parte dell'agenzia ANVUR; questo nonostante il database presenti comunque alcune problematiche, quale ad esempio la presenza di riviste dal dubbio valore scientifico.

La schermata iniziale di ricerca presenta un'impostazione molto simile a quella vista per Web of Science, con la differenza che in tal caso è possibile effettuare la ricerca per documenti, autori o affiliazione.

Il pannello principale è quello di ricerca per documenti, dove è possibile selezionare in quale campo cercare il valore di interesse, aggiungere altri campi di ricerca o filtri e passare alla schermata di ricerca avanzata.

Il risultato della ricerca viene presentato in una pagina simile a quella vista per il primo database, dal momento che viene sempre offerta la possibilità di affinare la ricerca sulla base di diversi parametri, nonché di analizzare i risultati ottenuti in virtù dei criteri di interesse.

In corrispondenza di ciascun record, oltre alle informazioni caratterizzanti lo stesso e precedentemente indicate, si trova un pulsante che consente di aprire la pubblicazione direttamente dal sito web dell'editore.

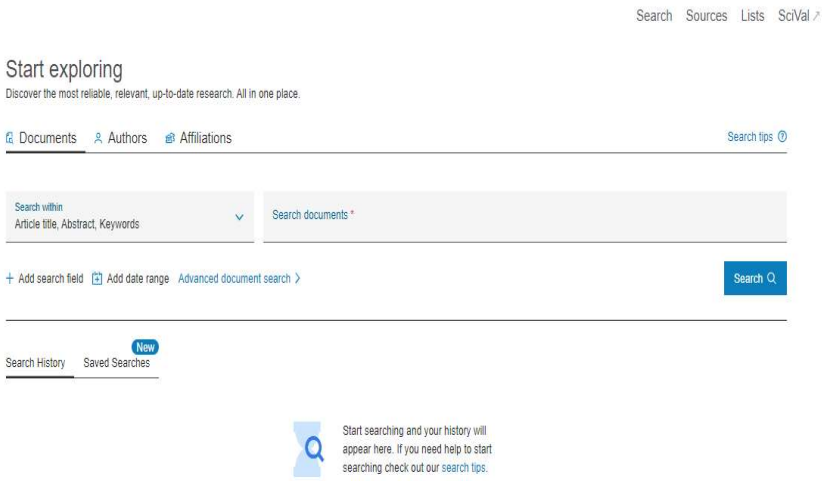


Figura 20: *Schermata iniziale di Scopus*

Cliccando sul titolo, si accederà alla pagina di Scopus contenente le informazioni di dettaglio sulla pubblicazione, mentre grazie alla funzione Export si potranno esportare i dati di interesse nel formato desiderato; nel nostro caso, si utilizzeranno file csv, poi convertiti in Excel.

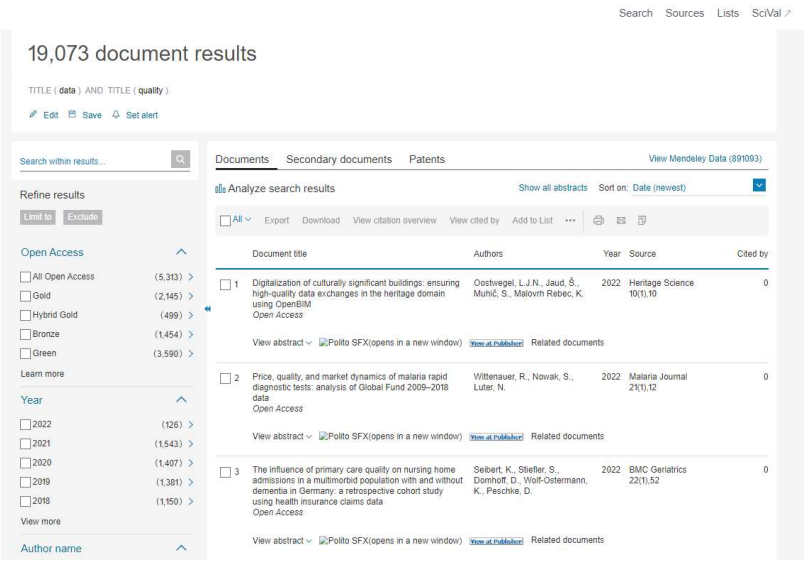


Figura 21: *Schermata di visualizzazione dei risultati della ricerca in Scopus*

Per concludere, si procederà ora ad un confronto tra i due database, anche sulla base della letteratura disponibile in materia.
Secondo Goodman e Deis (2005), i due database sono complementari e l'acquisto dell'uno o dell'altro è di fatto indifferente.
In realtà, ci sono alcune differenze che vale la pena sottolineare (Salisbury, 2009):

- Scopus copre una gamma più ampia di materiale rispetto a WoS, ma quest'ultimo risulta essere più completo. Inoltre, risulta anche che generalmente WoS storicizza articoli di più recente pubblicazione.
  Quindi, in termini di copertura e di attualità delle informazioni presenti, WoS risulta essere migliore di Scopus.
- Per quanto concerne il numero di referenze citate, in Scopus sono maggiori che in WoS in relazione al campione considerato; ciò è in parte dovuto al fatto che in Scopus sono presenti molteplici tipologie di materiale.
- In termini di facilità di utilizzo, entrambi i database raggiungono buoni standard, dal momento che hanno funzionalità di ricerca di base e avanzate ben organizzate, così come offrono agli utenti la possibilità di raffinare via via le loro ricerche.
  Anche l'analisi dei risultati è un'operazione agevole in entrambe le piattaforme.
- Entrambi i database possono fornire dati utili per attività di valutazione/costruzione di dataset ad hoc, ma dovrebbero rendere più agevoli le operazioni di export.

Ne consegue che vengano attribuiti su una scala da 1 a 5 i seguenti punteggi (**Tabella *3.2***):

Tabella 3.2: *Punteggi ottenuti da WoS e Scopus nell'analisi comparativa (Salisbury, 2009)*

	WOS	***SCOPUS***
COPERTURA E ATTUALITA'	5	4
NUMERO DI REFERENZE	4	4
FACILITA' DI UTILIZZO	5	5
ATTIVITA' OFFLINE	3	3
MEDIA	4 1/4	4

3.2 Applicazione della metodologia ai database bibliometrici

RACCOLTA DEI REQUISITI

La prima fase della metodologia prevede, una volta eseguita un'attenta analisi del database/ dei database di interesse, di procedere con la raccolta dei requisiti che gli stessi applicativi devono soddisfare.

Dal momento che i due DB qui considerati sono di fatto della stessa tipologia e vengono utilizzati per gli stessi scopi, dovranno rispondere alle medesime esigenze espresse dagli utenti e dunque quanto si procederà a descrivere di seguito è di fatto ugualmente valido per entrambi.

Sempre in un'ottica di semplificazione di tale prima fase, l'individuazione dei requisiti è avvenuta sulla base dell'esperienza personale, assumendo i punti di vista dei possibili utilizzatori del DB ed immaginando quelle che potessero essere le loro esigenze nell'interfacciamento con lo strumento. Inoltre, particolarmente di aiuto risulta essere stato anche lo standard SQuaRE che, nella ISO 25030, fornisce indicazioni circa il processo di estrazione dei requisiti.

Visto che il focus risulta essere sulla qualità dei dati, i requisiti che verranno di seguito elencati (***Tabella 3.3***) sono relativi a tale aspetto, vanno sotto il nome di DQR (Data Quality Requirements) e non guardano tanto alle caratteristiche funzionali relative alla struttura ed architettura del prodotto software in sé (DBMS).

Tabella 3.3: *Requisiti e loro descrizione*

	REQUISITO	DESCRIZIONE
CR1	Le informazioni riportate nei database corrispondono a quelle delle relative pubblicazioni e/o a quelle riportate sui siti web degli editori.	L'utente vuole avere la garanzia che i dati riportati in WoS o Scopus siano il più possibile aderenti alle informazioni ritenute vere e corrette. Queste potrebbero essere quelle riportate sulle stesse pubblicazioni, sui siti web degli editori, qualora disponibili oppure quelle recuperabili in rete da pagine comunque ritenute affidabili.
CR2	I record dei DB presentano dati in corrispondenza di tutti i campi che	Per l'utente è importante che le informazioni relative alle pubblicazioni siano complete e

	lo richiedono.	quindi che non manchino dati che in realtà dovrebbero essere presenti. Queste assenze potrebbero infatti impattare sulle successive attività di analisi che gli stakeholders vanno a compiere.
CR3	I record del DB sono identificabili in modo univoco.	È importante che sia presente un campo che consenta di identificare i record univocamente e nel quale siano vietati valori duplicati; ciò consente di evitare errori e confusioni nella ricerca delle informazioni di interesse.
CR4	I record non presentano delle contraddizioni al loro interno.	L'utente è interessato al fatto che non ci siano valori inconsistenti all'interno di uno stesso record: ad esempio, se la rivista X è stata pubblicata a partire dall'anno Y, sicuramente ci sarà un errore se viene invece storicizzato un record in cui un certo articolo risulta essere stato pubblicato su X nell'anno Y-10.
CR5	I record sono tra loro coerenti, in termini di informazioni storicizzate nei diversi campi.	L'utente desidera che, dato un certo campo, i valori memorizzati in relazione ai distinti record siano tra loro coerenti: se il campo è destinato a contenere l'anno di pubblicazione, trovare lì un numero compreso tra 1 e 10 è un errore, magari dovuto allo scambio di informazioni relative ad anno e volume della rivista.
CR6	Il DB nel suo complesso è stato certificato da un ente indipendente e riconosciuto.	In ottica qualità, per tutti i possibili utilizzatori è importante che il database sia stato validato nei suoi contenuti secondo un processo formalizzato e da un ente riconosciuto. Ciò dà opportune garanzie circa l'utilizzo dei dati nelle proprie attività operative.
CR7	I dati sono aggiornati/corrispondono alle ultime informazioni disponibili.	Qualora ci siano state delle modifiche/degli aggiornamenti alle informazioni contenute nei database, è importante che queste siano

		prontamente riportate nella base di dati. Da tenere in considerazione anche la tempestività con la quale WoS e Scopus modificano i dati errati segnalati dagli utenti.
CR8	Le informazioni sono espresse con linguaggi e simbologia tali da rendere la loro lettura ed interpretazione chiara ed immediata.	Per gli utenti è importante che i dati riportati siano espressi in un linguaggio e con simboli che rendano la loro fruizione semplice e senza problematiche.

IDENTIFICAZIONE DELLE DIMENSIONI SQuaRE DA VALUTARE

Per selezionare le caratteristiche tecniche da valutare con le prove previste nel piano di campionamento, si utilizzerà, come già specificato, il metodo del QFD: la Product Planning Matrix verrà allora utilizzata non come supporto alla progettazione della base dati, ma come strumento di definizione delle caratteristiche intrinseche dello standard SQuaRE da testare.

Per implementare correttamente lo strumento, è necessario definire, a partire dai requisiti, le caratteristiche tecniche: queste sono descrizioni chiare di come dovranno essere caratterizzati i dati e possono anche implicare la risoluzione di trade-off dal momento che capita spesso che i requisiti siano contrastanti tra loro.

Nel passaggio da requisito a caratteristiche tecniche, di fondamentale importanza è affidarsi al know-how degli esperti nel campo della data quality nonché basarsi, laddove disponibili, sulle norme e sugli standard diffusi nel campo in cui si sta lavorando: proprio per tale ragione faremo riferimento allo standard SQuaRE e a quelle che definisce come caratteristiche intrinseche (di cui è disponibile una descrizione di dettaglio nel paragrafo 1.4).

Nel nostro caso, ci siamo limitati alla costruzione del cuore della casa della qualità, ovvero la matrice delle relazioni poiché ciò ci consentirà di raggiungere l'obiettivo prefissato: dunque, per prima cosa si sono elencati i requisiti precedentemente definiti e le caratteristiche intrinseche nelle relative sezioni. A questo punto, si è proceduto all'attribuzione delle importanze dei bisogni secondo il criterio descritto in **Tabella 3.4**.

Tabella 3.4: *Classi di importanza dei requisiti e loro descrizione*

IMPORTANZA	DESCRIZIONE
5	Indispensabile
4	Molto importante
3	Importante
2	Preferibile
1	Trascurabile

L'importanza massima è stata attribuita all'aderenza dei dati riportati nei DB alle informazioni ufficiali disponibili, alla loro completezza, alla possibilità di identificare univocamente i record ed infine all'assenza di contraddizioni, sia all'interno di una stessa istanza che in relazione ad uno stesso campo per istanze distinte.

Con questi requisiti si va di fatto a coprire una buona parte degli aspetti inerenti la qualità dei dati e per gli utenti c'è la garanzia che il database consentirà loro di lavorare e trarre conclusioni sulla base di informazioni in input affidabili che non solo sono in linea con quelle ufficiali, ma anche prive di grossolani errori dovuti per lo più alla fase di caricamento.

Quindi, si è attribuito un livello di importanza pari a 4 all'aggiornamento dei dati e alla loro facilità di lettura ed interpretazione poiché sono comunque aspetti che non devono essere trascurati nella valutazione della qualità: la presenza di dati "al passo con i tempi" è senza dubbio importante, ma è un requisito che si lega anche ad aspetti tecnici relativi alle modalità di gestione degli update previste dal sistema informatico, aspetto difficilmente misurabile da un generico utilizzatore.

Per quanto riguarda invece la facilità di lettura ed interpretazione, essendo entrambi database particolarmente strutturati, è molto difficile che sia utilizzata una simbologia che renda non chiara la lettura e la fruizione dei diversi record.

In classe tre troviamo solo la certificazione da parte di un ente riconosciuto: questo è un requisito del tipo Passa/Non Passa che viene considerato soddisfatto qualora sia stato effettivamente eseguito un processo di validazione della base dati in accordo a standard comunemente riconosciuti.

Visto che anche la metodologia che si andrà ad applicare si basa su questa stessa normativa, rappresenterebbe nient'altro che un'ulteriore conferma, eseguita da un ente certificatore e non dal cliente, della qualità del database in esame.

Contestualmente, si è anche proceduto al calcolo dell'importanza relativa degli stessi requisiti, semplicemente dividendo il valore attribuito a ciascun termine per la somma delle importanze:

$$Importanza\ relativa_{\ CRi} = \frac{Importanza_{CRi}}{\sum_{i=1}^{n} Importanza_{CRi}}$$

A questo punto ci si è focalizzati sul popolamento della matrice centrale, seguendo la legenda mostrata in **Tabella *3.5***.

Tabella 3.5: *Tipi di correlazione previsti nella matrice delle relazioni*

SIMBOLO	CORRELAZIONE	PUNTEGGIO
●	Correlazione forte	9 pt
○	Correlazione media	3 pt
Δ	Correlazione debole	1 pt

Si è legato, con correlazione forte, l'aderenza dei dati riportati nei DB alle informazioni ufficiali all'accuratezza poiché questa può proprio essere quantificata in termini di numero di record presenti nel database che prevedono celle i cui valori coincidono con il loro valore vero: tanto più questo numero risulterà essere elevato, tanto più il DB sarà accurato.
L'esaustività dei dati è ovviamente legata alla completezza, ma anche all'accuratezza con un legame di tipo ○: ovviamente il database sarà tanto più accurato quanto più saranno effettivamente presenti le informazioni corrette.
La caratteristica tecnica della coerenza si lega con correlazione forte ai requisiti 3,4 e 5 che fanno di fatto riferimento a tre diversi tipi di coerenza testabile: la coerenza di chiave, la coerenza interna e la coerenza esterna.
Il requisito 6, relativo alla certificazione da parte di un ente riconosciuto, si lega fortemente alla credibilità, definita proprio come il grado con cui i dati presentano valori considerati come veri e credibili dagli utenti e debolmente alla conformità dal momento che, se i dati saranno stati opportunamente validati, è più probabile che rispettino standard, convenzioni o normative in vigore.
L'aggiornamento dei dati si lega con correlazione media all'accuratezza, alla completezza e alla coerenza: se i dati fossero aggiornati, in teoria dovrebbero anche esser stati corretti eventuali errori individuati dagli utenti e dunque il DB dovrebbe guadagnare punti in termini delle caratteristiche sopra menzionate.
La correlazione del requisito è debole anche con la tracciabilità: se si eseguono delle modifiche o degli aggiornamenti ai dati, è importante che siano presenti appositi campi che consentano di tener traccia di tutto ciò.

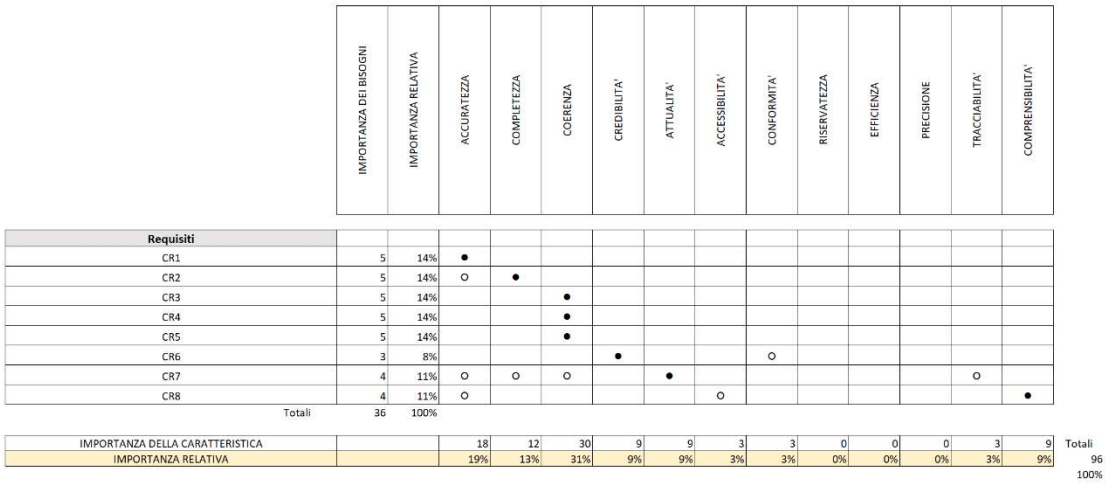
Infine, la correlazione è ovviamente forte con la caratteristica intrinseca dell'attualità che sottolinea proprio come i dati debbano avere dei valori temporalmente in linea con lo specifico contesto di utilizzo.
Infine, la facilità di lettura e comprensione si lega fortemente con la comprensibilità e con un legame di tipo ○ ad accuratezza ed accessibilità: tanto più i dati sono di immediata consultazione, tanto maggiore sarà l'accessibilità, così come se i dati sono più accurati è più facile che venga soddisfatto il requisito appena indicato.

Per ottenere dunque le importanze delle diverse caratteristiche tecniche, si è utilizzata la convenzione sopra indicata che consente di tradurre i simboli rappresentativi dei diversi tipi di correlazione in valori numerici. Si è quindi definito che:

$$Importanza_{ECj} = \sum_{i=1}^{n} R_{ij} * Importanza_{CRi}$$

$$Importanza\ relativa_{ECj} = \frac{Importanza_{ECj}}{\sum_{j=1}^{m} Importanza_{ECi}}$$

In **Figura *22*** viene riportata una vista complessiva della matrice delle relazioni, realizzata sulla base delle valutazioni appena descritte.



Requisiti	IMPORTANZA DEI BISOGNI	IMPORTANZA RELATIVA	ACCURATEZZA	COMPLETEZZA	COERENZA	CREDIBILITA'	ATTUALITA'	ACCESSIBILITA'	CONFORMITA'	RISERVATEZZA	EFFICIENZA	PRECISIONE	TRACCIABILITA'	COMPRENSIBILITA'
CR1	5	14%	●											
CR2	5	14%	○	●										
CR3	5	14%			●									
CR4	5	14%			●									
CR5	5	14%			●									
CR6	3	8%				●			○					
CR7	4	11%	○	○	○		●						○	
CR8	4	11%	○					○						●
Totali	36	100%												
IMPORTANZA DELLA CARATTERISTICA			18	12	30	9	9	3	3	0	0	0	3	9
IMPORTANZA RELATIVA			19%	13%	31%	9%	9%	3%	3%	0%	0%	0%	3%	9%

Totali 96 100%

Figura 22: *Implementazione della matrice delle relazioni per i dati dei database bibliografici*

Come si evince dal vettore contenente le importanze relative delle caratteristiche tecniche, sul podio troviamo la coerenza, quindi l'accuratezza ed infine la completezza. Sono allora proprio

queste tre le dimensioni che andremo a testare e che ci consentiranno di individuare i difetti in relazione ai diversi record.
Si è scelto di considerare le prime tre caratteristiche, che coincidono anche con quelle che possiedono un'importanza relativa superiore al 10%, poiché valutarle tutte sarebbe stato sicuramente time consuming, ma anche perché si ritiene che, ai fini della valutazione della qualità dei dati contenuti nei database, siano proprio quelli emersi gli aspetti più interessanti da valutare.
Infine, valutare aspetti quali l'attualità, la tracciabilità e l'efficienza avrebbe richiesto anche l'analisi dei sistemi software e dei metadati, elementi ai quali non è possibile accedere e che comunque richiedono conoscenze approfondite anche in relazione ai sistemi informatici.

DEFINIZIONE PROVE DA CONDURRE PER TESTARE LE DIMENSIONI SELEZIONATE

Una volta individuate le dimensioni inerenti la qualità dei dati che si vorranno testare, si è proceduto con la definizione delle prove operative da mettere in atto per ottenere i risultati di interesse.
La prova di completezza va in questo caso semplicemente a contare il numero di celle vuote nel database, dove per vuote si intendono quelle celle in cui ci si aspettava un valore in virtù del campo al quale fanno riferimento.
Nel nostro caso, teoricamente, non dovremmo averne perché, in relazione ai campi considerati nell'analisi, dovrebbero essere sempre previsti dei valori diversi da zero.
Nello specifico, i campi analizzati sono:

- Autori
- Titolo
- Anno di pubblicazione
- Titolo della fonte
- Volume
- Issue
- Pagina iniziale
- Pagina finale
- Tipo di documento

Il test è stato eseguito utilizzando una semplice funzione Excel che ha consentito di valutare se le celle del campione fossero o meno vuote. Ad esempio, in relazione al campo autore, la funzione risulta essere la seguente:

=SE (tbScopus[@Authors] = "";"VUOTA";"OK")

dove tbScopus[@Authors] rappresenta la cella corrispondente nel campione considerato.
Tale funzione è stata utilizzata per tutti i record e tutti i campi considerati; a questo punto, per ottenere una valutazione complessiva circa l'eventuale difettosità del record in esame in relazione alla completezza, si è utilizzata la funzione di seguito riportata:

=SE(O([@Authors] ="VUOTA"; [@Title]="VUOTA";[@Year]="VUOTA";[@[Source title]]="VUOTA";[@Volume]="VUOTA";[@Issue]="VUOTA";[@[Page start]]="VUOTA";[@[Page end]]="VUOTA";[@DOI]="VUOTA";[@[Document Type]]="VUOTA"); "X";"")

Il suo obiettivo è semplicemente quello di andare a marcare con una X quei record che presentano una o più celle vuote e che quindi presentano dei difetti in relazione alla dimensione appena testata.

La prova di coerenza di chiave, volta a valutare l'unicità e la presenza della chiave primaria, ha previsto da un lato l'utilizzo della formattazione condizionale per mettere in evidenza gli eventuali DOI duplicati e dall'altro ha consentito l'individuazione dei record privi di valori in questo campo. Si noti che l'assenza del DOI era sta già valutata con la prova di completezza, ma l'ulteriore considerazione di tale aspetto è voluta dal momento che tale campo assolve alla funzione di chiave primaria, elemento centrale in un database relazionale.

La prova di coerenza interna ha, invece, confrontato il numero di pagina iniziale e il numero di pagina finale, ma anche ri-evidenziato l'eventuale assenza di questi campi con la seguente funzione:

=SE (E([@[Page start]]="";[@[Page end]]=""); "X"; SE([@[Page start]]<=[@[Page end]];"";"X"))

Se i campi Page Start e Page End sono entrambi vuoti, viene inserita direttamente una X; in caso contrario, si valuta se il numero di pagina finale è maggiore o uguale al numero di pagina iniziale e, in caso di risposta affermativa, la prova sarà superata.

La prova di coerenza esterna si è focalizzata sul tipo di dato con il quale le diverse informazioni risultano essere storicizzate nel database: ad esempio in relazione al titolo, si è valutato se il suo tipo fosse NUMERO o STRINGA.

=SE (TIPO (tbScopus[@Title]) = 1;"NUMERO"; "STRINGA")

I tipi di dato che si sono considerati corretti in relazione ai diversi campi sono indicati in **Tabella 3.6**: quella fatta è ovviamente una semplificazione, volta a snellire e automatizzare le attività di analisi svolte.

Tabella 3.6: *Tipi di dato considerati corretti in relazione ai campi analizzati*

CAMPO	TIPO DI DATO
Autore	STRINGA
Titolo articolo	STRINGA
Anno di pubblicazione	NUMERO
Titolo della fonte	STRINGA
Volume	NUMERO
Issue	NUMERO
Pagina iniziale	NUMERO
Pagina finale	NUMERO
DOI	STRINGA
Tipo di documento	STRINGA

Dunque, per valutare quali record presentassero dei difetti in relazione alla coerenza esterna, si è confrontata la stringa derivante dalla concatenazione delle celle ottenute tramite l'impiego della funzione SE precedentemente riportata con la seguente stringa:

STRINGASTRINGANUMEROSTRINGANUMERONUMERONUMERONUMEROSTRINGASTRINGA

derivante dall'unione dei tipi di dato indicati in **Tabella 3.6**.

Si sono potute così individuare eventuali deviazioni rispetto ai tipi considerati come corretti.

In relazione all'accuratezza, si era specificato che si sarebbero ricercate le celle il cui contenuto non fosse considerato come accurato, sia semanticamente che sintatticamente. Dunque, la prova che si andrà ad eseguire prevede il confronto del campione estratto dal database in esame con un "riferimento assoluto" o *Golden Standard*.
Nel nostro caso, la sua formazione è avvenuta seguendo alcune linee guida:

- La definizione del contenuto dei record è in prima battuta avvenuto confrontando i record estratti, in relazione ad uno stesso campione, da Scopus e WoS; se i contenuti combaciavano, si è considerato ciò come una garanzia circa la correttezza delle informazioni riportate e dunque si è popolato l'oracolo con questi stessi valori.
  Viceversa, nel caso in cui fossero presenti delle difformità, si è proceduto alla consultazione dei siti web degli editori o, in loro assenza, alla ricerca di informazioni in rete che potessero consentire di individuare il "valore vero".
  Si tratta ovviamente di un processo estremamente lungo e complesso, soprattutto perché le difformità tra gli export sono molteplici e richiedono l'esecuzione di alcune operazioni preliminari di data preparation che di fatto hanno un impatto sui risultati ottenuti.
- I nomi degli autori sono stati riportati, laddove necessario e sia nei campioni che nel riferimento assoluto, nella forma:

  Cognome_1, Inziale_Nome_1; Cognome_2, Iniziale_Nome_2; etc

  O

  Cognome_1 Inziale_Nome_1; Cognome_2 Iniziale_Nome_2; etc
- Si sono considerati solo i titoli delle pubblicazioni in inglese, dal momento che WoS non riporta quelli in lingua originale;
- Nei titoli delle pubblicazioni si sono effettuate alcune sostituzioni in termini di simbologia utilizzata, per uniformare il contenuto delle diverse celle ed evitare che venissero considerate come non accurate a fronte di un minimo simbolo.
- Qualora non ci fosse una corrispondenza esatta tra i dati riportati nei due DB e non fosse possibile risalire ai valori veri da riportare nelle celle, si sono considerati come corrette le informazioni riportate nel database in esame.

Una volta effettuato il confronto con il riferimento assoluto in relazione a tutti i diversi campi previsti, si è proceduto con l'individuazione dei record difettosi, grazie all'ausilio della seguente funzione:

=SE(O([@Authors]="KO";[@[Article Title]]="KO";[@[Publication Year]]="KO";[@[Source Title]]="KO";[@Volume]="KO";[@Issue]="KO";[@[Start Page]]="KO";[@[End Page]]="KO";[@DOI]="KO";[@[Document Type]]="KO"); "X";"")

che consente, qualora un dato confronto non sia andato a buon fine, di marcare quel record perché caratterizzato da uno o più difetti di accuratezza.

Nel confronto occorre prestare particolare attenzione qualora non ci sia una corrispondenza esatta tra la posizione di un record nel database in esame e la posizione della corrispondente istanza nel riferimento assoluto; occorrerà in tal caso intervenire manualmente per inserire i riferimenti corretti.

Qualsiasi record che non supererà i test appena descritti dovrà essere selezionato poiché presenterà dei difetti in relazione alla caratteristica dell'accuratezza.

PROGETTAZIONE DEL PIANO DI CAMPIONAMENTO

Avendo definito con esattezza le prove che si eseguiranno per testare la qualità dei dati contenuti nei database bibliometrici, si può ora dare una definizione dei concetti di difetto e difettoso nello specifico contesto in esame.

Un difetto è una non conformità, ovvero una manifestazione del non raggiungimento delle specifiche di accuratezza, completezza o coerenza; ne consegue che un record sarà definito difettoso nel momento in cui non sarà conforme ad almeno una delle specifiche appena introdotte e quindi sarà caratterizzato dalla presenza di uno o più difetti.

Il controllo che andremo ad eseguire è di tipo campionario, ma ancor più precisamente di tipo statistico dal momento che si procederà alla determinazione del campione e dei criteri di accettazione sulla base della numerosità della popolazione e della sua tipologia.

L'obiettivo, come già introdotto, sarà quello di prendere una decisione circa l'accettazione o meno della base dati e ciò sarà fatto valutando la qualità dei dati contenuti nei database bibliometrici tramite l'impiego delle caratteristiche SQuaRE ritenute più importanti nell'ottica degli utilizzatori degli stessi strumenti.

Come prima ipotesi di lavoro, si è scelto di considerare come lotto in ingresso in relazione al quale prendere la decisione di accettazione o rifiuto l'intero database, che sia esso WoS e Scopus.

A questo punto, si è reso necessario definire la dimensione del campione n e il numero di accettazione c e sono state a tal proposito seguite due diverse strade:

1. **Progettazione di un piano di campionamento ad hoc**

Primo passo per procedere con la costruzione di un piano di campionamento, è quello di definire le posizioni di fornitore e committente e quindi i valori di α, β, AQL e LTDP**.**

Come indicato nel **CAPITOLO II**, nel paragrafo **2.3 La nuova metodologia di assessment della qualità**), il settaggio di questi parametri è avvenuto facendo delle ipotesi di lavoro che hanno cercato di tenere in considerazione il prodotto in esame e le sue peculiarità.

In primo luogo, si è scelta la seguente configurazione dei parametri di interesse:

Tabella 3.7: *Parametri piano ad hoc (Scenario ottimistico)*

PARAMETRO	***VALORE***
AQL	4%
α	5%
LTPD	6%
β	10%

Per quanto concerne AQL e LTPD, si è tenuto in considerazione il fatto che il tasso di difettosità dei database bibliografici calcolato in diverse pubblicazioni è compreso tra il 6 % e il 12%: prendendo dunque tale valore come espressione del punto di vista del cliente, si è settata la massima difettosità accettabile dallo stesso al valore minimo del 6%.

Dunque, poiché AQL deve essere inferiore a LTPD e non si avevano indicazioni da parte dei due DB circa la loro difettosità, lo si è posto al 4%.

Tale scelta implica l'adozione di un punto di vista ottimistico circa la massima difettosità ammessa da cliente e fornitore.

La definizione di α e β è avvenuta ammettendo che il rischio per il fornitore è più basso che per il cliente dal momento che il primo ha maggiori informazioni circa i dati presenti a sistema e le modalità relative al loro caricamento e gestione, mentre il secondo, parte debole, corre un maggior rischio di accettare come conformi dei lotti di qualità ritenuta insoddisfacente.
Come conseguenza di tale situazione, i valori di n e c risultanti ed ottenuti dal diagramma di Larson (**Figura *10***), sono:

n=800

c=40

Adottando invece un punto di vista pessimistico circa la massima difettosità ammessa, sono stati considerati i seguenti parametri:

Tabella 3.8: *Parametri piano ad hoc (scenario pessimistico)*

PARAMETRO	***VALORE***
AQL	9%
α	5%
LTPD	12%
β	10%

In tal caso, la massima difettosità ammessa dal cliente è posta pari al limite superiore dell'intervallo relativo al tasso di difettosità emerso dagli studi condotti in materia; di conseguenza, è stata fatta crescere anche la massima difettosità riscontrata dal fornitore.
L'impostazione dei valori α e β è stata eseguita seguendo la stessa ratio del caso precedentemente analizzato.
I valori risultanti sono:

n=800

c=90

Potremo dunque confrontare i due scenari appena proposti, ottenuti a parità di rischio ammesso da cliente e fornitore, per valutare possibili variazioni circa la decisione di accettazione o rifiuto dei database in esame.

2. Progettazione del piano secondo norma MIL STD 105E

Per completezza della trattazione, si è scelto di progettare i piani di campionamento anche sulla base della norma MIL STD 105E.

In tal caso, come primo step, si è selezionata la dimensione del lotto, superiore a 500001, e il livello di ispezione desiderato: per semplicità, si è scelto di operare con un'ispezione di tipo generale e inizialmente ridotta.

Ne consegue che il codice letterale che definisce la dimensione del campione è in tal caso "N" ed n è pari a 500.

Considerando AQL pari al 4%, come nello scenario ottimistico, otteniamo che il numero di accettazione è pari a 18 e il numero di rifiuto a 19.

Anche facendo salire il valore di AQL al 10%, numero di accettazione e numero di rifiuto non cambiano.

Altra soluzione, qui non trattata, sarebbe quella di scegliere di operare con un livello di ispezione sempre generale, ma normale; in tal caso, il campione dovrebbe essere costituito da 1250 record e, in relazione ai valori di AQL considerati, il numero di accettazione e di rifiuto sarebbero rispettivamente pari a 21 e 22.

Qualora i database vengano rifiutati nel caso di ispezione ridotta, è altamente probabile che siano rifiutati anche nel caso di ispezione normale dal momento che la dimensione del campione è più che raddoppiata, mentre il numero di accettazione sale solo di tre unità.

Al fine di agevolare le operazioni di analisi si sono estratti, dai database WoS e Scopus, due campioni casuali con dimensione di circa 800 record; quindi, per quei piani che richiedevano una dimensione del campione pari a 800, si sono considerate le estrazioni nella loro interezza, mentre per la prova con campione da cinquecento unità, si sono selezionati solo i primi cinquecento record.

Ciò ha consentito di snellire le operazioni di estrazione dei campioni da analizzare, particolarmente time consuming dal momento che si è cercato di ottenere, come verrà di seguito descritto, un campione che fosse il più rappresentativo possibile della popolazione e che quindi considerasse le peculiarità delle diverse discipline e dei diversi editori inclusi nei database.

Quanto finora descritto ha considerato i due database nella loro interezza come lotti di interesse in relazione ai quali prendere la decisione di accettazione o rifiuto; tuttavia, quando un determinato utente sceglie di accedere ad uno di questi database lo fa principalmente perché risulta essere interessato ad un certo sottoinsieme dei record. Potrebbe allora aver senso ipotizzare degli use case ed applicare i piani di campionamento, con le relative prove di testing delle dimensioni di interesse, a lotti così definiti.
Come esempio, si è considerato il seguente caso d'uso (**Tabella *3.9***):

Tabella 3.9: *Descrizione Use Case 1*

USE CASE 1	
Un valutatore vuole conoscere le pubblicazioni effettuate sulla rivista Production Planning & Control nell'anno 2018.	
PUBLICATION YEAR	2018
RIVISTA	Production Planning & Control
EDITORE	Taylor & Francis

In tal caso, il campione di interesse è costituito da novantasei record; applicando la norma MIL STD 105E e l'ispezione ridotta, risulta che il subset da analizzare dovrà possedere otto record e i numeri di accettazione e rifiuto saranno rispettivamente pari a 1 e 2.
Essendo tale caso molto semplice, si è scelto di applicare anche un'ispezione normale, in virtù della quale risulta che il campione è costituito da venti record e numero di accettazione e rifiuto sono pari a:

- 2 e 3 nel caso di AQL al 4%
- 5 e 6 nel caso di AQL al 10%

Altro esempio di use case potrebbe essere quello descritto in **Tabella *3.10***.

Tabella 3.10: *Descrizione Use Case 2*

USE CASE 2	
Uno studente vuole avere un quadro completo circa gli articoli pubblicati in tema di data quality negli ultimi due anni, nel campo dell'ingegneria e relativi agli aspetti di controllo della qualità.	
PARAMETRI	***VALORE***

TITOLO	Data quality
TIPOLOGIA	Articoli
PUBLICATION YEAR	2020-2022
AREA	Ingegneria

In tal caso il numero di record risulta essere compreso tra 281 e 500 e i parametri del piano di campionamento sono di seguito definiti:

- Nel caso di ispezione ridotta, il campione è costituito da 20 record e il numero di accettazione è pari a 1 (AQL=4%) o a 2(AQL=10%).
- Nel caso di ispezione normale, abbiamo un campione da cinquanta unità ed il numero di accettazione è pari a 5 (AQL=4%) o a 10 (AQL=10%).

ESECUZIONE DEL PIANO E DELLE PROVE

Si è dunque passati alla fase di implementazione vera e propria del piano di campionamento e delle relative prove per poter poi trarre le opportune conclusioni.

Campione 1

La costruzione del primo campione sul quale eseguire le varie analisi è stata effettuata a partire dall'elenco di editori coperti da Scopus, ricavato dal CiteScore 2020. Dunque, si è proceduto con l'estrazione casuale dei publishers da considerare nel campione con l'impostazione di una semplice funzione Excel:

INDICE ($A$2: $A$7044; CASUALE.TRA (1;7043))

dove:

$A$2: $A$7044 corrisponde all'intervallo di celle contenenti i nomi dei diversi editori

CASUALE.TRA (1;7043) è una funzione che consente di estrarre un numero casuale tra 1 e 7043.

Si noti che non si è impostato a priori un numero di editori da considerare, ma si sono estratti nomi fintanto che il numero di record non avesse raggiunto circa le ottocento unità, necessarie in virtù del piano di campionamento ad hoc precedentemente definito.

In conclusione, gli editori considerati sono riassunti in **Tabella *3.11***:

Tabella 3.11: *Editori selezionati per la costruzione del campione in Scopus*

EDITORI
University of Auckland

Statistics Canada
Wilmington Scientific Publishers
Institution of Chemical Engineers
Institute of Archealogy of the Academy of Sciences of the Czech Republic
Rockefeller University Press
Technical Publications Co.
Institute of Eastern Europe and Central Asia
Elsevier

A questo punto, per ciascun editore, si è casualizzata, laddove possibile, l'estrazione della disciplina di interesse, quindi della rivista ed infine dell'anno da considerare per la costruzione del campione.

Ad esempio, in relazione ad Elsevier, erano presenti riviste relative a ben 313 discipline diverse: da questo insieme, si è estratta l'area definita come *Business, Management and Accounting (all).* A questo punto, si sono selezionate le sole riviste attinenti a questo campo e si è estratta casualmente una di esse (*International Journal of Production Economics*).

Infine, consultando Scopus, si sono ricercati e riportati nel foglio Excel gli anni in cui risultavano esserci pubblicazioni sulla rivista in esame e si è estratto casualmente anche l'anno di interesse, che è risultato essere il 2017.

A questo punto, si è proceduta all'estrazione dei record caratterizzati da questi attributi dai due database.

Si noti che, come già precedentemente specificato, tale campione è stato costruito con circa 800 record in modo da poter essere utilizzato sia per il piano ad hoc che per il piano definito in accordo alla MIL STD 105E.

Nella **Tabella *3.12*** , si riportano i risultati delle diverse estrazioni eseguite e il numero dei record estratti in relazione a ciascun editore; per il dataset completo, si rimanda al file Excel in allegato (Campione 1_Definizione).

Laddove non si riportano indicazioni, è perché non aveva senso casualizzare l'estrazione poichè, ad esempio, era presente un'unica disciplina o un'unica rivista.

In colonna cinque, si riporta il numero di record trovati in Scopus e tra parentesi tonde () quanto estratto da WoS: questo perché il riferimento ufficiale, necessario per la prova di accuratezza, richiede il confronto tra le informazioni presenti nei due database.

Tabella 3.12: *Tabella riassuntiva del processo di definizione del campione 1*

EDITORE	***DISCIPLINA***	***RIVISTA***	***ANNO***	***NUMERO***

				RECORD
University of Auckland	Archeology (arts and humanities)	Journal of the Polynesian Society	2020	15 (20)
Statistics Canada	Modeling and Simulation	Survey Methodology	2017	15 (15)
Wilmington Scientific Publishers	//	Journal of Applied Analysis and Computation	2015	61 (61)
Institution of Chemical Engineers	Chemical Engineering (all)	Food and Bioproducts Processing	2009	43 (43)
Institute of Archealogy of the Academy of Sciences of the Czech Republic	//	Pamatky Archeologicke	2016	7 (21)
Rockefeller University Press	Biochemistry, Genetics and Molecular Biology (miscellaneous)	Life Science Alliance	2019	160 (148)
Technical Publications Co.	//	Control Engineering	2007	210 (378)
Institute of Eastern Europe and Central Asia	//	Journal of Eastern European and Central Asian Research	2020	33 (33)
Elsevier	Business, Management and Accounting	International Journal of Production Economics	2017	308 (308)

Una volta che il campione di partenza è stato costruito secondo le indicazioni sopra riportate, si è proceduto alla delimitazione dei set di record da considerare nei due piani di campionamento. Nel caso di progetto ad hoc, si sono considerati i primi 800 record, mentre nel caso di applicazione della MIL STD 105E i primi 500: si è scelto di operare secondo tale metodologia dal momento che si era già casualizzata l'estrazione dei record in partenza.

Nel caso di progettazione ad hoc del piano di campionamento, sia che venga adottato un punto di vista ottimistico che pessimistico, il campione di interesse risulta essere costituito da ottocento record.

Una volta che sono state eseguite tutte le prove (come mostrato nel dettaglio nel file Campione 1_Piano ad hoc), occorre tirare le somme e contare il numero di record difettosi, perché caratterizzati dalla presenza di uno o più difetti.
Considerando i risultati delle prove di testing delle varie dimensioni, risulta che il numero di record difettosi è pari a 663 e quindi, sia nel caso di scenario pessimistico (c=90) che ottimistico (c=40), non si dovrebbe accettare il database in esame.
La decisione circa l'accettazione è avvenuta riportando gli esiti delle singole prove su un foglio Excel ad hoc e definendo un giudizio complessivo grazie alla seguente funzione:

=SE (O (B2="X"; C2="X"; D2="X";E2="X"; F2="X"); "X";"")

dove le colonne B, C, D, E ed F riportano rispettivamente gli esiti delle prove di completezza, coerenza di chiave, coerenza interna, coerenza esterna ed accuratezza.

Anche qualora si scegliesse di applicare la MIL STD 105E e quindi di considerare un campione di 500 unità, sia nel caso di scenario ottimistico che pessimistico, si dovrebbe rifiutare la fornitura di Scopus dal momento che il numero di record difettosi risulta essere pari a 350 unità contro un numero di accettazione sempre pari a 18.

Campione 2

La costruzione del campione 2 è avvenuta cercando di ricalcare il più possibile quella eseguita per il campione 1, ma considerando come punto di partenza Web of Science; nello specifico, tuttavia, non avendo a disposizione una lista degli editori, la prima casualizzazione è stata effettuata in relazione alle discipline.
Anche in tal caso si sono via via estratte delle categorie fin tanto che la dimensione del campione non raggiungesse un valore prossimo ad ottocento record; dunque, si è sempre casualizzata anche l'estrazione della rivista per ciascuna disciplina e dell'anno di interesse per le singole riviste.
Nella **Tabella *3.13***, si riportano i risultati delle diverse estrazioni eseguite e il numero dei record estratti in relazione a ciascun editore; per il dataset completo, si rimanda al file Excel in allegato (Campione 2_Definizione).

Tabella 3.13: *Tabella riassuntiva del processo di definizione del campione 2*

DISCIPLINA	***RIVISTA***	***ANNO***	***NUMERO RECORD***
Area studies	Journal of Contemporary China	2019	108 (61)
Demography	Population Space and Place	2015	63 (59)
Computer Science, Hardware & Architecture	Journal of Computer and System Sciences	1994	120 (120)
Behavioral Science	Pharmacology Biochemistry and Behavior	2015	221 (219)
Mechanics	Communications in Nonlinear Science and Numerical Simulation	2007	125 (125)
History of social science	History of Education Review	2015	23 (16)
Telecommunications	IEEE Systems Journal	2015	145 (149)
Nursing	Japan Journal of Nursing Science	2007	17 (17)

Si noti che per questo secondo campione i valori riportati nella colonna "Numero record" sono relativi a Web of Science, mentre quelli tra parentesi tonde fanno riferimento a Scopus.
Una volta ottenuto il dataset dal quale estrarre i record per eseguire il piano di campionamento ad hoc ed il piano definito dalla MIL STD 105E, si sono implementate le prove con le stesse modalità sopra riportate e si è definito il numero di record difettosi.
I numeri di record difettosi riscontrati nell'ambito delle diverse prove sono riportati in **Tabella *3.14*** , dove non si sono riportati gli esiti della prova di accuratezza dal momento che la stessa non è stata eseguita.
Si può notare non solo come la maggior parte dei difetti siano relativi alla prova di completezza ma anche come, nonostante la diversa numerosità del campione (800 vs 500), i numeri di record difettosi riscontrati siano molto simili e in entrambi i casi superiori ai numeri di accettazione previsti nello scenario ottimistico e pessimistico.

Tabella 3.14: *Esiti delle prove condotte sul campione 2*

	COMP	*COEC*	*COEI*	*COERE*	**ACC**	**TOT**
PIANO AD HOC	224	2	0	2	**151**	**225**
MIL STD 105E	224	1	0	1	**82**	**245**

Campione Use Case 1

Per quanto concerne lo use case 1, in primis si è proceduto con il download dei record di interesse sulla base dei parametri indicati in **Tabella 3.9**, quindi si è passati alla definizione dei campioni sulle quali eseguire i test previsti.

In particolare, si sono estratti casualmente quattro campioni che fanno riferimento alle prove di seguito indicate:

- **PROVA A**

Nel caso di ispezione ridotta e per ognuno dei due database in esame, si sono estratti casualmente otto record dai novantasei di partenza e quindi si è proceduto all'esecuzione di tutte le prove.

I campi considerati, così come le modalità di implementazione delle prove in Excel, sono esattamente le stesse sopra descritte.

Nel caso di *Scopus,* i riferimenti dei record esaminati sono riportati in **Tabella 6.1** ed in tal caso tutte le prove (completezza, coerenza di chiave, coerenza interna, esterna ed accuratezza) hanno avuto esito positivo.

Dunque, poiché non si sono riscontrati difetti in nessuna delle dimensioni testate, il numero dei record difettosi risulta essere in tal caso nullo.

Per quanto riguarda l'implementazione delle prove, si rimanda al file Production Planning & Control_Scopus_Prova A.

In conclusione, visto che 0<1, pari al numero di accettazione previsto, il database risulta essere accettato per condurre analisi sulle pubblicazioni relative al caso di studio qui esaminato.

Per il database Web of Science, invece, il campione estratto è riportato in

Tabella *6.2*, mentre le informazioni di dettaglio sono nel file Production Planning & Control_WoS_Prova A.
In tal caso si è individuato un solo record difettoso (N° 7 della
Tabella *6.2*): il difetto che esso presenta è relativo alla dimensione dell'accuratezza e risulta essere dovuto al fatto che non vengono correttamente riportati i cognomi di due dei tre autori della pubblicazione.
Ciò di fatto causa una non corrispondenza nel confronto tra campione e riferimento assoluto, che comunque non impatta sulla decisione di accettazione delle informazioni in esame.

In conclusione, per entrambi i database e qualora si esegua un'ispezione ridotta, l'esito del controllo è positivo e conduce all'accettazione degli stessi DB; questo è in parte dovuto al fatto che l'ispezione ridotta sottintende, già in partenza, una certa fiducia nella fornitura ricevuta e nell'affidabilità del fornitore.

- **PROVA B**

Proprio per valutare cosa succeda qualora si scelga di adottare un tipo di ispezione normale, si è proceduto ad applicare, sempre in relazione al primo use case, questo secondo tipo di schema di campionamento.
In tal caso, i campioni estratti per i due database sono costituiti da venti record e riportati in **Tabella 6.3** e **Tabella 6.4** e nei relativi file di analisi in allegato (Production Planning & Control_Scopus_Prova B e Production Planning & Control_WoS_Prova B).

Nel caso di Scopus, solo il ventesimo record risulta essere difettoso a causa della presenza di un difetto di accuratezza relativo al tipo di pubblicazione.
Allora, a prescindere che AQL venga posto al 4% o al 10%, il numero di record difettosi è inferiore ai numeri di accettazione previsti, rispettivamente pari a 5 e 6 e, anche nel caso di ispezione normale, Scopus supera il controllo in accettazione eseguito dall'ipotetico cliente che vuole utilizzare le informazioni storicizzate al fine di condurre le sue attività di analisi.
L'esito dipende invece dal valore di AQL nel caso di Web of Science poiché i difetti di accuratezza vanno a rendere difettosi cinque distinti record: nel caso di AQL=4%, il database risulterà essere rifiutato dal momento che 5>2, mentre qualora il fornitore (e il cliente)

scelgano di accettare un più alto livello di difettosità massima, l'esito del controllo è positivo visto e considerato che il numero di accettazione è esattamente pari a 5.
In conclusione, in relazione al primo use case e alla conduzione di un'ispezione normale, gli esiti sono riassunti in **Tabella 3.15**.

Tabella 3.15: *Esito dell'applicazione dei piani di campionamento per lo use case 1 ed in caso di ispezione normale*

	AQL	
	4%	**10%**
SCOPUS	**ACCETTATO** ✓	**ACCETTATO** ✓
WOS	**RIFIUTATO** ✗	**ACCETTATO** ✓

Campione Use Case 2

Il secondo use case, volto a valutare la qualità dei dati estratti dai database bibliometrici proprio in tema di data quality, ha portato all'estrazione di quattro campioni dai due export ottenuti da Scopus e Web of Science poiché, anche per questa casistica, si è scelto di valutare la variazione degli esiti a seconda del tipo di ispezione scelto.

- **PROVA C**

Nel caso di ispezione ridotta, il campione estratto è costituito, per entrambi i database, da 20 record, anche in tal caso sottoposti alle prove di completezza, coerenza (interna, esterna e di chiave) e accuratezza.
A differenza dello use case 1, dove la fonte era una costante, si è ora considerata tra i campi anche la rivista sulla quale sono state effettuate le pubblicazioni.

Nel caso di Scopus (**Tabella 6.5**), già nella sola prova di completezza vengono individuati trentacinque difetti, che rendono ben quattordici record difettosi (
Tabella *3.16*).

Tutti i difetti relativi alla coerenza interna coincidono di fatto con difetti di completezza, poiché relativi a numeri di pagina mancanti.

Tabella 3.16: *Difetti riscontrati in relazione alle diverse dimensioni nella prova C per Scopus*

	Completezza	Coerenza di chiave	Coerenza interna	Coerenza esterna	Accuratezza	TOT
1						
2	X		X			X
3	X	X				X
4	X		X		X	X
5	X		X			X
6	X		X		X	X
7	X		X			X
8					X	X
9				X		X
10	X					X
11	X		X			X
12	X		X			X
13	X		X			X
14	X		X		X	X
15	X		X		X	X
16	X		X			X
17						
18						
19					X	X
20	X					X

Segue allora che Scopus risulta essere rifiutato, sia nel caso di AQL pari al 4%, che nel caso di AQL al 10%.

A tal proposito, vale la pena notare come, nel campione qui considerato, siano presenti anche articoli in press, per i quali è in corso la stampa e dunque non sono ancora note informazioni circa il volume, l'issue e l'impaginazione.

Ovviamente, escludere questi articoli potrebbe cambiare l'esito della prova ma comunque richiederebbe l'esecuzione di tutta una serie di operazioni, volte ad accertare lo stato delle diverse pubblicazioni, che ridurrebbero ulteriormente l'automaticità della procedura prevista.

Inoltre, si ipotizza che l'assenza di queste informazioni sia temporanea e dunque risolta una volta verificatasi l'effettiva pubblicazione degli articoli coinvolti.

Nel caso di WoS, il campione di riferimento è illustrato in **Tabella 6.6** e anche in tal caso, già dalle prime prove, il numero di record risultanti difettosi è particolarmente elevato (**Tabella *3.17***).

Tabella 3.17: *Difetti riscontrati in relazione alle diverse dimensioni nella prova C per WoS*

	Completezza	Coerenza di chiave	Coerenza interna	Coerenza esterna	Accuratezza	TOT
1	X		X		X	X
2	X		X		X	X
3	X		X			X
4	X		X			X
5						
6	X		X			X
7	X					X
8						
9						
10	X		X			X
11	X					X
12	X		X			X
13	X		X		X	X
14	X					X
15						
16						
17						
18	X		X		X	X
19				X	X	X
20						

Dunque, anche WoS risulta essere rifiutato, sia nel caso di AQL pari al 4%, che nel caso di AQL al 10%.

PROVA D

Avendo avuto esiti negativi nel caso di ispezione ridotta, ci si aspetta che, applicando una modalità di ispezione normale al campione derivante dal secondo use case, i risultati continuino ad implicare un rifiuto della fornitura.

Infatti, sia nel caso di Scopus che di WoS, i campioni di cinquanta unità risultano essere caratterizzati da un numero di record difettosi superiori ai numeri di accettazione previsti dalla MIL STD 105E in funzione di un AQL rispettivamente pari al 4% o al 10% (c=5 e c=10).

Per Scopus, il numero di difettosi risulta essere pari a 34, mentre per WoS, sempre nelle stesse condizioni, a 32.

Ne consegue allora che entrambi i database sono da rifiutare in accordo al piano di campionamento.

I file di dettaglio riportanti l'esecuzione della prova D sono in allegato e disponibili ai seguenti collegamenti:

- SCOPUS: Data Quality_Scopus_Prova D
- WOS: Data Quality_WoS_Prova D

3.3 Analisi dei risultati

Una volta condotte tutte le diverse prove previste ed aver ottenuto i diversi esiti circa l'accettazione o meno dei database in esame, occorre in primis fare un riepilogo circa le evidenze ottenute (**Tabella *3.18***) per poter poi procedere ad eventuali analisi quantitative e qualitative dei risultati.

Tabella 3.18: *Tabella riassuntiva degli esiti dei diversi piani di campionamento*

CAMPIONE	PIANO	SCOPUS	WOS
CAMPIONE 1	PIANO AD HOC - SCENARIO OTTIMISTICO	✗	*NA*
	PIANO AD HOC - SCENARIO PESSIMISTICO	✗	*NA*
	MIL STD 105 E - SCENARIO OTTIMISTICO	✗	*NA*
	MIL STD 105 E - SCENARIO PESSIMISTICO	✗	*NA*
CAMPIONE 2	PIANO AD HOC - SCENARIO OTTIMISTICO	*NA*	✗
	PIANO AD HOC - SCENARIO PESSIMISTICO	*NA*	✗
	MIL STD 105 E - SCENARIO OTTIMISTICO	*NA*	✗
	MIL STD 105 E - SCENARIO PESSIMISTICO	*NA*	✗
USE CASE 1	ISPEZIONE RIDOTTA – SCENARIO OTTIMISTICO	✓	✓

	ISPEZIONE RIDOTTA – SCENARIO PESSIMISTICO	✓	✓
	ISPEZIONE NORMALE – SCENARIO OTTIMISTICO	✓	✓
	ISPEZIONE NORMALE – SCENARIO PESSIMISTICO	✓	✗
USE CASE 2	ISPEZIONE RIDOTTA – SCENARIO OTTIMISTICO	✗	✗
	ISPEZIONE RIDOTTA – SCENARIO PESSIMISTICO	✗	✗
	ISPEZIONE NORMALE – SCENARIO OTTIMISTICO	✗	✗
	ISPEZIONE NORMALE – SCENARIO PESSIMISTICO	✗	✗

Quello che emerge è in primis come la maggior parte delle prove (17 su 24) abbiano esito negativo e conducano ad un rifiuto della base di dati analizzata: Scopus risulta essere rifiutato 8 volte su 12, mentre WoS 9 volte su 12 e dunque la prima base di dati sembrerebbe essere più performante in termini di qualità dei dati.

Si può inoltre constatare come entrambi i DB presentino esiti migliori, o comunque un minor numero di record difettosi, nel caso degli use case mentre per i campioni più grandi il numero di difettosi cresce notevolmente e supera di parecchio i numeri di accettazione di volta in volta previsti.

In particolare, si assiste all'accettazione dei DB nel momento in cui l'analisi che si vuole condurre risulta essere topic specific: ciò implica non solo un dataset di dimensioni particolarmente ridotte, ma anche la sua definizione tramite ben precisi parametri.

Nell'ambito delle diverse prove condotte, risultati particolarmente negativi vengono riscontrati nella prova di completezza, dove man mano che cresce la dimensione del campione

cresce il numero di record difettosi: i campi più coinvolti dall'assenza delle relative informazioni sono quelli di volume, issue, impaginazione e DOI.
Ciò potrebbe esser dovuto alla presenza di articoli in fase di stampa, per i quali non sono ancora stati definiti i sopracitati campi oppure ad articoli pubblicati non sulla stampa cartacea ma online, per i quali queste stesse informazioni non sono definite.
In relazione alla prova di coerenza di chiave, non si è mai verificato che due articoli avessero lo stesso DOI, anche perché esistono meccanismi di verifica dell'unicità di attribuzione di tale campo; più numerosi sono invece i casi di DOI assente, aspetto che, come già sopra specificato, è stato considerato nell'ambito della prova di completezza.
L'assenza del DOI potrebbe esser dovuta ad una mancata richiesta di codifica da parte dei pubblicatori oppure potrebbero essersi verificati dei problemi in fase di caricamento delle informazioni negli stessi database bibliometrici.
In relazione alla prova di coerenza esterna, si è considerato che ciascun campo dovesse contenere informazioni di un solo tipo (o stringa o numero): in realtà, vista la varietà ed eterogeneità delle riviste, si potrebbero ad esempio avere informazioni in formato stringa per quanto riguardo il volume o l'impaginazione.
Il fatto di aver allora introdotto questa esemplificazione, necessaria al fine di una maggiore automatizzazione delle attività di analisi, potrebbe aver incrementato il numero di difetti di coerenza esterna riscontrati.

Nella prova di accuratezza, sono stati riscontrati errori relativamente a:

- Spelling dei cognomi degli autori o di parole riportate nei titoli delle pubblicazioni. Tale errore era stato evidenziato anche da Franceschini et al., che avevano sottolineato come fosse piuttosto serio dal momento che non solo riguarda parole molto semplici, ma anche perché potrebbe essere facilmente risolvibile con sistemi di spell-checking;
- Assenza di simboli speciali in relazione a cognomi degli autori o titoli delle pubblicazioni: tale difetto è per lo più presente nei dati scaricati da Web of Science;
- Assenza di record relativi a pubblicazioni in realtà presenti nelle riviste considerate: ciò potrebbe essere dovuto ad errori o problematiche manifestatesi nella fase di uploading dei dati.
- Indicazione di pubblicazioni non effettivamente presenti nelle riviste analizzate: è infatti frequente che i database bibliometrici vadano a riportare tra i record articoli o altri tipi di pubblicazioni in realtà non afferenti ad una certa rivista.

- Confusione tra nomi e cognomi degli autori: tale errore potrebbe ad esempio essere dovuto ad errori commessi in fase di citazione della relativa pubblicazione ed essersi poi trascinato anche all'interno degli stessi database bibliometrici.
  Risulta essere particolarmente evidente nel caso di autori di origine asiatica.
- Diversità nelle scelte fatte dai DB in merito alla codifica delle informazioni, ad esempio in relazione ai titoli delle riviste

Tabella 3.19: *Differenza nelle indicazioni dei titoli delle pubblicazioni in Scopus e Web of Science*

SCOPUS	WEB OF SCIENCE
Applied Sciences (Switzerland)	APPLIED SCIENCES-BASEL
Sensors (Switzerland)	SENSORS
Computers and Industrial Engineering	COMPUTERS & INDUSTRIAL ENGINEERING

- Presenza di simbologia del tutto inconsistente in relazione al campo considerato, come ad esempio un "+" nel campo relativo alla pagina.

La maggior parte dei difetti riportati sono di tipo sintattico e non semantico e dunque non vanno ad inficiare completamente l'individuazione del record di interesse; resta comunque il fatto che sono errori che andrebbero corretti e che potevano tranquillamente essere evitati tramite appositi meccanismi di quality checking in fase di caricamento dei dati nei database in esame.
La presenza di difetti è infatti sempre e comunque nociva dal momento che ha inevitabilmente un impatto sulle valutazioni e conclusioni che verranno fatte sulla base dei dati.

4. CAPITOLO IV

4.1 Conclusioni

La tesi ha trattato della data quality e delle metodologie che possono essere utilizzate per valutarla in fase di acquisto. La definizione, in termini generali, delle fasi previste per l'assessment ha aperto la strada alla loro applicazione ad un caso di studio specifico, quello dei database bibliometrici, dal momento che si vuole fornire ai loro utenti una prassi oggettiva da seguire nella fase di selezione e successiva accettazione del prodotto da acquistare.

A tal proposito, si è combinato l'uso di standard di riferimento, come la normativa SQuaRE, e di tool qualitativi comunemente impiegati nel contesto produttivo classico, quali il QFD e i piani di campionamento: ciò rappresenta sicuramente una novità in materia di data quality assessment, dove le tecniche attualmente presenti risultano essere più incentrate sulle valutazioni qualitative e soggettive dei dati analizzati.

I due prodotti analizzati sono Web of Science e Scopus, in relazione ai quali si è scelto di valutare le dimensioni di completezza, coerenza ed accuratezza; gli esiti sono stati principalmente negativi, soprattutto quando si sono considerati i DB nella loro interezza.

Questo dipende sicuramente dalle caratteristiche che si è scelto di valutare; tuttavia, nel momento in cui il focus risulti essere sulle informazioni in sé, è evidente come non si possano tralasciare le dimensioni qui considerate.

Le analisi condotte hanno mostrato come la dimensione di completezza sia particolarmente critica anche se, come precedentemente specificato, l'elevata difettosità caratterizzante quest'area potrebbe essere dovuta ad aspetti propri delle pubblicazioni considerate nel campione esaminato (i.e. articoli in press, articoli online, etc).

Anche in relazione alla dimensione di accuratezza si sono riscontrati alcuni problemi, anche se non si proceduto all'applicazione di tale test in relazione a tutti i campioni esaminati. Questi difetti erano per lo più relativi ad aspetti sintattici che potevano essere evitati tramite l'implementazione di meccanismi ad hoc da parte dei provider dei DB bibliometrici.

Va inoltre specificato come i risultati ottenuti non dipendano solo dalla selezione delle dimensioni SQuaRE da valutare, ma anche dal settaggio dei parametri caratteristici del piano di campionamento: nonostante si siano valutati diversi scenari e fatte diverse ipotesi al fine di assegnare dei valori numeri ad AQL e LTPD, questi sono comunque stati definiti senza interpellare direttamente i soggetti coinvolti.

A prescindere dagli esiti, la metodologia ha comunque consentito di mettere in evidenza alcuni degli errori tipici riscontrabili nei database bibliometrici e quindi confermare il catalogo di difetti riportato in altre pubblicazioni in materia.
Ciò dovrebbe rappresentare un ulteriore stimolo per i provider dei DB ad intervenire sui record difettosi individuati e segnalati dagli utenti, al fine di ripristinare la qualità e quindi abbattere la difettosità in uscita rilevata dagli utenti.

Tra gli aspetti innovati che questo lavoro ha consentito di introdurre si citano:

- Utilizzo delle dimensioni SQuaRE come caratteristiche tecniche sulla base delle quali effettuare la valutazione dei database e dei dati in essi contenuti. Ciò consente di far riferimento ad aspetti comunemente riconosciuti e accettati che, in teoria, dovrebbero rappresentare le linee guida che chiunque abbia a che fare con un prodotto informatico dovrebbe seguire.
- Impiego dei tool qualitativi comunemente utilizzati in ambito produttivo in relazione a prodotti informatici/software in generale e ai dati in particolare.
- In aggiunta all'ultimo punto, vale la pena sottolineare l'applicazione dei piani di campionamento ai record di un database.

Tutto ciò è stato d'altra parte fatto al fine di aumentare la soddisfazione delle organizzazioni che concretamente e quotidianamente utilizzano i dati, nonché rendere le stesse e i destinatari finali degli output prodotti sulla base dei dati più fiduciosi in ciò con cui hanno a che fare.

4.2 Sviluppi futuri

Nonostante l'applicazione dei piani di campionamento, in combinazione con lo standard SQuaRE, risulti essere una novità nell'ambito della data quality, possono essere definite ulteriori strade percorribili in materia.

Dal punto di vista dell'ambito di applicazione della metodologia, si è fatta una semplificazione che prevede di considerare dati resi disponibili in forma strutturata in database relazionali; tuttavia, è sempre più comune la diffusione e l'utilizzo di dati non strutturati per i quali dovranno essere trovate apposite regole che consentano di definire comunque delle logiche di ispezione e quindi delle modalità tramite le quali eseguire più automaticamente possibile prove di testing delle dimensioni di interesse.

A tal proposito, sfidante sarebbe sicuramente l'estensione della metodologia ai big data, con l'ausilio di strumenti di ML e AI che consentirebbero, se ben sfruttati, di incrementarne le performance.

A prescindere dalla modalità con cui i dati sono organizzati, cercare di automatizzare il più possibile le prove condotte è un ulteriore obiettivo che varrebbe la pena porsi per snellire e incrementare le prestazioni della metodologia qui trattata: soprattutto in relazione alla prova di accuratezza, è richiesto un massiccio intervento umano che non solo rende la procedura poco efficiente, ma la espone anche ad un'alta probabilità di errore nella costruzione del riferimento assoluto.

Altro aspetto che vale la pena menzionare è il fatto che in tal caso si è adottato il punto di vista di un ipotetico cliente di un database che quindi acquista il prodotto fatto e finito, senza la possibilità di intervenire sui dati in esso memorizzati e quindi modificarli al fine di ridurre la difettosità inizialmente riscontrata.

Allora, nel momento in cui si possa avere accesso anche in scrittura ai dati, potrebbero essere applicati piani di campionamento con rettifica, attraverso i quali cercare di abbattere, attraverso la sostituzione dei record difettosi rilevati nelle diverse prove, la difettosità in uscita. A differenza di un piano con rettifica classico, tuttavia, viste e considerate le dimensioni dei lotti di partenza e le modalità di implementazione delle prove, la sostituzione dei record difettosi verrebbe comunque effettuata solo ed esclusivamente in relazione al campione esaminato, anche qualora si debba rifiutare il lotto.

Interessante sarebbe anche valutare come l'esito delle diverse prove condotte possa modificarsi al variare dell'istante temporale, al fine di riuscire a mettere in evidenza se

effettivamente i DB mettono in atto meccanismi che consentono di ridurre la difettosità sulla base delle segnalazioni effettuate dagli utenti.

Inoltre, l'analisi svolta ha considerato i piani di campionamento per attributi, ma potrebbero anche essere sfruttati piani per variabili dove, in relazione alle dimensioni che si desidera testare, dovrebbero essere definiti appositi limiti di specifica superiore (USL) o inferiore (LSL), a seconda che si imposti il problema in termini di massima difettosità riscontrabile o minimo livello desiderato in relazione alle diverse caratteristiche tecniche.

Dunque, in conclusione, studi futuri dovrebbero essere volti a rilasciare le assunzioni e semplificazioni introdotte nel presente lavoro, in modo da ottenere metodologie sempre più generali e applicabili ad una sempre più vasta gamma di casistiche.

5. RIFERIMENTI

Albano A., Ghelli G., Orsini R., 2019, Fondamenti di basi di dati

ASQ Statistics Division, 2004, Glossary and Tables For Statistical Quality Control, Fourth Edition

Batini C., Scannapieco M., 2003, Data quality: Concepts, Methodologies and Techniques, Springer Verlag

Batini C., Cabitza F., Cappiello C., Francalanci C., 2008, A comprehensive data quality methodology for Web and structured data, Int.J.Innov.Comput.Appl. 1,3,205-218

Batini C., Cappiello C., Francalaci C., Maurino A., 2009, Methodologies for data quality assessment and improvement, ACM Computing Surveys Vol.41, No.3

Che cosa è un database? https://www.oracle.com/it/database/what-is-database/

English L., 1999, Improving Data Warehouse and Business Information Quality, Wiley & Sons

Franceschini F., Maisano D., Mastrogiacomo L., 2016, The museum of errors/horrors in Scopus, Journal of Informetrics, 10, 1, 174-182

Franceschini F., Maisano D., Mastrogiacomo L., 2016, Empirical analysis and classification of database errors in Scopus and Web of Science, Journal of Informetrics, 10,4, 933-953

Franceschini F., Galetto M., Maisano D., Mastrogiacomo L., Ingegneria della qualità Applicazioni ed Esercizi (Quarta Edizione)

Goodman D., Louise D., 2005, Web of Science (2004 version) and Scopus, The Charleston Advisor, 5-21

Goodman D., Louise D., 2006. Update on Scopus and Web of Science, The Charleston Advisor, 15-18

Hawkins D., 1980, Identification of outliers, Chapman and Hall

ISO/IEC 12207:2008, 2008, Systems and software engineering — Software life cycle processes

ISO/IEC 25000:2014, 2014, Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Guide to SQuaRE

ISO/IEC 25012:2008, 2008, Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model

ISO/IEC 25030:2019, 2019, Systems and software engineering - Systems and software quality requirements and evaluation (SQuaRE) - Quality requirements framework

Jeusfeld M., Quix C., Jarke M., 1998, Design and analysis of quality information for data warehouses in Proceedings of the 9th International Conference on Information Quality

Knorr E., Ng R., 1998. Algorithms for mining distance-based outliers in large datasets in Proceedings of the 24th International Conference on Very Large Data Bases

Lee Y., Strong D., Kahn B., Wang R., 2002, AIMQ: A methodology for information quality assessment, Inform. Manage. 40, 2, 133-460

Marcus A., Maletic J., 2000, Utilizing association rules for identification of possible errors in data sets, Technical Report TR-CS-00-04, The University of Memphis, Division of Computer Science

Mongeon P., Paul-Hus A., 2016, The journal coverage of Web of Science and Scopus: A comparative analysis, Scientometrics, 106, 1, 213–228

Orduna-Malea E., Ayllón, J., Martín-Martín A., López-Cózar E., 2015, Methods for estimating the size of Google Scholar, Scientometrics, 104, 3, 931–949

Scopus https://www-scopus-com.ezproxy.biblio.polito.it/search/form.uri?display=basic#basic

UNI ISO 2859-1:2007, 2007, Procedimenti di campionamento nell'ispezione per attributi - Parte 1: Schemi di campionamento indicizzati secondo il limite di qualità accettabile (AQL) nelle ispezioni lotto per lotto

UNI EN ISO 8402:1995, 1995, Gestione per la qualità ed assicurazione della qualità. Termini e definizioni.

UNI EN ISO 9000:2015, 2015, Sistemi di gestione per la qualità - Fondamenti e vocabolario

UNI EN ISO 9001:2015, 2015, Sistemi di gestione per la qualità - Requisiti

UNI CEI ISO/IEC 25024:2016, 2016, Ingegneria del software e di sistema - Requisiti e valutazione della qualità dei sistemi e del software (SQuaRE) - Misurazione della qualità dei dati

USA Departement of Defence, MIL STD 105E – Sampling procedures and tables for inspection by attributes

Salisbury, 2009, Web of Science and Scopus:a Comparative review of Content and Searching Capabilities, The Charleston Advisor

Sporleder C., van Erp M., Porcelijn T., van den Bosch A., 2006, Spotting the 'Odd-one-out': Data-Driven Error Detection and Correction in Textual Databases in Proceedings of the Workshop on Adaptive Text Extraction and Mining

Wang R., 1998, A product perspective on total data quality management

Web Of Science https://www-webofscience-com.ezproxy.biblio.polito.it/wos/woscc/basic-sea

6. APPENDICE

Tabella 6.1*: Campione Prova A per Scopus*

	AUTHORS	TITLE	DOI
1	Sreedharan, V, Sunder, M	A novel approach to lean six sigma project management: a conceptual framework and empirical application	10.1080/09537287.2018.1492042
2	Raffoni, A, Visani, F, Bartolini, M, Silvi, R	Business Performance Analytics: exploring the potential for Performance Management Systems	10.1080/09537287.2017.1381887
3	Medeiros, M, Ferreira, L	Development of a purchasing portfolio model: An empirical study in a Brazilian hospital	10.1080/09537287.2018.1434912
4	Villa, A, Taurino, T	Event-driven production scheduling in SME	10.1080/09537287.2017.1401143
5	Seth, D, Nemani, V, Pokharel, S, Al Sayed, A	Impact of competitive conditions on supplier evaluation: a construction supply chain case study	10.1080/09537287.2017.1407971
6	Kim, B, Kwak, Y	Improving the accuracy and operational predictability of project cost forecasts: an adaptive combination approach	10.1080/09537287.2018.1467511
7	Love, P, Smith, J, Ackermann, F, Irani, Z, Teo, P	The costs of rework: insights from construction and opportunities for learning	10.1080/09537287.2018.1513177
8	Masi, D, Kumar, V, Garza-Reyes, J, Godsell, J	Towards a more circular economy: exploring the awareness, practices, and barriers from a focal firm perspective	10.1080/09537287.2018.1449246

Tabella 6.2: *Campione Prova A per WoS*

	AUTHORS	TITLE	DOI
1	Belhadi, A, Touriki, F, El Fezazi, S	Benefits of adopting lean production on green performance of SMEs: a case study	10.1080/09537287.2018.1490971
2	Erthal, A, Marques, L	National culture and organisational culture in lean organisations: a systematic review	10.1080/09537287.2018.1455233
3	Luo, W, Shi, Y, Venkatesh, V	Exploring the factors of achieving supply chain excellence: a New Zealand perspective	10.1080/09537287.2018.1451004
4	Mishra, J, Hopkinson, P, Tidridge, G	Value creation from circular economy-led closed loop supply chains: a case study of fast-moving consumer goods	10.1080/09537287.2018.1449245
5	Moazzam, M, Akhtar, P, Garnevska, E, Marr, N	Measuring agri-food supply chain performance and risk through a new analytical framework: a case	10.1080/09537287.2018.1522847

		study of New Zealand dairy	
6	Nounou, A	Developing a lean-based holistic framework for studying industrial systems	10.1080/09537287.2018.1517422
7	Plessner, M, Buhren, C, Frank, B	Market research with the aid of a smartphone application - a case study	10.1080/09537287.2017.1391345
8	Vlajic, J, Mijailovic, R, Bogdanova, M	Creating loops with value recovery: empirical study of fresh food supply chains	10.1080/09537287.2018.1449264

Tabella 6.3: *Campione Prova B per Scopus*

	AUTHORS	TITLE	DOI
1	Akwei, C, Zhang, L	Integrating risk and performance management in quality management systems for the development of complex bespoke systems (CBSs)	10.1080/09537287.2018.1520316
2	Altay, N, Gunasekaran, A, Dubey, R, Childe, S	Agility and resilience as antecedents of supply chain performance under moderating effects of organizational culture within the humanitarian setting: a dynamic capability view	10.1080/09537287.2018.1542174
3	Anaya-Arenas, A, Ruiz, A, Renaud, J	Importance of fairness in humanitarian relief distribution	10.1080/09537287.2018.1542157
4	Bai, C, Sarkis, J	Honoring complexity in sustainable supply chain research: a rough set theoretic approach (SI:ResMeth)	10.1080/09537287.2018.1535133
5	Baş Güre, S, Carello, G, Lanzarone, E, Yalçındağ, S	Unaddressed problems and research perspectives in scheduling blood collection from donors	10.1080/09537287.2017.1367860
6	Bhattacharya, A, David, D	An empirical assessment of the operational performance through internal benchmarking: A case of a global logistics firm	10.1080/09537287.2018.1457809
7	Bodaghi, B, Palaneeswaran, E, Abbasi, B	Bi-objective multi-resource scheduling problem for emergency relief operations	10.1080/09537287.2018.1542026
8	El-Akruti, K, Kiridena, S, Dwight, R	Contextualist-retroductive case study design for strategic asset management research	10.1080/09537287.2018.1535145
9	Gijo, E, Palod, R, Antony, J	Lean Six Sigma approach in an Indian auto ancillary conglomerate: a case study	10.1080/09537287.2018.1469801
10	Kiil, K, Dreyer, H, Hvolby, H, Chabada, L	Sustainable food supply chains: the impact of automatic replenishment in grocery stores	10.1080/09537287.2017.1384077
11	Kim, B, Kwak, Y	Improving the accuracy and operational predictability of project	10.1080/09537287.2018.1467511

		cost forecasts: an adaptive combination approach	
12	Liu, H, Love, P, Smith, J, Irani, Z, Hajli, N, Sing, M	From design to operations: a process management life-cycle performance measurement system for Public-Private Partnerships	10.1080/09537287.2017.1382740
13	Love, P, Smith, J, Ackermann, F, Irani, Z	The praxis of stupidity: an explanation to understand the barriers mitigating rework in construction	10.1080/09537287.2018.1518551
14	Moazzam, M, Akhtar, P, Garnevska, E, Marr, N	Measuring agri-food supply chain performance and risk through a new analytical framework: a case study of New Zealand dairy	10.1080/09537287.2018.1522847
15	Raval, S, Kant, R, Shankar, R	Lean Six Sigma implementation: modelling the interaction among the enablers	10.1080/09537287.2018.1495773
16	Shamsuzzaman, M, Alzeraif, M, Alsyouf, I, Khoo, M	Using Lean Six Sigma to improve mobile order fulfilment process in a telecom service sector	10.1080/09537287.2018.1426132
17	Villa, A, Taurino, T	Event-driven production scheduling in SME	10.1080/09537287.2017.1401143
18	Yang, M, Smart, P, Kumar, M, Jolly, M, Evans, S	Product-service systems business models for circular supply chains	10.1080/09537287.2018.1449247
19	Zhang, J, Chen, X, Fang, C	Transmission of a supplier's disruption risk along the supply chain: a further investigation of the Chinese automotive industry	10.1080/09537287.2018.1470268
20	Ziaee Bigdeli, A, Baines, T, Schroeder, A, Brown, S, Musson, E, Guang Shi, V, Calabrese, A	Measuring servitization progress and outcome: the case of 'advanced services'	10.1080/09537287.2018.1429029

Tabella 6.4: *Campione Prova B per WoS*

	Authors	**Title**	**DOI**
1	Akwei, C; Zhang, L	Integrating risk and performance management in quality management systems for the development of complex bespoke systems (CBSs)	10.1080/09537287.2018.1520316
2	Batista, L; Bourlakis, M; Liu, Y; Smart, P; Sohal, A	Supply chain operations for a circular economy	10.1080/09537287.2018.1449267
3	Braglia, M; Castellano, D; Frosolini, M; Gallo, M	Overall material usage effectiveness (OME): a structured indicator to measure the effective material usage within manufacturing processes	10.1080/09537287.2017.1395920
4	Caffieri, J; Love, P; Whyte, A; Ahiaga-Dagbui, D	Planning for production in construction: controlling costs in major capital projects	10.1080/09537287.2017.1376258
5	Dwaikat, N; Money, A; Behashti, H; Salehi-Sangari,	How does information sharing affect first-tier suppliers'	10.1080/09537287.2017.1420261

	E	flexibility? Evidence from the automotive industry in Sweden	
6	Egbunike, O; Purvis, L; Naim, M	A systematic review of research into the management of manufacturing capabilities	10.1080/09537287.2018.1535147
7	Goncalves, P; Castaneda, J	Stockpiling supplies for disaster response: an experimental analysis of prepositioning biases	10.1080/09537287.2018.1542173
8	Heaslip, G; Kovacs, G; Haavisto, I	Innovations in humanitarian supply chains: the case of cash transfer programmes	10.1080/09537287.2018.1542172
9	Kelly, S; Wagner, B; Ramsay, J	Opportunism in buyer-supplier exchange: a critical examination of the concept and its implications for theory and practice	10.1080/09537287.2018.1495772
10	Macchion, L; Da Giau, A; Caniato, F; Caridi, M; Danese, P; Rinaldi, R; Vinelli, A	Strategic approaches to sustainability in fashion supply chain management	10.1080/09537287.2017.1374485
11	Romero-Silva, R; Santos, J; Hurtado, M	A note on defining organisational systems for contingency theory in OM	10.1080/09537287.2018.1535146
12	Samson, D; Gloet, M	Integrating performance and risk aspects of supply chain design processes	10.1080/09537287.2018.1520314
13	Smith, M; Paton, S; MacBryde, J	Lean implementation in a service factory: views from the front-line	10.1080/09537287.2017.1418455
14	Srai, J; Tsolakis, N; Kumar, M; Bam, W	Circular supply chains and renewable chemical feedstocks: a network configuration analysis framework	10.1080/09537287.2018.1449263
15	Sreedharan, V; Sunder, M	A novel approach to lean six sigma project management: a conceptual framework and empirical application	10.1080/09537287.2018.1492042
16	Tezel, A; Koskela, L; Aziz, Z	Lean thinking in the highways construction sector: motivation, implementation and barriers	10.1080/09537287.2017.1412522
17	Viveros, P; Nikulin, C; Lopez-Campos, M; Villalon, R; Crespo, A	Resolution of reliability problems based on failure mode analysis: an integrated proposal applied to a mining case study	10.1080/09537287.2018.1520293
18	Yadav, G; Seth, D; Desai, T	Application of hybrid framework to facilitate lean six sigma implementation: a manufacturing company case experience	10.1080/09537287.2017.1402134
19	Yang, M; Smart, P; Kumar, M; Jolly, M; Evans, S	Product-service systems business models for circular supply chains	10.1080/09537287.2018.1449247
20	Zhang, M; Lettice, F; Chan, H; Nguyen, H	Supplier integration and firm performance: the moderating effects of internal integration and trust	10.1080/09537287.2018.1474394

Tabella 6.5: *Campione Prova C per Scopus*

	Authors	Title	DOI
1	Zhao, C., Yang, S., McCann, J.A.	On the Data Quality in Privacy-Preserving Mobile Crowdsensing Systems with Untruthful Reporting	10.1109/TMC.2019.2943468
2	Mak, H.W.L., Lam, Y.F.	Comparative assessments and insights of data openness of 50 smart cities in air quality aspects	10.1016/j.scs.2021.102868
3	Mahalakshmi, P., Kaul, A., Aggarwal, Y.	Comparative analysis of frequent pattern mining algorithm on water quality data	
4	Jiang, S., Ni, C., Chen, G., Liu, Y.	A novel data fusion strategy based on multiple intelligent sensory technologies and its application in the quality evaluation of Jinhua dry-cured hams	10.1016/j.snb.2021.130324
5	Shu, Q., Lu, S., Xia, M., Ding, J., Niu, J., Liu, Y.	Enhanced feature extraction method for motor fault diagnosis using low-quality vibration data from wireless sensor networks	10.1088/1361-6501/ab5cca
6	Alshalalfeh, A., Shalalfeh, L.	Bearing fault diagnosis approach under data quality issues	10.3390/app11073289
7	Ma, Q., Li, H., Thorstenson, A.	A big data-driven root cause analysis system: Application of Machine Learning in quality problem solving	10.1016/j.cie.2021.107580
8	Zavoleas, Y., Hank Haeusler, M., Dunn, K., Bishop, M., Dafforn, K., Schaefer, N., Sedano, F., Daniel Yu, K.	Designing bio-shelters: Improving water quality and biodiversity in the bays precinct through dynamic data-driven approaches	10.14627/537690053
9	Phadikar, A., Mandal, H., Chiu, T.-L.	A novel QIM data hiding scheme and its hardware implementation using FPGA for quality access control of digital image	10.1007/s11042-019-08392-5
10	Yu, Y., Yu, J.J.Q., Li, V.O.K., Lam, J.C.K.	A Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data	10.1109/ACCESS.2020.2988684
11	Rezaie-Balf, M., Attar, N.F., Mohammadzadeh, A., Murti, M.A., Ahmed, A.N., Fai, C.M., Nabipour, N., Alaghmand, S., El-Shafie, A.	Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach	10.1016/j.jclepro.2020.122576
12	Sun, C., Yin, Y., Kang, H., Ma, H.	A Quality-Related Fault Detection Method Based on the Dynamic Data-Driven Algorithm for Industrial Systems	10.1109/TASE.2021.3139766
13	Yaghoubi, B., Hosseini, S.A., Nazif, S., Daghighi, A.	Development of reservoir's optimum operation rules considering water quality issues and climatic change data analysis	10.1016/j.scs.2020.102467
14	Soomets, T., Uudeberg, K.,	Validation and comparison of	10.3390/s20030742

	Jakovels, D., Brauns, A., Zagars, M., Kutser, T.	water quality products in baltic lakes using sentinel-2 msi and sentinel-3 OLCI data	
15	Fizza, K., Jayaraman, P.P., Banerjee, A., Georgakopoulos, D., Ranjan, R.	Evaluating Sensor Data Quality in Internet of Things Smart Agriculture Applications	10.1109/MM.2021.3137401
16	Schwartz, Y., Godoy-Shimizu, D., Korolija, I., Dong, J., Hong, S.M., Mavrogianni, A., Mumovic, D.	Developing a Data-driven school building stock energy and indoor environmental quality modelling method	10.1016/j.enbuild.2021.111249
17	Kim, I.S., Shen, Y.-D.	Study on the welding quality using big data	10.5302/J.ICROS.2021.21.0183
18	Zhang, C.W., Pan, R., Goh, T.N.	Reliability assessment of high-Quality new products with data scarcity	10.1080/00207543.2020.1758355
19	Chen, L., Wu, M.	Intelligent workshop quality data integration and visual analysis platform design [智能车间质量数据集成及可视化分析平台设计]	10.13196/j.cims.2021.06.010
20	Fiok, K., Karwowski, W., Gutierrez-Franco, E., Liciaga, T., Belmonte, A., Capobianco, R., Saeidi, M.	Automated Detection of Leadership Qualities Using Textual Data at the Message Level	10.1109/ACCESS.2021.3072372

Tabella 6.6: *Campione Prova C per WoS*

	Authors	***Title***	***DOI***
1	Zhao, C., Yang, S., McCann, J.A.	On the Data Quality in Privacy-Preserving Mobile Crowdsensing Systems with Untruthful Reporting	10.1109/TMC.2019.2943468
2	Mak, H.W.L., Lam, Y.F.	Comparative assessments and insights of data openness of 50 smart cities in air quality aspects	10.1016/j.scs.2021.102868
3	Mahalakshmi, P., Kaul, A., Aggarwal, Y.	Comparative analysis of frequent pattern mining algorithm on water quality data	
4	Jiang, S., Ni, C., Chen, G., Liu, Y.	A novel data fusion strategy based on multiple intelligent sensory technologies and its application in the quality evaluation of Jinhua dry-cured hams	10.1016/j.snb.2021.130324
5	Shu, Q., Lu, S., Xia, M., Ding, J., Niu, J., Liu, Y.	Enhanced feature extraction method for motor fault diagnosis using low-quality vibration data from wireless sensor networks	10.1088/1361-6501/ab5cca
6	Alshalalfeh, A., Shalalfeh, L.	Bearing fault diagnosis approach under data quality issues	10.3390/app11073289
7	Ma, Q., Li, H., Thorstenson,	A big data-driven root cause	10.1016/j.cie.2021.107580

	A.	analysis system: Application of Machine Learning in quality problem solving	
8	Zavoleas, Y., Hank Haeusler, M., Dunn, K., Bishop, M., Dafforn, K., Schaefer, N., Sedano, F., Daniel Yu, K.	Designing bio-shelters: Improving water quality and biodiversity in the bays precinct through dynamic data-driven approaches	10.14627/537690053
9	Phadikar, A., Mandal, H., Chiu, T.-L.	A novel QIM data hiding scheme and its hardware implementation using FPGA for quality access control of digital image	10.1007/s11042-019-08392-5
10	Yu, Y., Yu, J.J.Q., Li, V.O.K., Lam, J.C.K.	A Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data	10.1109/ACCESS.2020.2988684
11	Rezaie-Balf, M., Attar, N.F., Mohammadzadeh, A., Murti, M.A., Ahmed, A.N., Fai, C.M., Nabipour, N., Alaghmand, S., El-Shafie, A.	Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach	10.1016/j.jclepro.2020.122576
12	Sun, C., Yin, Y., Kang, H., Ma, H.	A Quality-Related Fault Detection Method Based on the Dynamic Data-Driven Algorithm for Industrial Systems	10.1109/TASE.2021.3139766
13	Yaghoubi, B., Hosseini, S.A., Nazif, S., Daghighi, A.	Development of reservoir's optimum operation rules considering water quality issues and climatic change data analysis	10.1016/j.scs.2020.102467
14	Soomets, T., Uudeberg, K., Jakovels, D., Brauns, A., Zagars, M., Kutser, T.	Validation and comparison of water quality products in baltic lakes using sentinel-2 msi and sentinel-3 OLCI data	10.3390/s20030742
15	Fizza, K., Jayaraman, P.P., Banerjee, A., Georgakopoulos, D., Ranjan, R.	Evaluating Sensor Data Quality in Internet of Things Smart Agriculture Applications	10.1109/MM.2021.3137401
16	Schwartz, Y., Godoy-Shimizu, D., Korolija, I., Dong, J., Hong, S.M., Mavrogianni, A., Mumovic, D.	Developing a Data-driven school building stock energy and indoor environmental quality modelling method	10.1016/j.enbuild.2021.111249
17	Kim, I.S., Shen, Y.-D.	Study on the welding quality using big data	10.5302/J.ICROS.2021.21.0183
18	Zhang, C.W., Pan, R., Goh, T.N.	Reliability assessment of high-Quality new products with data scarcity	10.1080/00207543.2020.1758355
19	Chen, L., Wu, M.	Intelligent workshop quality data integration and visual analysis platform design [智能车间质量数据集成及可视化分析平台设计]	10.13196/j.cims.2021.06.010
20	Fiok, K., Karwowski, W.,	Automated Detection of	10.1109/ACCESS.2021.3072372

Gutierrez-Franco, E., Liciaga, T., Belmonte, A., Capobianco, R., Saeidi, M.	Leadership Qualities Using Textual Data at the Message Level

7. ALLEGATI

ALLEGATO 1: Applicazione del QFD per la definizione delle caratteristiche tecniche da testare con la metodologia proposta

ALLEGATO 2: Costruzione del campione 1 per la valutazione di Scopus, nel caso di considerazione del database nella sua interezza

ALLEGATO 3: Applicazione di un piano di campionamento ad hoc al campione 1

ALLEGATO 4: Applicazione di un piano di campionamento costruito in accordo alla MIL STD 105E al campione 1

ALLEGATO 5: Costruzione del campione 2 per la valutazione di WoS, nel caso di considerazione del database nella sua interezza

ALLEGATO 6: Applicazione di un piano di campionamento ad hoc al campione 2

ALLEGATO 7: Applicazione di un piano di campionamento costruito in accordo alla MIL STD 105E al campione 2

ALLEGATO 8: Applicazione della prova A nel caso di use case 1 e Scopus

ALLEGATO 9: Applicazione della prova B nel caso di use case 1 e Scopus

ALLEGATO 10: Applicazione della prova A nel caso di use case 1 e WoS

ALLEGATO 11: Applicazione della prova B nel caso di use case 1 e WoS

ALLEGATO 12: Applicazione della prova C nel caso di use case 2 e Scopus

ALLEGATO 13: Applicazione della prova D nel caso di use case 2 e Scopus

ALLEGATO 14: Applicazione della prova C nel caso di use case 2 e WoS

ALLEGATO 15: Applicazione della prova C nel caso di use case 2 e WoS

8. RINGRAZIAMENTI

Grazie alla TERRA fertile che mi ha fatto crescere e che ha saputo far germogliare al meglio le mie scelte.
Ricca di amore, rispetto, valori, ascolto, mi ha permesso di piantare solide radici e di crescere in altezza e ampiezza per cercare di raggiungere qualsiasi cielo avessi il desiderio di toccare.
Grazie all'ARIA fresca che mi circonda e che mi fa respirare quella felicità e libertà che sono impossibili da descrivere a parole: sa farmi ridere, piangere, divertire, riflettere.
Mi ha imparato da un lato a resistere e dall'altro a prendere fiato.
Ci sono venti stabili che mi portano con loro da tempo e altri che invece mi hanno travolto solo recentemente: hanno saputo sciogliere i miei nodi e aiutato a capire da dove venissi e dove volessi andare.
Grazie al FUOCO della conoscenza che da sempre mi arde dentro.
Una fiamma viva che non ho mai smesso di alimentare grazie all'incontro con tutti coloro che hanno nelle loro mani l'abilità di controllarla e farla crescere in grandezza ed intensità.
La fortuna ha voluto che incontrassi grandi ed uniche fiamme che hanno buttato legna al mio falò e fatto capire di non smettere mai di incendiarmi per ciò in cui credo.
Come l'ACQUA che cambia e resta sempre la stessa, vorrei rendere ogni giorno un'occasione per essere sempre più la persona che voglio essere: gli ostacoli lungo il percorso, i ruscelli che in me confluiscono, tutti coloro che desiderano bere dalla mia fonte sono al tempo stesso un'occasione ed un regalo che non può che rendermi ogni giorno più ricca.