

La conservazione delle basi di dati

Aprile 2023

Sommario

Premessa	4
Obiettivi del documento	4
La preservazione delle informazioni digitali	6
1.1. La conservazione delle risorse digitali: Il modello OAIS	8
La conservazione delle informazioni strutturate di rilevanza giuridica, documentale e archivistica	10
1.2. Il punto di vista giuridico nel quadro normativo europeo e nazionale	11
1.3. La conservazione permanente delle basi di dati a fini storici e scientifici	13
1.4. Forme di rappresentazione e <i>digital curation</i> delle basi di dati	15
La conservazione dei dati transazionali	20
1.5. I dati transazionali: loro caratteristiche e problematiche di conservazione a lungo termine .	20
1.6. Metodi per la conservazione delle basi di dati a lungo termine	21
1.6.1. Staticizzazione	21
1.6.2. Emulazione	22
1.6.3. Migrazione	22
1.6.4. Normalizzazione	23
1.6.5. Migrazione/normalizzazione di basi di dati relazionali.....	23
1.6.6. Analisi del contesto internazionale per la conservazione delle basi di dati	24
1.7. Razionale per la valutazione della complessità del DB origine per selezionare una soluzione per la conservazione nel tempo e vincoli di cui tenere conto	26
1.7.1. La struttura della documentazione della base di dati in esercizio	26
1.7.2. Documentazione della business logic codificata nella base di dati in esercizio	27
1.7.3. Documentazione dei casi d'uso della base di dati in esercizio.	28
1.7.4. Informazioni statistiche sulla base di dati in esercizio.....	29
1.7.5. Aspetti tecnologici della base di dati di esercizio	29
1.7.6. Modalità di utilizzo futuro della base di dati storicizzata	30
1.7.7. Valutazione della qualità dello schema e dei dati delle basi di dati da storicizzare.....	30
La conservazione di basi dati secondo lo standard SIARD	32
1.8. Funzionalità specifiche del software	33
1.9. Testing di importazione da SQL Server	33
La conservazione dei registri e dei repertori	34

1.10.	Conservazione dei repertori informatici notarili	34
1.11.	Conservazione dei registri di protocollo.....	37
1.12.	Conservazione dei verbali di esame	40
Allegato - Sperimentazione di SIARD		42
1.13.	Basi di dati utilizzate	42
1.14.	Test di normalizzazione	43
1.15.	Importazione da file SIARD	45
1.15.1.	Destinazione: SQLServer	45
1.15.2.	Destinazione: MySQL:	48
1.15.3.	Destinazione Oracle:	49
1.15.4.	Destinazione PostgreSQL:	51

Premessa

Nell'ambito del Gruppo di lavoro coordinato da AgID sui Poli di Conservazione, è stato attivato un sottogruppo specifico che si è posto, tra gli obiettivi, quello di una prima riflessione sulla “preservazione delle basi di dati”. Tale sottogruppo, coordinato dal prof. Antonino Mazzeo dell'Università Federico II di Napoli, a valle di numerose riunioni di approfondimento sul tema, ha redatto il presente studio, a cui hanno contribuito numerosi esperti delle aree archivistica, informatica e giuridica, e membri del Gruppo di lavoro AgID sui Poli di Conservazione, nella veste di esperti di dominio. In particolare, si riportano di seguito i nominativi di coloro che hanno contribuito alla redazione del presente documento: dott.ssa Patrizia Gentili ed ing. Luigi Avena (AGID); proff. Sonia Bergamaschi e Luca Gagliardelli (Univ. Di Modena e Reggio Emilia), proff. Sergio Flesca e Domenico Talia (Univ. Della Calabria), prof.ssa Mariella Guercio (Univ. Di Roma La Sapienza), prof. Adriano Peron (Univ. Di Trieste), dott. Gabriele Bezzi (Polo Archivistico Emilia Romagna), Costantino Landino (Istituto Centrale degli Archivi). Come esperti di dominio partecipanti al gruppo di lavoro sui poli di conservazione hanno, inoltre, contribuito i dottori Cinzia Amici (Regione Marche), Caterina Pugliese e Francesca Quai (SOGEI) e Francesca Siliprandi (Comune di Verona).

Obiettivi del documento

Stante la diffusione dei servizi digitali e delle tecnologie informatiche disponibili su larga scala e in continua crescita, la “memoria” della nostra società si baserà sempre più sulla selezione e conservazione di enormi quantità di informazioni digitali di vario tipo, rendendo quindi sempre più importante il tema generale della “preservazione¹ della memoria del tempo in cui viviamo”. La preservazione delle informazioni è, certamente, una delle aree più sfidanti che stanno impegnando e che impegneranno sempre più i Governi, al fine di sviluppare soluzioni tecnologiche, normative e organizzative per la memorizzazione di tale enorme patrimonio di informazioni e mettere in essere concrete “buone pratiche”.

Il fine è di sensibilizzare tutti gli attori coinvolti sulle tematiche della preservazione delle informazioni ed il legislatore per aggiornare la vigente normativa sulla conservazione, incentrata principalmente sul documento informatico, verso una più generale regolamentazione della tenuta nel tempo delle varie tipologie di informazioni digitali (dati trattati nei sistemi informatici degli enti pubblici e privati, nei web e nei social, etc.).

In Italia, la tematica della preservazione delle informazioni è, da tempo, seguita principalmente da gruppi di ricerca ed esperti, operanti nelle università e centri nazionali di archivistica e informatica. Tali gruppi operano in seno a progetti nazionali, internazionali o comunitari sul tema della preservazione e partecipano anche alle attività di normazione portate avanti da specifiche organizzazioni di *preservation coalition*.

¹ In questo documento si fa uso del termine preservazione per indicare la tenuta delle informazioni nelle sue varie accezioni (in inglese *information preservation*) e del termine conservazione per la tenuta dei documenti informatici e delle basi di dati così come utilizzato dalla normativa italiana.

Limitata appare, invece, l'attività svolta in tal senso dalle istituzioni, che interessata principalmente ai temi dell'e-government, ha privilegiato la regolamentazione dei processi amministrativi basati solo sul documento informatico e non, come detto, la preservazione delle informazioni in tutte le varie rappresentazioni che oggi l'informatica e internet ci rendono disponibili.

Da tale scenario ne consegue che gli enti pongono, allo stato, poco interesse alla definizione e pianificazione delle azioni miranti alla preservazione del patrimonio informativo e, in special modo, dei dati trattati nei sistemi transazionali presenti nei loro sistemi informatici. Si evidenzia, quindi, necessità di costituire al più presto gruppi di competenza multidisciplinari sulla preservazione dell'informazione e, in particolare, sulla conservazione delle basi di dati relazionali a supporto dei sistemi transazionali che vedono coinvolti come *stakeholder* sia enti pubblici sia quelli privati, al fine di cominciare a fornire nel breve concrete risposte per evitare la perdita delle informazioni.

Il presente studio, con riferimento anche al contesto internazionale, si propone di fornire un primo quadro sintetico sulla preservazione delle informazioni in generale e, nello specifico, sulla conservazione dei dati trattati nei sistemi transazionali. Lo studio intende sensibilizzare i vari soggetti interessati a tale tematica, fornendo alcune indicazioni di natura metodologica e operativa su come approcciare e pianificare la problematica della conservazione delle basi di dati utilizzate nei sistemi informatici.

Il documento, dopo una premessa generale sulla preservazione delle diverse tipologie di informazioni di interesse, illustra nel dettaglio le principali problematiche relative alla conservazione delle basi dati relazionali e alla digitalizzazione e alla tenuta di registri e repertori, evidenziandone le possibili soluzioni e riportando taluni significativi casi d'uso. Il documento è organizzato in cinque capitoli e un allegato, in particolare nei vari capitoli sono trattati i seguenti argomenti:

- capitolo 1 “La preservazione delle informazioni digitali” inquadra, con riferimento allo scenario internazionale, il generale problema della preservazione delle informazioni digitali, evidenziandone le esigenze, l'importanza, le problematiche tecniche, le difficoltà e criticità di trattamento. Presenta, inoltre, gli aspetti salienti dello standard OAIS ampiamente utilizzato come modello di riferimento per la conservazione degli oggetti digitali;
- capitolo 2 “La conservazione delle informazioni strutturate di rilevanza giuridica, documentale e archivistica” tratta, con riferimento al panorama italiano e alla normativa sul documento informatico, degli aspetti salienti, di natura giuridico-archivistica, della conservazione delle informazioni e, in particolare, della conservazione delle basi di dati;
- capitolo 3 “La conservazione dei dati transazionali” affronta le problematiche informatiche legate alla conservazione delle basi di dati e, in particolare, delle basi di dati relazionali, illustrando i principali metodi e i principali progetti sviluppati di recente in ambito internazionale e disponibili anche come applicazioni commerciali;
- capitolo 4 “La conservazione di basi dati secondo lo standard SIARD” si focalizza sull'uso di SIARD, descrivendone il modello e il formato, le funzionalità principali e i risultati delle sperimentazioni effettuate applicate a concreti casi d'uso (i dettagli di tali sperimentazioni, sono riportati nell'allegato);
- capitolo 5 “La conservazione dei registri e dei repertori” approfondisce gli aspetti della digitalizzazione dei registri e dei repertori, di fatto sono assimilabili a semplici basi di dati), evidenziandone i vantaggi in termini di dinamicità ed efficienza ottenibili. In particolare, sono riportati esempi di digitalizzazione dei repertori notarili e dei verbali digitali degli esami

universitari e la digitalizzazione dei registri di protocollo alla luce dei vincoli giuridici e organizzativi posti dalla normativa vigente che li regola.

La preservazione delle informazioni digitali

I processi conservativi presentano un'elevata complessità sia per la pluralità delle forme di produzione dei contenuti digitali da parte delle organizzazioni e delle persone sia per la differenziazione delle modalità di rappresentazione digitale e di condivisione e trasmissione, ma anche per la presenza di numerosi standard tecnici specifici. La conservazione a lungo termine dei contenuti digitali è un processo che deve partire dalla fase della loro formazione al fine di assicurarne in modo sostenibile e accurato l'autenticità, l'affidabilità e l'accesso nel tempo.

Conservare risorse digitali di qualunque natura, soprattutto se hanno valore giuridico in sé o contengono dati e relazioni giuridicamente rilevanti (tali cioè da richiedere interventi atti ad assicurarne valore legale ed efficacia probatoria nel tempo), implica un'accurata documentazione dei processi specifici relativi alla loro formazione (informazioni del contesto di produzione e provenienza) e alle fasi di gestione e di eventuale trasformazione che si sono rese necessarie per il loro trattamento². Negli ultimi anni, si è inoltre precisato anche nel nostro ordinamento giuridico che non solo i documenti (nelle varie accezioni definite nel CAD e nelle Linee guida sulla formazione, gestione e conservazione dei documenti informatici, di seguito Linee guida AgID), ma anche i dati debbano essere oggetto di trattamento adeguato in termini di formazione e conservazione, assicurando la persistenza di caratteristiche di integrità, autenticità e accuratezza.

Non si fa qui riferimento esclusivamente ai metadati che devono accompagnare la produzione dei documenti e la loro tenuta nel tempo ai sensi dell'allegato 5 delle Linee guida AgID, ma si considerano anche i dati organizzati in forma strutturata, presenti ormai da decenni tra i patrimoni di informazioni generate in ambito sia pubblico sia privato e identificati con l'espressione inclusiva di basi di dati.

Il tema della conservazione delle basi di dati in realtà è da tempo oggetto di attenzione da parte degli archivisti e, in generale, delle istituzioni di memoria, a livello internazionale e nazionale. Basti ricordare l'iniziativa condotta nel 2003 sul tema della conservazione delle basi di dati dai ricercatori del progetto europeo ERPANET (1999-2004).

Già allora, vent'anni fa, si riconosceva l'importanza delle basi di dati e la necessità di un investimento adeguato destinato alla loro conservazione, tanto più indispensabile quanto più

² In relazione alla normativa utile cfr. Progetto Poli di conservazione - definizione di un modello di riferimento per i Poli di Conservazione e della relativa rete nazionale, 22 giugno 2021, <https://www.agid.gov.it/it/piattaforme/conservazione/poli-conservazione>. Per la letteratura di riferimento e i progetti realizzati cfr. L. Duranti, R. Preston, *International research on permanent authentic records in electronic systems (InterPARES) 2: experiential, interactive, dynamic records*, Padova, Cleup, 2008; C. Landino e P. A. Marzotti, *Memorie dinamiche. La conservazione dei database e il web archiving*, Torino, Edizioni ANAI, 2018. Per una guida ai problemi tecnici cfr. Artefactual Systems and Digital Preservation Coalition, *Preserving Databases*, July 2021, <http://doi.org/10.7207/twgn21-06>.

complessi erano e sono i nodi conservativi da affrontare, considerata la loro natura dinamica sia nell'accumulazione dei contenuti sia nelle modalità flessibili che devono essere garantite a fini di interrogazione e fruizione. Il rapporto conclusivo del seminario organizzato nel 2003 a Berna, nell'ambito del citato progetto ERPANET, concludeva riconoscendo che la loro conservazione avrebbe comportato una serie elevata di problemi unici e impegnativi (<https://www.erpanet.org>). Così unici e impegnativi che ben poco è stato pianificato e, ancor meno, realizzato da allora, anche perché le tecnologie abilitanti o non erano mature o economicamente insostenibili. Anzi, gran parte delle domande che si posero i ricercatori nei primi anni 2000 sono le stesse cui si è cercato di dare risposta in questa sede. Esse riguardano le strategie e le policy che dovrebbero guidare la conservazione archivistica delle basi di dati, la struttura e le fasi del processo di conservazione, i problemi della selezione, la qualità e quantità della documentazione da acquisire, le modalità e i profili di accesso e infine, ma non meno importante, le forme possibili e opportune per la cooperazione interdisciplinare, in particolare tra archivisti e informatici.

Il tempo passato ha, casomai, accresciuto la rilevanza e l'urgenza delle questioni allora sollevate. Non sono mancate del tutto esperienze e progetti che, tuttavia, non hanno finora condotto a risposte esaurienti e consolidate. Gli standard all'epoca appena approvati si sono, comunque, anche se ancora parzialmente, rafforzati. Si evidenzia, in particolare, il modello OAIS che costituisce il punto di partenza anche di questo lavoro, poiché esso offre un quadro logico di riferimento, in grado di considerare e trattare, in modo aperto e allo stesso tempo strutturato, tutte le possibili tipologie di informazioni che la tecnologia attuale e futura ci consente di produrre e sedimentare.

Al fine di completare il quadro internazionale di riferimento, merita infine sottolineare che il tema è stato oggetto di interesse da parte della comunità internazionale e della Commissione europea. Quest'ultima ha costituito un organismo indipendente denominato "Digital Information Life Cycle Interoperability Standards Board (DILCIS Board)"³ con il compito di sviluppare e mantenere specifiche tecniche pubbliche, implementazioni personalizzate e casi di studio. In particolare, il DILCIS Board ha sviluppato il progetto E-ARK4ALL che costituisce il principale punto di riferimento europeo per l'implementazione di soluzioni applicative per l'e-archiving, individuando la soluzione al problema della conservazione e dell'accesso ai dati per raggiungere una ampia interoperabilità fra soluzioni. Fra le linee guida proposte e di interesse per l'ambito di questo studio sono, inoltre, da considerare le "Specification for the E-ARK Content Information Type Specification for Electronic Records Management Systems (CITS ERMS)"⁴ e le specifiche per il formato SIARD⁵, progetto derivato da ERPANET.

³ <https://dilcis.eu/>

⁴ <https://dilcis.eu/content-types/cserms>

⁵ <https://dilcis.eu/content-types/siard>

1.1. La conservazione delle risorse digitali: Il modello OAIS

Lo standard OAIS⁶ costituisce da molti anni (ufficialmente dal 2003 in quanto standard ISO 14721:2003, in realtà dal 1997 come raccomandazione del CCSDS – Consultative Committee for Space Data System) il modello generale di riferimento per la conservazione di qualunque tipo di risorsa digitale (dagli oggetti più semplici formati da singoli file a quelli ben più impegnativi costituiti dai siti web, dai software, dalle basi di dati di cui si tratta nel presente studio).

L'utilità di questo modello longevo e universalmente apprezzato e adottato, rivisto nel 2012 e oggetto di una nuova edizione in corso di approvazione in sede ISO, deriva dalla sua efficacia nel definire sia gli aspetti logici sia quelli fisici della conservazione, consentendo allo stesso tempo di dettagliare ogni elemento delle risorse da conservare senza trascurare i vincoli e le relazioni che ne garantiscono l'intelligibilità, oltre a fornire strumenti di verifica della loro integrità e autenticità.

Non è questa la sede per descrivere la norma ISO nella sua interezza, né per riflettere sulle peculiarità che sono all'origine del suo successo. In questo ambito merita, tuttavia, sottolineare le qualità principali che consentono all'architettura OAIS di sostenere con efficienza anche il processo di conservazione di oggetti molto complessi quali, ad esempio le basi di dati. In primo luogo, la struttura definita nel modello permette di inserire in un quadro generale, inclusivo e ricorsivo, anche l'analisi degli elementi minimi da individuare e trattare in un processo conservativo: è possibile gestire i singoli frammenti dell'intera struttura in modo rigoroso senza rinunciare a tenere traccia di tutti i vincoli che identificano e qualificano l'insieme. Il vincolo principale di cui tenere conto è la definizione dei bisogni e delle capacità di comprensione degli utenti, indicati con il termine di *comunità di riferimento*, la cui definizione assume quindi rilievo centrale: un obbligo tassativo per chiunque progetti e gestisca sistemi e servizi di conservazione.

Ai fini della conservazione a lungo termine delle basi di dati, alcuni concetti e termini fondamentali di OAIS sono di indubbia utilità, non solo perché adottati anche dal legislatore italiano nella definizione del sistema di conservazione nazionale, ma anche e soprattutto perché individuano e descrivono le principali componenti da considerare a fini conservativi per questa specifica tipologia di risorse.

I termini e i principi definiti da OAIS e ritenuti centrali in questo lavoro riguardano i seguenti elementi che saranno più volte citati nelle sezioni e capitoli che seguono e che meritano quindi di essere qui riportati:

- *Data object*: oggetto dati inteso semplicemente come oggetto costituito da dati; può avere natura digitale ma anche materiale (in quest'ultimo caso è un oggetto con proprietà fisiche osservabili); potrebbe essere un bitstream;
- *Information object*: oggetto informativo inteso come insieme di *Data object* e *Representation information*; è il contenuto informativo che costituisce l'obiettivo originario della conservazione reso comprensibile all'utente; in un processo conservativo ogni *Data object* dovrà essere sempre

⁶ Lo standard nella versione attuale del 2012 è gratuitamente disponibile nella forma di raccomandazione del CCSDS – Consultative Committee for Space Data System al sito <https://public.ccsds.org/Pubs/650x0b1s.pdf>. La versione approvata come standard ISO è a pagamento: ISO 14721:2012 Space data and information transfer systems — Open archival information system (OAIS) — Reference model.

interpretato attraverso informazioni significative associate (per esempio la tabella ASCII che descrive il rapporto tra la sequenza di bit costituita dal *Data Object* e il simbolo indicato nella tabella medesima);

- *Information package*: il pacchetto informativo inteso come aggregato logico che include l'*Information object*, le successive informazioni sulla conservazione (*Preservation description information – PDI*) e le informazioni che identificano e collegano le componenti destinate a essere gestite nei processi di conservazione (per esempio le informazioni sulle directory). Gli *information package* sono di tre tipi di versamento (SIP), di archiviazione (AIP) e di distribuzione (DIP)⁷.
- *Preservation description information – PDI*: le informazioni necessarie alla conservazione dell'*Information object* che ogni sistema di conservazione deve associare per garantire alla comunità di riferimento la comprensione dei contenuti informativi; includono informazioni relative a *Provenance, Reference, Context, Fixity, Rights*;
- *Representation information*: le informazioni che associano a ogni *Data object* concetti significativi di diversa natura (informazioni strutturali, semantiche, software o altro ancora) in grado di rendere comprensibile il contenuto informativo destinato alla conservazione (per esempio un file word è un *Data object* la cui rappresentazione a video è assicurata da meccanismi informatici di decodifica).

In base al modello OAIS, è evidente che la conservazione non è mai riducibile ai *Data object*, ma deve sempre includere le informazioni descrittive (tra cui *Representation Information, PDI, Information package*). Tali elementi informativi, inoltre, devono essere inseriti in forme appropriate a fini di robusta auto-documentazione all'interno dei pacchetti destinati alla conservazione e non possono essere affidati esclusivamente a un'applicazione esterna che descriva i pacchetti medesimi. Merita sottolineare che il principio dell'indipendenza dei processi conservativi dalle tecnologie e dalle applicazioni trova forza e garanzia proprio nell'accuratezza e nella completezza dei pacchetti informativi di archiviazione (AIP).

⁷ Submission Information Package (SIP): An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.

Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

Dissemination Information Package (DIP): An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS.

La conservazione delle informazioni strutturate di rilevanza giuridica, documentale e archivistica

Per la conservazione di una base di dati, è importante affrontare la questione della sua conservazione a partire dall'analisi del contesto della sua formazione e dall'analisi della tipologia documentaria cui fa riferimento. È inoltre opportuno ricordare, come si preciserà meglio in seguito, che le Linee guida AgID riconoscono tale fattispecie nel momento in cui precisano che i dati conservati in una base di dati possono essere prodotti a partire dalla "registrazione informatica delle informazioni risultanti da transazioni o processi informatici o dalla presentazione telematica di dati attraverso moduli o formulari resi disponibili all'utente" (par. 2.1.1 lett. c) e che "la generazione o il raggruppamento anche in via automatica di un insieme di dati o registrazioni, provenienti da una o più basi dati, anche appartenenti a più soggetti interoperanti, secondo una struttura logica predeterminata e memorizzata in forma statica" (par. 2.1.1 lett. d) determina la produzione di un documento informatico. Ciò indica che per il legislatore italiano una base di dati è uno strumento potenzialmente in grado di raccogliere e aggregare informazioni necessarie a produrre in via automatica documenti informatici.

Una base di dati è, tuttavia, sia una struttura logica di memorizzazione sia uno strumento per la gestione dei dati e delle loro associazioni. È necessario, pertanto, assicurare la capacità di mantenere i dati anche successivamente alle eventuali integrazioni o modifiche, oltre alla capacità di memorizzare le interazioni tra utenti, sistemi e dati (ad esempio su log file) e di preservare da sovrascrittura sia i dati sia le informazioni sulle interazioni. È, inoltre, indispensabile definire regole per l'estrazione, in una forma statica, dei dati memorizzati (che possono essere viste, report e anche potenziali documenti) e tenere presenti le soluzioni tecniche implementate sino a oggi, in grado di garantire l'accesso nel tempo alle informazioni, mantenendone l'integrità e l'autenticità.

La tipologia di informazioni ampiamente utilizzata dai soggetti pubblici e privati, affrontata con particolare attenzione in questo studio, è quella dei dati transazionali, cioè dei dati trattati dalle applicazioni informatiche nelle interazioni con una base di dati. Tali informazioni sono spesso di tipo amministrativo e assoggettate, nel loro trattamento e tenuta, a vincoli di norma e a regolamenti di tipo giuridico. La progettazione e la gestione delle basi di dati non sempre sono accompagnate dalla definizione di specifiche strategie di conservazione, che tengano anche conto della natura mutevole e frammentata di molti sistemi di basi di dati e di metodi per l'eventuale anonimizzazione e/o pseudo-anonimizzazione dei dati ai fini della loro fruizione. In generale, mentre i soggetti privati gestiscono e conservano i dati raccolti di routine per scopi esclusivamente aziendali, finalizzati alla gestione dei loro processi organizzativi e di produzione, i soggetti istituzionali dovrebbero adottare misure atte a garantire la conservazione a lungo termine dei dati in ottemperanza a leggi e regolamenti. Ad oggi, in Italia, pochissimi sono attivi in questo ambito, con l'eccezione delle comunità scientifiche e accademiche che da tempo sviluppano e adottano metodologie finalizzate alla qualità e alla conservazione dei dati per supportare la citazione e la riproducibilità della ricerca nel tempo.

I dati transazionali, creati attraverso le interazioni con una base di dati, possono provenire da una vasta gamma di fonti e rappresentare molti tipi diversi di informazioni. I metodi utilizzati per creare o raccogliere tali dati hanno un impatto diretto, di natura tecnologica e/o organizzativa, sul modo in cui essi possano essere successivamente conservati. I dati transazionali sono preziosi solo se possono essere trasformati in informazione e conoscenza in grado di persistere nel tempo. A tale fine, la loro

gestione e la loro qualità devono includere anche informazioni aggiuntive sui dati stessi, come il contesto, la provenienza e le informazioni sui diritti, le restrizioni sull'accesso ai dati e gli eventuali accordi di licenza. Fortunatamente, la maggior parte di questi dati sono conservati in basi di dati relazionali e gestite mediante un linguaggio di programmazione SQL (Structured Query Language) standard, cosa che ne garantisce elevate forme di interoperabilità.

La tenuta nel tempo delle basi di dati implica riflessioni di natura giuridica, archivistica e tecnologica, per valutare da un lato il mantenimento del valore legale e della rilevanza giuridica dei dati trattati e dall'altro la fruizione nel lungo periodo delle informazioni conservate. Tali aspetti sono di seguito approfonditi, mentre l'aspetto tecnologico, che certamente deve tenere conto di essi, sarà trattato nei capitoli successivi.

1.2. Il punto di vista giuridico nel quadro normativo europeo e nazionale

Per affrontare la natura giuridica di una base di dati e dei suoi contenuti è necessario risalire alle definizioni che sono alla base del nostro sistema giuridico, con particolare riferimento alle indicazioni del regolamento europeo sul documento elettronico e al Codice dell'amministrazione digitale.

Sul piano generale, è possibile affermare, come si dirà meglio in seguito, che il documento informatico contemplato nel sistema giuridico italiano costituisce una fattispecie del concetto di documento elettronico definito nel regolamento europeo eIDAS ed è l'elemento centrale per la trasformazione digitale delle attività amministrative. Grazie a tale approccio, che risale alla fine degli anni Novanta, è stato possibile equiparare i processi documentali cartacei a quelli informatizzati, assicurando la piena corrispondenza tra il documento analogico e quello informatico.

In particolare, ai sensi del regolamento eIDAS, il documento elettronico è formulato in termini generali e aperti: "qualsiasi contenuto conservato in forma elettronica". Purtroppo, le esemplificazioni che completano questa definizione non sono né chiare né esaustive e lasciano margini ampi di incertezza e ambiguità, come sempre capita quando si trasforma una disposizione generale in un elenco, come avviene nel caso della norma europea. Al punto 46 il regolamento, comunque, precisa che a un documento elettronico non sono negati gli effetti giuridici e l'ammissibilità come prova in procedimenti giudiziari per il solo motivo della sua forma elettronica.

Più accurata e precisa è la definizione che il CAD utilizza per individuare il documento informatico come fattispecie del documento elettronico: "rappresentazione informatica di atti, fatti o dati giuridicamente rilevanti". È la funzione di rappresentazione che è all'origine della rilevanza giuridica dei contenuti rappresentati, siano essi atti, fatti o dati, come del resto affermato dal legislatore anche nella definizione di documento amministrativo (informatico) fornita dal DPR 445/2000 con specifico riferimento al settore pubblico: "ogni rappresentazione comunque formata del contenuto di atti, anche interni, delle pubbliche amministrazioni o, comunque, utilizzati ai fini dell'attività amministrativa".

Il combinato disposto delle due definizioni che sono alla base dei nostri sistemi documentali e della normativa tecnica consente di sostenere che anche i dati, se adeguatamente *rappresentati*, possono assumere rilevanza giuridica ed essere, quindi, oggetto di conservazione a supporto del processo decisionale. A questo proposito intervengono anche le Linee guida AgID, entrate in vigore nel gennaio

2022, allorché definiscono le diverse modalità di formazione dei documenti informatici (2.1.1) e includono, tra le altre, la possibilità che essi siano costituiti:

- c) dalla memorizzazione su supporto informatico in formato digitale delle informazioni risultanti da transazioni o processi informatici o dalla presentazione telematica di dati attraverso moduli o formulari resi disponibili all'utente;
- d) dalla generazione o raggruppamento anche in via automatica di un insieme di dati o registrazioni, provenienti da una o più banche dati, anche appartenenti a più soggetti interoperanti, secondo una struttura logica predeterminata e memorizzata in forma statica.

Le due fattispecie qui elencate sono, a loro volta, oggetto di approfondimento per quanto riguarda le modalità da seguire per assicurare la loro immodificabilità e integrità. Si prevede in alternativa:

- l'apposizione di una firma elettronica qualificata, di una firma digitale o di un sigillo elettronico qualificato o firma elettronica avanzata,
- la registrazione nei log di sistema dell'esito dell'operazione di formazione del documento informatico, compresa l'applicazione di misure per la protezione dell'integrità delle basi di dati e per la produzione e conservazione dei log di sistema,
- la produzione di una estrazione statica dei dati e il trasferimento della stessa nel sistema di conservazione.

Anche l'allegato 2 che accompagna le Linee guida AgID offre elementi di rilievo in questo ambito, poiché identifica i formati dedicati al trasporto di dati strutturati, riferendosi con questa accezione anche a formati la cui tipologia di contenuto non sia predeterminata a priori. Le basi di dati sono una delle fattispecie applicative tipiche in questo ambito oltre che ormai largamente diffuse.

Pur rientrando pienamente nell'ambito delle basi di dati ora ricordate, è opportuno sottolineare che il legislatore ha trattato in modo specifico e distinto il caso della base di dati costituita dal registro di protocollo informatico. Pur prevedendone la conservazione obbligatoria, la norma originaria (il D.P.R. 28 dicembre 2000, n. 445) di fatto obbliga tale conservazione congelando i dati prodotti giornalmente in un formato statico non riconducibile alla conservazione del documento informatico inteso come base di dati. Le ragioni sono molteplici, legate soprattutto alla necessità di assicurare la funzione di protezione del registro di protocollo più che la conservazione dei contenuti informativi pregiati che quel registro contiene. Si tratta di un nodo la cui soluzione ottimale richiederebbe una revisione normativa, considerata l'importanza dei contenuti strutturati di cui il registro è costituito. Non mancano, tuttavia, possibili iniziative in grado, anche all'interno del quadro normativo vigente, di proporre soluzioni di compromesso, come quella di affiancare gli obblighi attuali con una procedura di conservazione periodica (annuale) del registro nella forma di base di dati, assicurando in tal modo la tenuta sicura, non modificabile e certa dei dati immessi e il loro pieno trattamento informatico. Secondo questa modalità, si potrebbe generare da tale sistema opportunamente qualificato, sempre annualmente (o comunque, periodicamente), un pacchetto di versamento contenente il documento informatico, firmato dal responsabile del protocollo, con tutte le registrazioni del periodo. La questione è sicuramente complessa tanto da richiedere un approfondimento tecnico cui è dedicato in questo studio il capitolo 5.

1.3. La conservazione permanente delle basi di dati a fini storici e scientifici

A partire dal riconoscimento del valore giuridico delle basi di dati e del loro trattamento a fini probatori, la conservazione delle basi di dati, in quanto parte del patrimonio archivistico storico di un soggetto produttore, richiede una particolare attenzione e requisiti ulteriori rispetto a quelli individuati finora, poiché la dimensione conservativa a fini storici e scientifici implica la possibilità che le risorse digitali in questione debbano essere mantenute nel tempo prolungato e indefinito della conservazione permanente. In questa dimensione diventa essenziale e alquanto problematico, in primo luogo, definire e in secondo luogo tutelare l'integrità e l'autenticità dei contenuti e delle informazioni di relazione, di processo e di provenienza, a fronte di sfide ulteriori legate all'obsolescenza delle tecnologie, alla necessità di esportare i materiali in nuovi ambienti, di disporre di piattaforme in grado di assicurare la fruizione a persone estranee all'ambiente di produzione della base di dati in contesti assai diversi da quelli originari.

Tali esigenze conservative restano comunque ancorate alla natura giuridica del bene di cui si è già detto in precedenza e sono quindi legate ai vincoli di integrità e autenticità non facilmente definibili se il sistema e i contenuti sono dinamici e consentono modifiche e sostituzioni. Anche per questa ragione, nel caso delle basi di dati, la prospettiva archivistica di lungo termine mette in luce con chiarezza (ancor più che per gli altri documenti informatici) la necessità di gestire la conservazione come una funzione *olistica* che non può limitarsi ai soli contenuti (ai dati), ma deve anche considerare e trattare con organicità e precocemente tutte le componenti costitutive del sistema, oltre a comprenderne e a renderne esplicite le informazioni esterne che definiscono e descrivono il contesto di provenienza, gli obiettivi e le funzioni specifiche, i tracciati che guidano la formazione delle tabelle, gli interventi di manutenzione evolutiva da documentare sia all'interno del sistema che attraverso i materiali di corredo.

La dimensione problematica dello studio tra natura dinamica e variabile della base di dati e l'esigenza di stabilità della conservazione di contenuti e risorse rilevanti a fini giuridici trova, quindi, proprio nel caso delle basi di dati una sfida di notevole impegno. La ricerca che ha con maggior rigore affrontato questi aspetti mettendo in evidenza la rilevanza – proprio a fini di conservazione storica di questi patrimoni – delle informazioni di processo è il progetto InterPARES, in particolare lo studio condotto da Luciana Duranti e Ken Thibodeau nel 2006⁸.

I risultati di quell'indagine hanno consentito di individuare le diverse tipologie di sistemi interattivi in grado di produrre documenti d'archivio e i requisiti da rispettare per ciascuna di esse a fini conservativi. La definizione di sistema interattivo utilizzata dagli autori include i casi più complessi e dinamici, nei quali ogni input esterno è in grado di determinare una risposta o un'azione da parte del sistema stesso. Casi, quindi, che costituiscono una sfida rilevante ai principi di integrità e immodificabilità che sono alla base della conservazione anche in ambiente digitale.

La prospettiva seguita dai ricercatori per inquadrare la funzione *documentaria* dei sistemi informatici, riferita alla loro capacità probatoria nel lungo periodo, ha richiesto una serie di riflessioni sulle diverse tipologie di codifica dei dati che possono caratterizzare i documenti che ne derivano: si sono distinti i

⁸ L. Duranti, K. Thibodeau, The concept of records in interactive, experiential and dynamic environment.: the view of InterPARES, IN "Archival Science", 2006, 6, pp. 13-68.

dati di contenuto, i dati relativi alla forma del documento e i dati di composizione che consentono al sistema informatico di produrre informazioni e documenti con forma e contenuti specifici. Si è riconosciuto che contenuto, forma e dati di composizione possono essere contenuti in un'unica componente o appartenere a componenti diverse. Ed è questo il caso delle basi di dati che memorizzano le specifiche delle forme documentarie in un file di report che non include dati di contenuto, a loro volta immagazzinati nelle tavole della base di dati, mentre i dati di composizione, che mappano gli elementi da includere nel report in relazione al modello logico dei dati stessi, sono a loro volta trattati e mantenuti separatamente. Per ottenere e conservare la specifica istanza di una base di dati, sono quindi necessari dati di composizione integrativi che includono tanto i formati utilizzati, quanto gli schemi concettuali all'origine della base di dati. I ricercatori di InterPARES hanno tuttavia considerato anche la necessità di comprendere, soprattutto nei sistemi dinamici e interattivi, tra le componenti da considerare a fini di conservazione, anche le istruzioni che guidano la selezione e la presentazione dei contenuti in una determinata forma.

È soprattutto ai sistemi interattivi che il progetto ha dedicato particolare attenzione, considerata la crescita dirompente di tecnologie dinamiche caratterizzate da regole variabili per la formazione e presentazione dei dati. Le basi di dati di questa natura costituiscono risorse digitali dotate di valore probatorio di lungo periodo i cui contenuti e/o la cui forma non sono statici, poiché sono il risultato di dati che cambiano anche frequentemente come nel caso di sistemi il cui disegno consente l'aggiornamento, la sostituzione delle informazioni e non le storicizzano, anche se oggetto di modifica, a differenza di quanto avviene per esempio nel sistema di registrazione informatica dei documenti.

L'interrogativo in questi casi è radicale: si tratta di sistemi che possono creare documenti informatici in grado di produrre i loro effetti giuridici e mantenere efficacia probatoria nel tempo? La risposta complessa che il progetto InterPARES propone merita di essere presentata in dettaglio anche se le esperienze documentarie oggetto di conservazione nei progetti internazionali ed europei sono finora limitate alle basi di dati statiche, di cui si mantengono i contenuti (dati primari e tabelle) ma quasi mai le interazioni, affidandosi a strumenti di normalizzazione basati su formati standard stabili e riconosciuti.

La risposta che i ricercatori propongono sposta l'attenzione dal singolo documento alla base di dati, che dovrà essere quindi oggetto di conservazione nella sua interezza: la possibilità data all'utente di formare un documento a partire da una selezione variabile di contenuti non nega il principio dell'integrità e dell'immodificabilità, purché si riconosca che è tutta la base di dati a costituire il *documento/archivio*. Ricorrendo all'esempio di un catalogo di prodotti in vendita online e di una selezione di dati ottenuta da un utente, Duranti e Thibodeau chiariscono infatti che:

"in such cases, while what is presented to the user at any moment may appear to be a document, it is instead part of the only existing document, the catalog [database]. The full content of an online sales catalog comprises all data stored and available for presentation to a user".

"In tali casi, mentre ciò che viene presentato in ogni momento all'utente può apparire come un documento, in realtà è parte dell'unico documento esistente, il catalogo [database]. L'intero contenuto di un catalogo di vendite online comprende tutti i dati memorizzati e disponibili per la presentazione ad un utente."

Gli oggetti digitali che rendono possibile la selezione di contenuti specifici sono componenti digitali del documento, la cui integrità e persistenza non sono negate. Gli ambienti digitali interattivi – concludono i ricercatori – forniscono una struttura che permette di selezionare i dati e variarne la sequenza senza inficiare il principio della stabilità dei contenuti e della forma. Il concetto stesso di documento informatico, tuttavia, richiede una nuova definizione che è basata sulla distinzione tra la sua *manifestazione* e la memorizzazione delle sue componenti digitali e ha in InterPARES la seguente formulazione: “le componenti memorizzate rendono possibile la riproduzione del documento, ma non lo costituiscono”. È possibile conservare il documento come *manifestazione*, grazie alla capacità di riproduzione resa possibile dalla tenuta delle sue componenti digitali che non includono quindi solo i dati, ma anche le regole di processamento, comprese quelle che guidano le variazioni consentite dal sistema.

Acquista rilievo crescente in questa nuova realtà la documentazione integrativa, che un recente standard collegato al modello OAIS e sviluppato per documentare in modo permanente i complessi sistemi informatici delle agenzie spaziali ha definito come *additional information*⁹, tanto più significativa quanto più complessa e *disruptive* è la tecnologia disponibile.

Quel che è certo è che gli archivisti, da decenni, hanno riconosciuto la varietà tipologica della produzione documentaria e archivistica che la dimensione digitale rende possibile e hanno sviluppato modelli concettuali per comprenderne la natura e non perderne il controllo e le tracce. In particolare, in Italia hanno anche elaborato una strategia che identifica gli strumenti utili per governare con flessibilità le trasformazioni. Tra questi, meritano una specifica attenzione il piano di classificazione e il piano di conservazione che consentono di dar conto dell’esistenza e stabilire la durata e le politiche di conservazione anche delle basi di dati di valore archivistico permanente. Si tratta di un percorso ancora iniziale che pochissimi enti hanno intrapreso. È stato e sarà tuttavia fondamentale che la normativa tecnica, in particolare le Linee guida AgID entrate in vigore nel 2022 abbiano riconosciuto formalmente la rilevanza di queste attività e reso già oggi possibile accompagnare con indicazioni operative, direttive interne, workflow documentati l’utilizzo di tecnologie di valore strategico, garantendo il trattamento conservativo di una enorme e significativa quantità di informazioni di elevato valore storico e scientifico.

1.4. Forme di rappresentazione e *digital curation* delle basi di dati

La conservazione delle basi di dati richiede, come già detto, la gestione di un insieme di processi e di procedure finalizzate a mantenere l’autenticità, l’integrità, l’affidabilità, la leggibilità, la reperibilità dei dati memorizzati. Si tratta di aspetti che devono essere affrontati nella duplice prospettiva archivistica e informatica in ragione della specifica natura dei documenti e delle risorse digitali da conservare.

Le prime soluzioni elaborate, tradotte da diversi anni in applicazioni concrete, risolvono il problema della conservazione attraverso l’estrazione dalla base di dati delle informazioni da conservare e la

⁹ Si tratta delle raccomandazioni in corso di approvazione a cura del CCSDS, *Information Preparation to Enable Long term use* – IPELTU (CCSDS 653.0-R-1).

conseguente produzione di documenti statici (viste). Successivamente, si è cominciato a porre maggiore attenzione al problema della definizione di schemi concettuali specifici per la produzione di documenti dinamici assimilabili alla documentazione, spostando l'attenzione dalla conservazione della forma fisica dei documenti digitali alla conservazione dei dati logicamente strutturati, ossia ai caratteri intrinseci di questi documenti.

Tali caratteristiche sono fondamentali nel determinare l'interesse per queste risorse nel momento del passaggio dalla fase attiva alla fase di deposito e alla fase di conservazione storica, ricercando soluzioni efficaci sia a fini di consultabilità delle informazioni contenute che per la tutela dell'integrità e dell'autenticità di quelle informazioni. Sin dal momento della creazione di un oggetto digitale dovrebbe essere chiaro se sia necessario o se esista la volontà di garantire l'integrità, l'autenticità e l'accesso alle informazioni e alle meta-informazioni in esso contenute. Soltanto in questo modo, infatti, è possibile progettare le caratteristiche che questi oggetti digitali devono possedere per essere inseriti in un processo di conservazione efficace.

In particolare, i sistemi informatici di conservazione devono possedere, analogamente a quanto avviene per la conservazione documentale, proprietà di tenuta integra e sicura dei dati e devono essere certificati rispetto a ben consolidati standard industriali e normativi che ne possano qualificare tutte le funzionalità nel tempo.

Il conservatore dovrà ricorrere a tecnologie e processi organizzativi in grado di garantire la tenuta di tutti i dati significativi della base di dati garantendone la ricerca e la consultazione dei dati ivi registrati, per tutto il periodo di tenuta ed il loro trattamento, se richiesto.

La qualificazione del sistema è dichiarata dal conservatore che, sotto la propria responsabilità, dovrà asserire che il sistema di conservazione (fatto da un'infrastruttura, da un'organizzazione e da personale qualificato di adeguato profilo professionale, così come previsto dalle norme o da piani di sicurezza e qualità propri dell'organizzazione del conservatore) e descritto nel relativo MO, è rispondente ai requisiti di cui alle norme e Linee guida AgID attuali o a loro future variazioni.

Un adeguato intervento per la conservazione dovrà richiedere una completa e accurata documentazione dei processi specifici nei quali sono coinvolti gli oggetti digitali destinati a una conservazione di medio o lungo periodo. Il contesto di produzione, il contesto di provenienza degli oggetti e dei contenuti digitali, le fasi di gestione delle risorse devono essere mantenuti nel tempo.

Si tratta di attività complesse che, nel caso della conservazione di una base di dati, richiedono innanzitutto una valutazione di merito oltre che impegnativi interventi di natura tecnica. Stabilire il valore di una base di dati, ossia stabilirne natura, struttura e dimensioni, è determinante per definire le modalità di intervento che ne garantiscano una conservazione efficace, soprattutto nel caso in cui si tratti di un sistema informativo di valore archivistico, tale cioè da contenere dati e relazioni giuridicamente rilevanti o che abbiano valore di testimonianza, tanto da richiedere una specifica attenzione per la conservazione della loro autenticità.

In questa prospettiva, è necessaria la progettazione di forme documentarie auto-consistenti, ovvero corredate da informazioni che descrivano le componenti del documento o dell'entità digitale che entreranno a far parte dell'archivio e che saranno trasferiti nel sistema di conservazione. Questo tipo di analisi deve accompagnare un'analisi archivistica di più ampio respiro, che consideri i documenti e le entità digitali nel contesto in cui sono stati formati e gestiti, in relazione tra loro e in relazione con i vari elementi che compongono un sistema documentario complesso.

È necessario ripensare i processi di formazione, gestione e conservazione della documentazione e dell'archivio digitale. I soggetti abilitati a operare in ognuna di queste fasi di vita del documento e dell'archivio digitale dovranno quindi intervenire documentando i processi di gestione degli oggetti digitali, descrivendone le componenti grazie all'aggiunta di informazioni di processo, utili a garantire l'autenticità, l'integrità e l'affidabilità dell'oggetto digitale sia in fase di gestione che in fase di conservazione.

Nell'analisi del processo di conservazione delle basi di dati è necessario garantire quelle caratteristiche di integrità, autenticità e accesso alle informazioni e alle meta-informazioni in esso contenute che possono permettere una conservazione efficace, provvedendo ad effettuare una "Digital Curation"¹⁰ che tenga conto non solo di tipologie documentarie specifiche (registri di protocollo; repertori di documenti omogenei per tipologia come delibere o contratti, etc.), ma che affronti la gestione in un'ottica interdisciplinare (Archivistica, Giuridica, Informatica).

I metadati che descrivono le componenti semantiche dei contenuti di una base di dati, degli oggetti digitali e delle informazioni che dovranno essere trasferiti al sistema di conservazione, devono essere accuratamente definiti. Questa analisi deve essere accompagnata da un'analisi archivistica di più ampio respiro, che consideri questi elementi nel più ampio contesto in cui sono stati formati e gestiti, in relazione tra loro ed in relazione con i vari elementi che compongono un sistema documentario.

È necessario, inoltre, progettare attentamente i processi di formazione, gestione e conservazione delle basi di dati e dei suoi contenuti. I soggetti abilitati ad operare in ognuna di queste fasi del ciclo di vita, infatti, devono intervenire documentando i processi di gestione, descrivendone le componenti intellettuali grazie all'aggiunta di metadati di processo, utili a garantire l'autenticità, l'integrità e l'affidabilità sia in fase di gestione che in fase di conservazione.

Devono essere valorizzati gli aspetti procedurali, ossia l'insieme di regole, istruzioni, azioni e responsabilità che determinano la formazione di una base di dati.

L'utilità di informazioni di corredo è ancora maggiore nel caso degli archivi digitali, sebbene in un ambiente digitale sia necessario governare la formazione di tali dati di processo ricorrendo a modelli e standard. È evidente quindi che sia necessario individuare, all'interno dell'organizzazione che gestisce l'archivio, delle figure adeguatamente competenti e in grado di attuare tutte le operazioni necessarie alla gestione della conservazione digitale.

Uno degli elementi fondamentali del metodo adottato per realizzare i processi di conservazione descritti è certamente l'attività di documentazione. La documentazione delle azioni compiute tanto sui dati, quanto sulla struttura e sulle funzioni della base di dati originaria è importante per garantire l'integrità della base di dati. Tali azioni sono state infatti memorizzate attraverso opportuni metadati PREMIS.

L'attività di documentazione sta alla base anche della capacità di garantire l'affidabilità delle basi di dati conservate, dal momento che è l'unico modo per tenere traccia di tutte le azioni compiute (o non

¹⁰ Con il termine *digital curation* si indicano la selezione, l'acquisizione, la preservazione, la manutenzione, la raccolta e l'archiviazione delle risorse digitali.

compiute) sui contenuti della base di dati originaria e delle scelte operate per determinare dati storicizzati, dati storicizzabili e dati non storicizzati.

Nell'attività di documentazione rientrano anche la creazione di "check list" della documentazione disponibile per spiegare il funzionamento della base di dati e raccontare quale sia stato il suo utilizzo nel tempo, fondamentali per garantire la leggibilità dei dati nel lungo periodo tanto quanto la documentazione generata durante il processo di conservazione digitale.

La reperibilità delle basi di dati esaminate è, invece, garantita dalla soluzione tecnica adottata, che si caratterizza per la scelta del formato SIARD, uno standard non proprietario, aperto e trasparente, che possiede le caratteristiche necessarie a garantire la completa documentazione del processo.

La centralità dell'attività di documentazione di ogni aspetto e di ogni elemento della conservazione digitale delle basi di dati evidenzia come l'attività di conservazione non si esaurisca nella produzione di oggetti o contenuti digitali da archiviare, ma come questa vada intesa come un vero e proprio 'processo', articolato in diverse attività coordinate tra le quali sono comprese anche quelle dedicate a gestire l'accesso ai documenti o ai contenuti o alla base di dati storicizzata. In questo senso la conservazione digitale a lungo termine deve essere un processo aperto, soggetto a continui interventi di verifica di integrità dei dati, di aggiornamento delle procedure, di controllo dell'efficienza dei sistemi.

La strategia di intervento per la conservazione digitale di una base di dati può essere diversificata in relazione a vari elementi sia di tipo tecnico (architettura hardware e software, architettura logica, qualità e consistenza dei dati, capacità della base di dati, etc.) che di tipo gestionale (risorse, finalità), ferma restando in ogni caso l'attenta considerazione di tutti gli elementi di contesto e di provenienza.

Il processo di conservazione digitale di una base di dati va implementato in modo diverso a seconda che si debba intervenire su una base di dati "chiusa", non più utilizzata o su di una base di dati "attiva".

Una base di dati in disuso, perché obsoleta o per avere esaurito la propria funzione di aggregatore di informazioni specifiche, potrebbe essere considerata come quella che in archivistica si definisce una "serie chiusa"; come una serie che non dovrà più essere incrementata di nuovi documenti o un registro che non dovrà più essere utilizzato per la registrazione di atti (registro chiuso), una base di dati chiusa può essere considerata come un complesso di record non più passibile di incrementi. Questa caratteristica, valutata contestualmente ad altri elementi, potrebbe far propendere ad esempio per la scelta di conservare i soli contenuti della base di dati, estraendone soltanto i dati considerati rilevanti (o che in fase progettuale sono stati giudicati meritevoli di una conservazione a lungo termine), e di gestirli indipendentemente dalla base di dati originale.

Diverso il caso delle base di dati che definiamo "attive", che non hanno concluso cioè il loro ciclo di vita, essendo ancora utilizzate e costantemente incrementate. In questo caso si può scegliere tra due strategie diverse, già richiamate all'inizio di questo capitolo: in questo caso è infatti possibile ricorrere sia all'estrazione dei dati, sia alla conservazione dell'intera base di dati.

Esiste tuttavia una terza strategia, ovvero quella di prevedere un processo di conservazione digitale a lungo termine già nella fase di progettazione delle basi di dati e sistemi di gestione documentale, cominciando a implementare il processo di conservazione già a partire dalla fase di formazione e di gestione della base di dati e del sistema documentale (fase "attiva"). Questo significa individuare quali siano gli elementi mancanti o difficilmente recuperabili tra quelli necessari a documentare i processi di

formazione, gestione e conservazione di questi sistemi, a testimoniare l'integrità e l'autenticità, a garantirne la riproducibilità nel tempo.

Nella progettazione delle basi di dati quali elementi integranti di sistemi di gestione documentale, è imprescindibile prevedere tecnologie finalizzate all'integrazione e gestione degli eventi e degli agenti legati ai processi di formazione e gestione dell'archivio. Queste tecnologie possono consistere, ad esempio, in modelli del processo (workflow), meccanismi per la raccolta di dati di esecuzione del processo, *access control list* molto sofisticate in grado di gestire le autorizzazioni in base a utenze profilate su funzioni e processi.

Estremamente utili sono i meccanismi di gestione e monitoraggio delle autenticazioni, attraverso l'uso di tecnologie specifiche come le firme digitali o, ad esempio, attraverso l'implementazione di procedure di certificazione della titolarità dei processi e la produzione di metadati specifici.

Le procedure di gestione dei record devono essere più precise e vincolanti, in modo da consentire non soltanto di gestire la sedimentazione dell'archivio in modo corretto (attraverso protocollazione, classificazione e fascicolazione), ma che rendano manifesto anche il vincolo esistente tra i documenti e agenti (produttori e fruitori) ed eventi e processi, attraverso lo sviluppo di tecnologie in grado di tracciare ogni elemento utile ad una migliore gestione e conservazione dell'archivio.

I processi di produzione e gestione dei record dovranno prevedere meccanismi che consentano la registrazione automatica, o comunque la possibilità di inserimento manuale ma contestuale, di informazioni operative descrittive e dello stato di ogni singola istanza (oggetti, agenti, eventi, etc.), delle *intellectual entity* PREMIS e delle loro relazioni per gestire la corretta interpretazione dei documenti sottoposti a conservazione.

Particolare attenzione deve essere prestata anche alla documentazione dei meccanismi di aggregazione dei dati memorizzati nella base di dati, sviluppati al fine di produrre viste o altre forme manifeste dei record: struttura della base di dati, logica di aggregazione dei campi, mapping tra campi di input e campi di output, meccanismi di generazione delle viste e loro memorizzazione.

Non da ultimo, in ordine di importanza, sarà necessario implementare meccanismi di elaborazione e di memorizzazione delle informazioni di contesto, fondamentali per portare a termine con successo gli obiettivi della conservazione digitale a lungo termine. Ci riferiamo a sistemi di registrazione classificazione e repertorizzazione dei record; vincoli archivistici (espliciti e no) con altri record e record set; dati identificativi e descrittivi di enti produttori e soggetti conservatori, sistemi di gestione della persistenza di informazioni esterne al sistema, come URL, Schemi XSD, ontologie, etc.; utilizzo di standard; sistemi di gestione e memorizzazione dei file di log e di altri dati di transazione, al fine di garantire autenticità e integrità dei dati.

La progettazione delle attività propedeutiche alla realizzazione di una strategia di conservazione digitale avviata precocemente, già in fase di implementazione dell'archivio, sarà possibile grazie all'esperienza ed alle conoscenze maturate nel corso della realizzazione di processi di conservazione come quelli descritti nei casi di studio esposti nel presente documento.

Attuare queste strategie di conservazione digitale delle informazioni significa avere la possibilità di contare su professionalità specifiche, versate e nel settore dell'archivistica e nel settore dell'informatica, altamente specializzate e in grado di lavorare in sinergia.

La conservazione dei dati transazionali

In questo capitolo saranno discussi gli aspetti di natura prevalentemente tecnica riguardanti la conservazione delle basi di dati, assumendo come interesse specifico le basi di dati relazionali create e gestite mediante sistemi transazionali. I sistemi per la gestione delle basi di dati relazionali hanno costituito, infatti, il modello di riferimento de facto fino all'avvento dei paradigmi NoSQL.

La discussione che segue propone una breve introduzione sulle caratteristiche delle basi di dati transazionali, sulle problematiche relative alla loro conservazione a lungo termine e sui diversi modelli utilizzati per la loro memorizzazione relazionale e NoSQL. Successivamente, la trattazione sarà focalizzata esclusivamente sulle basi di dati relazionali che adottano un modello dei dati relazionale (classicamente rappresentato mediante strutture tabellari con associazioni tra le tabelle realizzate mediante chiavi primarie e chiavi esterne che le referenziano) e una manipolazione della base di dati (inserimento, cancellazione e modifica) incapsulata nel costrutto transazionale. Si tralasceranno le successive evoluzioni verso le proposte NoSQL.

Tradizionalmente, un sistema per la gestione delle basi di dati transazionali offre garanzie qualitative sulla manipolazione della base di dati garantite dalle proprietà di atomicità, consistenza, isolamento e persistenza delle transazioni. L'aspetto transazionale ha dunque impatto quasi essenzialmente sull'aspetto dinamico della gestione delle basi di dati. Diverso sarebbe il caso in cui l'oggetto da storicizzare è la transazione stessa, vale a dire la storia delle variazioni delle basi di dati e degli utenti responsabili delle transazioni.

Dato che in questo documento trattiamo la conservazione della base di dati e non delle transazioni, verranno trattate diverse tecniche per la conservazione di dati relazionali.

1.5. I dati transazionali: loro caratteristiche e problematiche di conservazione a lungo termine

La maggior parte dei dati prodotti ogni giorno nel mondo sono di tipo transazionale. I dati transazionali sono tutti quei dati generati tramite una transazione, ovvero da un'unità logica di elaborazione che corrisponde ad una serie di operazioni fisiche elementari (letture o scritture) su di una base di dati. Ogni transazione garantisce il rispetto di quattro proprietà indicate con l'acronimo ACID:

- Atomicità: la transazione deve essere eseguita come se fosse un'operazione atomica;
- Consistenza: la transazione deve lasciare la base dati in uno stato consistente;
- Isolamento: ogni transazione è eseguita indipendentemente dalle altre, quindi, deve essere gestita la concorrenza tra varie transazioni;
- Durabilità: gli effetti di una transazione che sia conclusa correttamente devono essere rese persistenti nel tempo (anche in presenza di guasti).

Ogni interazione con una base di dati è una transazione. I dati transazionali sono pervasivi nella nostra società e sono alla base di tutti i servizi rivolti a cittadini, imprese e istituzioni.

Tutti i dati transazionali raccolti sono associati a dei metadati che li descrivono, come, ad esempio, il loro schema, il formato, la provenienza e tutti gli aspetti legati al loro trattamento (ad esempio, i dati

sensibili devono essere trattati seguendo norme come la GDPR). I metadati associati ai dati (ossia i dati transazionali originali) contribuiscono a creare un'informazione significativa da cui è possibile generare conoscenza. Per questo motivo nella conservazione a lungo termine è necessario non solo conservare i dati, ma anche tutti i metadati ad essi associati al fine di conservare il valore informativo della base di dati.

Un'altra problematica legata alla conservazione riguarda i sistemi di gestione delle basi di dati e l'infrastruttura hardware e software di supporto. Essendo questi sistemi in continua evoluzione tecnologica, è necessario garantire nel tempo la fruibilità dei dati indipendentemente dalle infrastrutture utilizzate per la loro elaborazione.

Nelle basi di dati relazionali i dati sono memorizzati in tabelle: ogni tabella è identificata da un nome, ha un insieme di colonne, ognuna con un tipo di dato fissato, e un insieme di righe che rappresentano l'istanza dei dati. Le basi di dati relazionali rispettano le proprietà ACID e sono interrogate tramite un linguaggio standard chiamato SQL (Structured Query Language) condiviso fra tutti i sistemi di gestione di basi di dati relazionali.

Le basi di dati relazionali risultano essere ancora i sistemi più utilizzati per la memorizzazione di dati transazionali, soprattutto nell'ambito degli Enti pubblici in cui non sono richieste le caratteristiche dei sistemi NoSQL ed è importante il mantenimento delle proprietà ACID, per questo motivo di seguito tratteremo la conservazione delle basi di dati di tipo relazionale.

1.6. Metodi per la conservazione delle basi di dati a lungo termine

I principali metodi utilizzabili per la conservazione delle basi di dati a lungo termine possono essere classificati in: staticizzazione con generazione di documenti informatici, emulazione, migrazione, normalizzazione. Per ciascuno di essi è riportata di seguito una breve descrizione.

1.6.1. Staticizzazione

Con tale metodo si procede alla staticizzazione della base di dati, nel rispetto di regole preventivamente definite per la loro conservazione, con la produzione a tempi definiti di viste e report come documenti informatici che andranno generati e conservati secondo la normativa vigente.

Per un efficace processo di staticizzazione sono essenziali i seguenti passi: definire, sin dalla progettazione del sistema informatico, i documenti informatici da generare e da conservare e i loro formati; pianificare i punti di sincronizzazione periodica in cui operare, con l'estrazione dei dati e la loro aggregazione in documenti informatici; definire i meccanismi per la generazione anche dei metadati necessari alla conservazione dei documenti generati.

1.6.2. Emulazione

Con tale metodo si usa un emulatore del sistema originale, in particolare, per una base di dati relazionale si intende preservare la base di dati e il DBMS originali.

L'emulazione comporta la sostituzione di componenti software e/o hardware dello stack tecnologico della base di dati con software che simula il funzionamento di queste parti; ad esempio, l'uso di una macchina virtuale per imitare l'hardware mantenendo intatto il resto dello stack tecnologico.

Questa strategia mantiene l'ambiente originale della base di dati e si connota come vantaggio perché permette l'emulazione di tutte le funzionalità del sistema informativo originale. Tuttavia, esso può costituire nel tempo uno svantaggio in quanto le differenze tecnologiche e le evoluzioni nell'interazione uomo macchina e macchina-macchina possono rendere difficile per il consumatore utilizzare il sistema nel suo stato originale. I problemi nella tecnologia originale possono anche essere accidentalmente preservati (ad esempio rischi di sicurezza noti) e le restrizioni di accesso potrebbero ostacolare l'accesso ai consumatori in futuro.

L'emulazione può anche introdurre problemi legati alle licenze software e ai diritti di proprietà intellettuale, poiché il sistema operativo, il DBMS e altre applicazioni coesistenti, possono avere licenze associate che negano il diritto di duplicare, accedere o utilizzare la tecnologia.

L'emulazione può essere molto complessa da implementare in contesti in cui le basi di dati sono distribuite su una rete. In questi casi, l'intera configurazione di rete deve essere preservata dall'ambiente di emulazione.

1.6.3. Migrazione

La migrazione comporta il trasferimento da un DBMS a uno alternativo più aggiornato rispetto alle necessità funzionali e tecnologiche. La migrazione coinvolge potenzialmente sia la ristrutturazione dei dati e metadati della base di dati, sia la potenziale re-implementazione della *business logic* storicizzata (parte delle regole e vincoli di integrità storicizzate) nella base di dati di partenza (ad esempio, la parte di *trigger* e procedure memorizzate utilizzati molto spesso per garantire la correttezza e la qualità dei dati).

Il vantaggio principale della migrazione è la possibilità di rendere fruibile la base di dati in un sistema aggiornato rispetto alle esigenze tecnologiche e di fruizione dei dati.

Lo svantaggio principale è il costo e la potenziale difficoltà tecnica quando l'ambiente di destinazione non sia un'estensione conservativa dell'ambiente di partenza. In particolare, quando il modello dei dati di partenza e di destinazione non siano compatibili e i linguaggi di programmazione che codificano la *business logic* storicizzata non siano supportati nel sistema di gestione di destinazione, costringendo in tal caso ad una re-ingegnerizzazione.

Ulteriore elemento negativo della migrazione è che il processo di migrazione deve essere calibrato per uno specifico ambiente di partenza e uno specifico ambiente di destinazione.

1.6.4. Normalizzazione

La normalizzazione può essere vista come una forma di migrazione verso un formato standard per la base di dati. Come nel caso della migrazione essa coinvolge potenzialmente sia la struttura dei dati sia i metadati della base di dati sia la *business logic* storicizzata.

Un vantaggio della normalizzazione rispetto alla migrazione è che l'ambiente di normalizzazione ha un modello dei dati usualmente inclusivo rispetto alle variazioni di un modello dei dati diffuso quale quello relazionale. È inoltre disponibile un linguaggio standard di interrogazione della base di dati normalizzata (SQL) che non dipende dalle varianti.

Un secondo vantaggio è che, essendo fissato il possibile ambiente di destinazione, è possibile definire dei pattern di migrazione dagli ambienti sorgenti più diffusi all'ambiente normalizzato (molti-a-uno), consentendo ampio riuso delle procedure di normalizzazione in contesti in cui la sorgente usi la medesima tecnologia.

La normalizzazione soffre, invece, delle stesse difficoltà della migrazione per quanto riguarda la *business logic* storicizzata in forma procedurale, poiché gli ambienti normalizzati correntemente disponibili non offrono soluzioni per la *business logic* procedurale.

1.6.5. Migrazione/normalizzazione di basi di dati relazionali

La conservazione delle basi di dati relazionali con il metodo della normalizzazione:

- permette di circoscrivere il problema della conservazione considerando inizialmente dati che hanno le caratteristiche di forte strutturazione e omogeneità, lasciando allo studio futuro i casi più complessi di dati semi-strutturati ed eterogenei;
- l'aderenza significativa delle basi di dati relazionali a uno standard linguistico accettato permette di disegnare approcci adeguati a un ampio spettro di sistemi per la gestione di basi di dati;
- inoltre, nonostante la recente diffusione di paradigmi e metodologie che superano o integrano il modello relazione standard essa costituisce ancora il modello di riferimento nella grande maggioranza delle applicazioni tradizionali.

Di seguito si riportano diverse considerazioni riguardo agli approcci di normalizzazione e migrazione:

- l'approccio di migrazione/normalizzazione funziona esportando alcune proprietà della base di dati originale dal suo DBMS in un altro DBMS o formato di file più adeguato alla conservazione a lungo termine;
- il DBMS di destinazione deve essere scelto con cura per garantire il successo di questa strategia;
- il DBMS o il formato di file scelto dovrebbe essere maturo, aperto, ampiamente adottato, ben supportato dalla comunità e trasparente. Esso dovrebbe soddisfare i requisiti di conservazione e i futuri utilizzi stabiliti durante la pianificazione della conservazione.

I vincoli da considerare sono i seguenti:

- la compatibilità con le versioni precedenti di un DBMS comporta l'utilizzo di versioni software e/o hardware più recenti per aprire, accedere e leggere una base di dati creata utilizzando una versione precedente;
- l'interoperabilità comporta la riduzione della possibilità di obsolescenza garantendo che un determinato file sia accessibile con più di una combinazione di software e hardware;
- la conversione agli standard comporta il trasferimento dell'archiviazione dei dati da un formato proprietario a un formato aperto, più facilmente accessibile e ampiamente utilizzato.

La migrazione, così come l'emulazione, richiedono anche di effettuare la raccolta di documentazione sull'ambiente tecnologico di una base di dati. Questa documentazione può includere: manuali del DBMS, informazioni sull'applicazione dell'utente finale, specifiche del formato di file, dettagli sul sistema operativo e descrizioni dell'hardware. La documentazione può anche includere informazioni su altre applicazioni che coesistono nello stesso ambiente IT. Una delle difficoltà inerenti a questo approccio sta nel conoscere l'intera portata della documentazione che sarà necessaria per capire il contenuto in futuro.

Alla luce dell'analisi riportata, l'approccio di normalizzazione rappresenta l'attuale migliore pratica quando si tratti di preservare una base di dati.

1.6.6. Analisi del contesto internazionale per la conservazione delle basi di dati

Essendo le basi di dati relazionali uno dei sistemi più utilizzati per la memorizzazione dei dati, il tema della loro conservazione è stato ampiamente affrontato in ambito internazionale e diverse nazioni hanno sviluppato tecniche e strumenti atti alla loro conservazione.

La maggior parte delle tecniche utilizzate si basa sull'esportazione dei dati in formato testuale utilizzando standard Open Source (es. CSV, XML, SQL) che consentono di creare documenti di testo in grado di contenere dati e metadati estratti dal sistema originale che possono essere facilmente conservati. L'obiettivo finale è quello di conservare dati e metadati in modo trasparente e indipendente dal sistema originale, così da poterli reintegrare in futuro su di un qualsiasi DBMS.

Riportiamo di seguito alcuni progetti e strumenti che sono stati sviluppati per la conservazione delle basi di dati relazionali.

Archival Data Description Markup Language (ADDML)

ADDML¹¹ è uno standard basato su linguaggio XML sviluppato dall'Archivio Nazionale Norvegese. Il formato memorizza in un file XML i metadati della base di dati originale, la sua struttura e i dati.

¹¹ <http://xml.ra.se/addml/>

Questo formato è stato adottato anche dall'Archivio Nazionale Svedese che ha sviluppato un software chiamato KRAM che consente di convertire basi di dati relazionali in formato ADDML validati (es. verifica che i dati siano conformi ai metadati).

Tutta la documentazione e i software sono disponibili solo in lingua norvegese, risultando difficilmente utilizzabili.

Standard Data Format for Preservation (SDFP)

SDFP¹² è un formato di archiviazione sviluppato come parte del progetto "Migration to Intermediate XML for Electronic Data (MIXED)" dal DANS (Data Archiving and Networked Services), l'archivio di dati scientifici olandese. Il formato è basato su XML e consente l'archiviazione di diversi tipi di file, tra cui le basi di dati relazionali.

Gli autori del progetto hanno sviluppato anche un tool chiamato MIXED¹³ distribuito come servizio web che consente di convertire file colonnari (Excel, CSV) in formato SDFP, al momento della stesura di questo documento il servizio non risulta attivo.

Software Independent Archiving of Relational Databases (SIARD)

SIARD¹⁴ è un formato Open Source sviluppato dall'Archivio Federale Svizzero specificatamente per l'archiviazione di basi di dati relazionali. Il formato SIARD utilizza un file XML compresso con algoritmo ZIP per memorizzare tutti i dati e i metadati di una base dati relazionale.

L'Archivio Federale Svizzero mette a disposizione un software denominato SIARD Suite dal quale è possibile esportare/importare una base di dati nel/dal formato SIARD. Il software è ben documentato in lingua inglese ed è distribuito in varie lingue, per questo è stato adottato per una sperimentazione illustrata nella sezione 0.

Software commerciali

Esistono svariati software commerciali atti alla conservazione delle basi di dati. Lo scopo principale di questi software è rimuovere i dati inattivi da una base di dati in modo da alleggerirne il funzionamento. Un esempio è **CHRONOS**¹⁵, un software che archivia i dati di una base di dati per consentirne un'usabilità a lungo termine. Il software crea archivi ZIP che contengono i dati all'interno di file CSV, consentendo la loro lettura senza la necessità del DBMS originale. I metadati vengono salvati in file XML che contengono i nomi degli attributi, i tipi di dato e i vincoli relazionali.

¹² <http://www.ijdc.net/article/view/195/260>

¹³ <https://sites.google.com/a/datanetworkservice.nl/mixed/>

¹⁴ <https://www.bar.admin.ch/bar/it/home/archiviazione/strumenti/siard-suite.html>

¹⁵ <https://www.csp-sw.com/chronos-en>

1.7. Razionale per la valutazione della complessità del DB origine per selezionare una soluzione per la conservazione nel tempo e vincoli di cui tenere conto

Un aspetto critico nella gestione di un prodotto software, come nel caso di una base di dati, è la presenza di una documentazione adeguata che accompagni ogni fase del suo ciclo di vita. Solo una documentazione adeguata può permettere l'eventuale recupero di parte delle funzionalità del sistema archiviato qualora questo si renda necessario. Un prodotto software in esercizio può contare sul passaggio di consegne e sul trasferimento di esperienza degli operatori che si avvicinano nell'amministrazione del prodotto. Un prodotto archiviato deve poter contare solo sulla documentazione esistente per l'interpretazione concettuale delle strutture memorizzate e per le finalità di utilizzo intese da progettisti e amministratori.

La documentazione che dovrebbe accompagnare la base di dati da storicizzare riguarda:

- La struttura della base di dati in esercizio;
- La business logic codificata;
- I casi d'uso;
- Le Informazioni statistiche e gli aspetti tecnologici della base di dati in esercizio.

Di seguito vengono analizzati i punti menzionati.

1.7.1. La struttura della documentazione della base di dati in esercizio

In questa sezione viene tratteggiata in forma essenziale la struttura della documentazione che dovrebbe essere associata alla base di dati in esercizio. Ad essa andrebbe successivamente allegata la documentazione di supporto alla storicizzazione della base di dati e alle eventuali attività di trasformazione della base di dati propedeutiche alla storicizzazione. La documentazione dovrebbe tener traccia sia della struttura della base di dati stessa sia degli applicativi o delle forme d'uso della base di dati. Per quanto riguarda la base di dati sono consigliati usualmente i seguenti documenti:

- descrizione del livello concettuale della base di dati;
- descrizione del livello logico/fisico della base di dati.

La descrizione del livello concettuale è focalizzata sulle entità informative memorizzate, sulla loro struttura e sulle loro associazioni. È usualmente sintetizzata mediante un tradizionale schema Entità-Associazioni o alternativamente mediante un *Class Diagram* di UML. Tali schemi possono essere supportati da documenti semi-strutturati esplicativi (Dizionari delle entità e delle associazioni) e business logic (*esplicitazione di eventuali vincoli semantici derivati dal dominio di interesse utili a fissare regole di consistenza per i dati, inclusi i vincoli d'integrità dichiarativi esprimibili in linguaggio SQL*).

Nel caso la documentazione del livello concettuale non fosse già disponibile o non fosse allineata alla base di dati esistente, essa dovrebbe essere ricostruita mediante una operazione di "Reverse engineering", a partire dallo schema della base di dati. *In particolare, il dizionario dei vincoli statici può essere ricostruito parzialmente a partire dall'implementazione dei vincoli statici della base di dati.*

La descrizione del livello logico/fisico è focalizzata sulla codifica relazionale delle entità e delle associazioni e sulla loro specifica implementazione nel DBMS di esercizio. Qualora la documentazione del livello logico/fisico non fosse già disponibile essa potrebbe essere desunta dalla definizione SQL delle tabelle (CREATE TABLE).

La descrizione del livello logico disponibile in SQL comprende:

- l'elenco delle tabelle;
- per ogni tabella, l'elenco degli attributi e tipi di dati associati;
- gli identificativi delle tabelle desumibili (se presenti) dalle dichiarazioni di chiave primarie (PRIMARY KEY) o da vincoli di unicità (UNIQUE);
- la dipendenza tra tabelle realizzata dalle coppie chiave esterna-chiave primaria desumibili dalle dichiarazioni di chiave esterna (FOREIGN KEY).

Nella documentazione della base di dati dovrebbero essere individuate con chiarezza le tabelle contenenti i dati significativi, distinguendole dalle tabelle di supporto (ad esempio le viste materializzate) o dalle tabelle obsolete (utilizzate per scopi tecnici nella vita della base di dati) eventualmente presenti. Esempi di tabelle obsolete sono gli *snapshot* di tabelle che vengono creati in alcune occasioni per necessità di backup.

1.7.2. Documentazione della business logic codificata nella base di dati in esercizio

In questa sezione viene fatto cenno alla struttura della documentazione che descrive la business logic codificata nella base di dati in esercizio. Essa comprende tra l'altro: l'implementazione di vincoli di consistenza che la base di dati in esercizio deve soddisfare (non rappresentabili come vincoli dichiarativi in SQL); il computo di dati derivati a partire dai dati memorizzati.

L'assunto principale del problema della conservazione della base di dati è che il suo contenuto abbia carattere statico e non più modificabile. Essendo la base di dati non modificabile e partendo dal presupposto che l'immagine della base di dati da storicizzare soddisfi i vincoli di consistenza, la documentazione dei vincoli di consistenza in essa presenti ha il fine di dare una descrizione semantica più puntuale delle strutture dati ospitate. L'implementazione dei vincoli può assumere forma diversa in ragione delle diverse tipologie di vincoli:

- di dominio;
- di unicità;
- di chiave primaria e di chiave esterna;
- di riga e tabella;
- correlazioni tra tabelle .

I vincoli possono essere definiti mediante linguaggio SQL in modalità dichiarativa (per lo più nel contesto della dichiarazione della tabella) o in via procedurale (Stored procedure e Trigger). I vincoli definiti mediante SQL possono essere agevolmente specificati nella dichiarazione dello schema della base di dati (CREATE o ALTER TABLE, CREATE DOMAIN etc., con particolare riguardo alle clausole CHECK e CONSTRAINT delle definizioni). Si ritengono particolarmente significativi a scopo di documentazione i

vincoli di chiave primaria ed esterna e i vincoli di dominio che restringono gli insiemi di valori ammissibili associati agli attributi.

I vincoli che mettono in corrispondenza tabelle diverse esulano dalla definizione delle tabelle e, pertanto, sono i più complessi da riconoscere e documentare. Essi sono frequentemente implementati in termini procedurali mediante TRIGGER e STORED PROCEDURE (se implementati internamente al sistema per la gestione della base di dati e non esternamente nella business logic degli applicativi Client). In questo caso, la documentazione dovrebbe associare la descrizione testuale dei vincoli ai trigger o alle procedure che li implementano.

È importante evidenziare che nel caso plausibile in cui la scelta di storicizzazione della base di dati sia indirizzata ad una soluzione di standardizzazione di struttura e contenuti (soluzione SIARD ad esempio) gran parte delle definizioni di vincoli non troverà realizzazioni nelle strutture ospitanti. *In modo particolare, di trigger e procedure - basati su corpi procedurali che dipendono in modo costitutivo da linguaggi di programmazione spesso proprietari e inevitabilmente destinati a una forte dinamica evolutiva - si potrà mantenere traccia solo se opportunamente descritti nella documentazione.*

Un aspetto importante riguarda la documentazione dei dati calcolati a partire dai dati memorizzati e gli algoritmi di calcolo. Poiché le procedure memorizzate non saranno preservate nella storicizzazione della base di dati si pone il problema di come preservare la possibilità di derivare l'informazione qualora essa possa essere successivamente richiesta (ad esempio mediante l'estensione degli schemi con la memorizzazione opportuna dei dati calcolati).

1.7.3. Documentazione dei casi d'uso della base di dati in esercizio.

In questa sezione viene fatto cenno alla struttura della documentazione che descrive gli scenari d'uso della base di dati in esercizio. La modalità di interazione degli applicativi e degli utenti con la base di dati è importante perché permette di comprendere come i dati sono fruiti dagli attori esterni e perché indica le possibili richieste a cui la base di dati storicizzata potrebbe dover rispondere anche in futuro.

Passo preliminare all'analisi dei casi d'uso è l'individuazione degli attori possibilmente coinvolti nell'uso della base di dati:

- categorie di utenti e diritti/limitazioni di accesso a tabelle;
- applicativi che si interfacciano con la base di dati.

Per i diversi casi di interazione è significativo recuperare le interrogazioni che utenti/applicativi inviano alla base di dati (poiché la base di dati avrà una immagine statica sono meno rilevanti le operazioni di manipolazioni di dati quali INSERT, UPDATE, DELETE). La documentazione sulle interrogazioni in esercizio è di estrema rilevanza per comprendere le modalità d'uso della base di dati. Esse costituiscono un banco di prova per poter verificare che la base di dati storicizzata sia ancora in grado di rispondere alle richieste informative che la base di dati in esercizio era in grado di soddisfare.

Le interrogazioni possono essere recuperate in diverse sorgenti informative:

- interrogazioni memorizzate;
- interrogazioni associate a viste e viste materializzate;

- interrogazioni comprese nel corpo procedurale di procedure memorizzate e di trigger;
- se disponibili, interrogazioni inoltrate da applicativi client mediante API al sistema per la gestione della base di dati.

1.7.4. Informazioni statistiche sulla base di dati in esercizio.

Si ravvisa l'opportunità di documentare l'immagine stabile della base di dati da storicizzare con l'informazione statistica dei dati. L'informazione è in genere facilmente recuperabile consultando il dizionario della base di dati che ogni DBMS mette a disposizione per la consultazione. L'informazione è utile per avere delle metriche quantitative sulla base di dati e per condurre, eventualmente, misure di conformità tra la base di dati in esercizio e la base di dati storicizzata (ad esempio, la coincidenza di numero di righe delle tabelle corrispondenti, del numero di valori distinti associati ad un attributo etc.).

Un esempio (non esaustivo) delle informazioni statistiche comprende:

- per ogni tabella il numero delle registrazioni;
- per ogni attributo di una tabella:
 - il numero dei valori distinti associati all'attributo;
 - il numero dei valori NULL associati all'attributo;
 - il valore minimo e massimo associati all'attributo.

1.7.5. Aspetti tecnologici della base di dati di esercizio

Rilevante è l'informazione sulle caratteristiche tecniche del DBMS che ha in carico la base di dati in esercizio. In particolare:

- tipo e versione del DBMS;
- linguaggi di scripting per le procedure memorizzate e/o trigger.

Come accennato nella descrizione generale del modello migrazione/normalizzazione è verosimile che serva definire un progetto di mappatura ad hoc per ogni DBMS verso l'ambiente di normalizzazione (ad esempio SIARD).

L'informazione sul DBMS fornisce implicitamente il grado di aderenza dei linguaggi di definizione e manipolazione della base di dati rispetto allo standard SQL. Ogni DBMS dichiara nella sua documentazione le difformità rispetto allo standard (l'adesione non è mai di fatto completa). Inoltre, frequentemente essi introducono:

- varianti sintattiche di costrutti presenti nello standard SQL;
- tipi di dati, operatori e funzioni non presenti nello standard SQL (caso molto frequente è quello legato ai tipi di dati di base).

È di particolare rilievo documentare le istruzioni del linguaggio di definizione e manipolazione dei dati del DBMS sorgente che non abbiano corrispondenza diretta o indiretta nel linguaggio di definizione e manipolazione normalizzato.

1.7.6. Modalità di utilizzo futuro della base di dati storicizzata

Una chiara previsione sulle modalità di utilizzo futuro della base di dati storicizzata è un requisito fondamentale per una migrazione/normalizzazione efficace e fruibile.

La previsione riguarderà: gli attori coinvolti (utenti, applicativi, etc.); i casi di uso che coinvolgono recupero di dati dalla base di dati storicizzata.

Un modo ragionevole per circostanziare le previsioni di utilizzo futuro è indicare l'insieme delle interrogazioni a cui la base di dati dovrà rispondere.

Esse possono comprendere:

- un insieme rappresentativo delle interrogazioni già servito dalla base di dati in esercizio e già elencate in fase di documentazione nella Sezione 4.3.3 (si osservi che le query della base di dati in esercizio dovranno essere opportunamente traslate per operare nella struttura normalizzata della base di dati storicizzata);
- un eventuale insieme di interrogazioni corrispondenti a finalità di recupero estese rispetto a quelle della base di dati in esercizio (ad esempio recupero di valori aggregati per supporto decisionale).

La disponibilità di un insieme di interrogazione di riferimento costituisce un efficace banco di test per verificare la correttezza e completezza dell'operazione di migrazione/normalizzazione.

L'esecuzione delle query sulla base di dati in esercizio e delle query traslate sulla base di dati normalizzate dovrà restituire lo stesso risultato.

1.7.7. Valutazione della qualità dello schema e dei dati delle basi di dati da storicizzare

Un passo preliminare alla migrazione/normalizzazione è la valutazione della qualità dei dati della base di dati in esercizio. Il primo ovvio criterio è il soddisfacimento dei vincoli di consistenza usualmente associati ai dati per garantirne l'integrità strutturale:

- vincoli di totalità ed unicità dei campi associati ad una chiave primaria;
- vincoli di integrità referenziale per le chiavi esterne.

Il soddisfacimento dei vincoli non è scontato se la base di dati ha operato in esercizio con i vincoli disabilitati, ma esistono consolidate tecniche di *data profiling* per abilitarli.

Un passo aggiuntivo può riguardare aspetti qualitativi interni alle singole tabelle. A titolo di esempio e senza pretesa di esaustività:

- duplicazioni di righe;
- rispetto dei vincoli di dominio associati agli attributi;
- standardizzazione dei valori associati agli attributi;
- rispetto della struttura per attributi che ospitano tipi di dati strutturati codificati in stringhe di caratteri (ad esempio, indirizzi, codici, date, marche temporali).

Operazioni di miglioramento della qualità dei dati sono ritenuti indispensabili quando la base di dati storicizzata debba rispondere a interrogazioni che recuperano dati in forma aggregata utilizzati possibilmente per supporto decisionale (esistono linee guida articolate in tal senso per le fasi di trasformazione dei dati nei processi ETL preliminari al popolamento di un Data Warehouse a partire da sorgenti operazionali dei dati).

Un passo di valutazione progettuale aggiuntivo può riguardare gli aspetti di complessità strutturale della base di dati da storicizzare. Una degli approcci peculiari dei sistemi transazionale che adottano il modello relazionale dei dati è la frammentazione (normalizzazione) di entità concettualmente unitarie in tabelle legate da corrispondenze tra chiavi primarie e referenze mediante chiavi esterne. La finalità è l'eliminazione delle ridondanze e dei problemi connessi alla gestione delle eventuali duplicazioni nella manipolazione dei dati (inserimenti, cancellazioni, variazioni).

Una misura di complessità della base di dati in esame potrebbe essere legata al numero di *join* (giunzioni) necessarie a ricomporre la frammentazione di entità logiche (ad esempio, il numero di join per le interrogazioni significative già documentate nella Sezione 4.3.3 o le interrogazioni previste nella Sezione 4.3.6).

Poiché nella fase di migrazione/normalizzazione si può assumere che la base di dati sia e rimanga statica, è possibile riflettere sulla possibilità di ricomporre la frammentazione (denormalizzazione) per ricomporre in un'unica tabella tutte le informazioni collegate ad una entità concettuale, evitando sequenze complesse di join nelle interrogazioni della base di dati storicizzata.

La conservazione di basi dati secondo lo standard SIARD

In questo capitolo analizziamo SIARD, un sistema di conservazione di basi di dati relazionali sviluppato dall'Archivio Federale Svizzero, basato sul modello di normalizzazione presentato nel capitolo precedente.

SIARD propone uno standard di normalizzazione per conservare i dati di una base di dati relazionale, intendendo conservati dati che siano ancora fruibili dopo un periodo di 50-100 anni.

SIARD è un formato unificato per eseguire il dump di una base di dati (è un file contenente il contenuto e la struttura di una base di dati). È possibile usarlo per finalità di backup (copia di riserva) in un formato standard e consente di avere una singola applicazione per eseguire normalizzazione/denormalizzazione di una base di dati. In questo modo solo questa applicazione deve essere mantenuta aggiornata per consentire il collegamento a nuovi DBMS quando variano le specifiche.

SIARD è concepito secondo l'idea che una base di dati è preservata se una query di selezione eseguita sulla base di dati originale e su quella ripristinata danno lo stesso risultato; si considerano solo query di interrogazione (SELECT) perché l'obiettivo è solo quello di conservare i dati.

Specifiche/Informazioni:

- memorizza una base di dati in un file singolo;
- l'utente che esegue l'archiviazione dovrebbe avere i massimi privilegi, SIARD non può leggere ciò che l'utente non vede;
- è basato su standard open;
- il backup deve contenere tutti i dati;
- il backup deve contenere tutti i metadati primari (nomi delle tabelle, nomi colonne, tipo di dato);
- il backup dovrebbe (opzionale) contenere FK, AK, PK, sono considerate opzionali perché la loro presenza non cambia il risultato di una SELECT;
- i tipi di dati ripristinati non è detto che siano identici, ma devono essere comparabili, es. CHAR/VARCHAR;
- oggetti Binary (BLOB) hanno problemi ad essere archiviati, un oggetto BLOB si presume possa essere un insieme di bit letto da un software; quindi, non è un oggetto destinato a una lunga conservazione;
- tutti i tipi di dati SQL possono essere mappati in tipi di dato XML (XML viene usato per memorizzare i metadati, quindi le informazioni sui nomi di colonna e tipo);
- le viste non vengono riportate perché il linguaggio SQL può differire da DBMS a DBMS, si memorizza (se possibile) la query che genera la vista, i campi e il tipo di dato;
- stored procedure e trigger non vengono riportati, in quanto la loro definizione è dipendente dal DBMS in cui sono implementati, non viene riportato nemmeno il codice per la loro generazione.

1.8. Funzionalità specifiche del software

Il software SiardGui è un tool che permette di importare ed esportare, tramite un'interfaccia grafica, basi di dati relazionali nel formato SIARD.

Lo strumento, inoltre, offre l'opportunità di leggere file in formato SIARD per poterne esplorare il contenuto. Il software supporta vari DBMS da cui esportare/importare basi di dati, in particolare: Microsoft Access, DB/2, H2 Database, PostgreSQL, MySQL, Oracle e SQLServer; la sperimentazione si è focalizzata in particolare sugli ultimi quattro.

1.9. Testing di importazione da SQL Server

Utilizzando diverse basi di dati di testing sono stati eseguiti dei test di esportazione/importazione del formato SIARD, al fine di valutare il mantenimento della struttura, dei vincoli relazionali e dei dati presenti all'interno della base di dati originarie.

In particolare, dato che il principio della conservazione tramite normalizzazione non garantisce la conservazione di tutte le componenti della base di dati, la sperimentazione si è focalizzata su conservazione di:

- vincoli di chiave: Primary Key e Foreign Key;
- vincoli sui valori: CHECK, NOT NULL;
- tipi di dato "numerici" e di stringhe (in particolare mantenimento delle differenze, ad esempio, fra INT e SMALLINT, o fra VARCHAR e NVARCHAR);
- viste;
- Binary Large Objects (BLOB).

Il procedimento seguito è il seguente:

1. utilizzando SiardGui ci si collega alla base di dati tramite una connessione JDBC;
2. si crea un backup della base di dati normalizzata secondo il formato SIARD;
3. utilizzando il file in formato SIARD si ripristina una copia della base di dati;
4. si verificano eventuali differenze tra la base di dati originale e quella ripristinata;

Nell'allegato "Sperimentazione SIARD" sono riportati tutti i dettagli del processo di normalizzazione seguito per talune basi di dati d'esempio utilizzati nelle prove effettuate.

La conservazione dei registri e dei repertori

La tenuta di registri e repertori digitali rappresenta di certo un concreto caso, seppur semplice, di conservazione di una base di dati. Un tale approccio, qualora perseguito, certamente favorirebbe un'elevata efficienza nei processi di trattamento dei registri e nella conservazione dei loro dati, organizzati in tabelle di una base di dati con record strutturati per memorizzare i vari campi che interessano le singole linee di un registro. In tal modo, come già precedentemente evidenziato, un registro digitalizzato non si limiterebbe a riprodurre con il ricorso a un documento informatico la tradizionale organizzazione documentale cartacea, così come oggi, in base alle vigenti norme, è ampiamente praticato.

In un sistema informatico, un registro digitale è implementabile mediante una o più tabelle di una base di dati in cui memorizzare le informazioni proprie del registro e a cui possono essere aggiunte eventuali informazioni di servizio utili a renderne possibile il trattamento informatico e la consultazione. Esso può essere modellato logicamente come una struttura dati regolare, fatta di record (righe/linee di scritture nel registro) che memorizzano eventi totalmente (o parzialmente) ordinabili, fissati nel tempo mediante forme di marcatura temporale (ad es. un timestamp). L'epoca di registrazione di ciascun evento è successiva al tempo di creazione/attivazione del registro (l'istante zero di attivazione del registro è il primo timestamp apposto all'atto della creazione del registro).

Il timestamp associato a ciascuna linea da memorizzare, unitamente all'identificatore unico del registro ad esso collegato inscindibilmente, rappresenta l'analogo digitale della numerazione delle pagine e della loro vidimazione manuale, da parte di una terza parte fidata che inizializza e valida il registro cartaceo.

Al fine di certificare l'inserimento della linea dati nel registro del soggetto a cui esso è collegato, si ricorre all'uso di un'evidenza da parte del titolare che ne lascia la traccia certa.

Un sistema di log traccia tutti gli eventi significativi legati ai vari versamenti in conservazione delle linee di un registro. Anche i file di log dovranno presentare le caratteristiche di integrità e inalterabilità, ed essere protetti con idonee misure contro ogni uso improprio e dovranno essere conservati a norma.

Particolare attenzione, stante lo specifico quadro normativo, va posta alla conservazione dei registri di protocollo, come sarà discusso di seguito, stante i vincoli normativi a cui devono essere assoggettati.

Di seguito, a titolo di esempio, sono analizzati tre casi di conservazione di registri: i repertori notarili, i registri di protocollo e il registro dei verbali d'esame universitari. Altre tipologie di registri e repertori sono simili e facilmente modellabili a partire da questi esempi.

1.10. Conservazione dei repertori informatici notarili

Di seguito si illustra sinteticamente un possibile modello di digitalizzazione dei repertori notarili che vede principalmente coinvolti come stakeholder oltre al notaio titolare del repertorio, l'organizzazione notarile (Consiglio Nazionale del Notariato e Distretti Notarili) e il Ministero della Giustizia-Archivi Notarili. Il modello qui esposto riporta i punti salienti fin qui emersi nell'ambito di uno studio-progetto

più ampio, in corso di sviluppo presso il notariato, di digitalizzazione e conservazione dei registri in generale, applicato prioritariamente alla digitalizzazione dei repertori.

In base alla legge notarile i notai devono tenere i repertori degli atti tra vivi, di ultima volontà e dei protesti.

In essi un notaio deve annotare, entro il giorno successivo e per ordine di numero, tutti gli atti ricevuti (compresi gli atti rilasciati e quelli esenti da registrazione). Tale attività è, allo stato, effettuata dai notai utilizzando un tradizionale registro cartaceo numerato e vidimato in tutte le pagine.

La legge notarile, all'art.66-bis comma 1 e 2 prevede già la possibilità di ricorrere alla digitalizzazione e tenuta dei repertori digitali:

“Tutti i repertori e i registri dei quali è obbligatoria la tenuta per il notaio sono formati e conservati su supporto informatico, nel rispetto dei principi di cui al decreto legislativo 7 marzo 2005, n. 82.

Il notaio provvede alla tenuta dei repertori e dei registri di cui al comma 1 avvalendosi della struttura di cui all'articolo 62-bis.”

Per l'attuazione di tale norma è richiesta la definizione di regole tecniche che indichino le caratteristiche della soluzione da implementare. Ovviamente, tale digitalizzazione riguarda gli aspetti principali dell'informatizzazione e della conservazione del registro repertorio.

La soluzione più ovvia per tale informatizzazione è quella di replicare, con ricorso a un documento informatico, la struttura e il processo di creazione e tenuta del repertorio documentale, e cioè di generare un analogo digitale del repertorio e delle sue pagine e procedere alla tradizionale conservazione a norma di un documento informatico costituito da una o più pagine del repertorio stesso, sottoscritte con firma digitale qualificata del notaio.

Purtroppo, tale soluzione non consentendo efficientemente di disporre della struttura dei dati che compongono ciascuna linea del repertorio (o addirittura non rendendone possibile del tutto l'accesso) fa sì che non possono essere effettuati molti dei trattamenti informatici richiesti dai vari stakeholder, facendo perdere, pertanto, tutti i benefici attesi con la digitalizzazione. La soluzione proposta di conservazione delle basi di dati fa sì che, a parità di garanzie di tenuta e sicurezza delle informazioni in essere con la conservazione del documento informatico, si possa operare con una piena informatizzazione transazionale di tali registri.

Un repertorio notarile è composto da una serie di linee di registro ognuna associata ad un singolo atto notarile. Ogni linea è univocamente individuata da chiavi identificative che la associano al repertorio di uno specifico notaio operante in un distretto notarile (al cambiamento di distretto ad un notaio è assegnato un nuovo registro di repertorio, per cui nella sua vita professionale ad un notaio possono essere associati più registri).

Il processo di gestione del repertorio notarile con ricorso ad una base di dati può essere così sinteticamente descritto:

- il notaio, in base alla legge notarile, predispone righe di indice di repertorio da inviare in conservazione sul suo registro e contenenti tutti gli estremi dell'atto. La riga ha come chiave il numero di repertorio progressivo, così come avviene attualmente con il processo cartaceo;

- il sistema di conservazione inserisce un timestamp (Il timestamp assicura l'ordinamento totale di tutti gli eventi riga repertorio) e aggiunge la riga alla tabella repertorio associata al notaio;
- a distanza anche di giorni il notaio completa l'evento repertorio inserendo le altre indicazioni richieste e invia la linea in conservazione, con la stessa chiave/numero repertorio usata nell'indice, annotando la linea di indice precedentemente registrata. Di fatto, si realizza in tal modo un registro dinamico utilizzando la tecnica delle annotazioni di riga che, ad una riga radice, aggiunge nel tempo ulteriori righe ad essa strettamente collegate;
- il mese successivo, come da vincoli di legge, entro il 26 di ogni mese, il notaio procede a congelare e convalidare le immissioni delle righe di repertorio trascritte, firmando, ad esempio, con la propria firma digitale la parte di repertorio del mese corrente;
- a partire dalla base dati repertorio conservata, apposite funzioni applicative provvedono a predisporre l'estratto repertoriale mensile e ad effettuare controlli e conteggi (tali funzionalità sono disponibili sia per i notai sia per gli operatori degli Archivi Notarili sia per altri soggetti accreditati per effettuare controlli. È data, inoltre, la possibilità di verificare in finestre temporali preindicate, le vari registrazioni relative a ciascun notaio di un distretto e a fare ricerche e visualizzazioni dei dati conservati).

Tale processo deve essere supportato da una specifica piattaforma qualificata, accessibile dai vari soggetti interessati che implementa tutte le funzionalità applicative e di base (componenti di servizio, identificativi, autorizzativi, documentali, etc.) necessari.

Le funzionalità principali per la gestione del repertorio sono le seguenti:

- inizializzazione registro/repertorio;
- movimentazione registro/repertorio;
- consolidamento in conservazione;
- consultazione ed estrazione di linee in un periodo prefissato;
- reportistica con generazione anche di pagine di repertorio come documenti informatici da inviare periodicamente (annualmente) nel sistema di conservazione documentale.

Gli attori coinvolti nei processi di gestione del repertorio sono:

- lo studio notarile: notaio e collaboratori autorizzati a movimentare il registro repertorio del notaio;
- la società Notartel del notariato (limitatamente alle funzioni di amministrazione di sistema;
- Il Consiglio Nazionale del Notariato e i Distretti notarili;
- gli Archivi Notarili (limitatamente alle funzioni ispettive);
- le SW-House notarili (limitatamente all'integrazione con i software di studio per l'alimentazione dei dati).

Tali attori, per poter operare e accedere alle specifiche funzionalità di ciascun registro devono essere assoggettati a regime identificativo e autorizzativo a livello di sicurezza medio-alto mediante un sistema IAM che soprassiede all'utilizzo del servizio.

1.11. Conservazione dei registri di protocollo

L'innovazione tecnologica ha trasformato il protocollo da strumento statico a dinamico, in grado di svolgere più funzioni e di essere integrato con altri sistemi elettronici, interni o esterni all'ente che lo utilizza. Esso, quindi, mantiene la sua originaria funzione legale, ma può anche diventare parte di un sistema informativo più complesso, che gestisce documenti e procedimenti e interagisce con altre applicazioni.

La rilevanza del registro di protocollo discende dalla sua storia e dalla sua natura giuridica di atto pubblico di fede privilegiata¹⁶. Ha svolto nella sua forma cartacea tradizionale e continua a svolgere nella sua evoluzione informatica un ruolo fondamentale di garanzia di imparzialità e trasparenza dell'azione amministrativa, attraverso la funzione delicata di registrazione dei documenti per attestare data e provenienza certa dei documenti registrati e indicare il momento di ingresso degli stessi in un archivio pubblico¹⁷.

Il tradizionale supporto cartaceo del registro di protocollo è stato sostituito, a partire dagli anni Ottanta del XX secolo, da supporti magnetici (nastri e dischetti), e si è cominciato a redigere il registro con l'ausilio di elaboratori elettronici. In questa fase, quindi, il registro si è trasformato in una banca dati, e, talvolta, la sostituzione dei registri di protocollo cartacei ha prospettato il rischio di una progressiva perdita di valore probatorio dello stesso registro. Nel passaggio da cartaceo a digitale in un certo senso il registro di protocollo si è 'perso' nel mare della gestione documentale, andando a far parte di un sistema complessivo di gestione in cui ai documenti sono associati elementi di profilazione (metadati). In altre parole, si può ritenere che se in un archivio cartaceo il registro di protocollo viene prima prodotto e poi conservato, all'interno dei sistemi di protocollo informatico, invece, il registro si presenta nella forma di una banca dati costituita da un insieme di singole registrazioni cronologiche di protocollo, identificate da un numero progressivo univoco prodotte durante l'anno solare.

La registrazione di protocollo può quindi essere intesa come un insieme di dati di cui bisogna garantire l'integrità nel tempo per mantenerne il valore probatorio di fede privilegiata.

Si è posto quindi il problema di come e quando consolidare il registro nella sua forma permanente e di quali informazioni comprendere in esso o come tracciare le eventuali modifiche successive definendo requisiti funzionali minimi di sicurezza.

¹⁶ Tale qualità è stata più volte ribadita dalla giurisprudenza. Valga per tutte la sentenza del Consiglio di Stato 1993, I, 838.

¹⁷ Sulla funzione e storia del registro di protocollo, dalla sua nascita, risalente alla fine del XVIII secolo, la sua diffusione nel nostro paese derivata soprattutto dalle invasioni napoleoniche, con i principi importati direttamente dalla Rivoluzione Francese, fino all'introduzione del protocollo informatico, si rimanda ai riferimenti bibliografici e al testo del saggio di Gianni Penzo Doria, *Profili archivistici del protocollo informatico*, in "Archivi & Computer", n. 1/2005 disponibile anche in rete nel sito di Procedamus: <https://www.procedamus.it/images/materiali/GPDProfiliarchivisticiidelprotocollo2005.pdf>.

In questa sede si ricorda solo il "regolamento per gli uffici di registratura e di archivio nelle diverse Amministrazioni centrali" (Regio Decreto 25 gennaio 1900, n. 35) che ha rappresentato in ambito statale un importante elemento organizzativo e funzionale per tutto il XX secolo, non sempre riconosciuto dagli stessi addetti al servizio di protocollo. Il R.D. 25 gennaio 1900 n.35 è stato formalmente abrogato con il D.lgs. 13 dicembre 2010 n. 212 ma, di fatto, superato dall'entrata in vigore del DPR 428/98 prima e poi del DPR 445/2000.

Nell'ordinamento italiano, specifiche disposizioni di legge hanno normato le caratteristiche della registrazione e le funzionalità dei sistemi di registrazione (si veda in particolare il Decreto del Presidente della Repubblica del 28 dicembre 2000, n. 445, artt. 53-55 e 57), dettando regole volte a disciplinare la produzione di uno strumento che, come detto, si qualifica come atto pubblico di fede privilegiata. La norma italiana definisce anche che il sistema di gestione informatica dei documenti “deve consentire la produzione del registro giornaliero di protocollo, costituito dall'elenco delle informazioni inserite con l'operazione di registrazione di protocollo nell'arco dello stesso giorno”¹⁸. Pertanto, il registro giornaliero deve contenere, in modo ordinato e progressivo, l'elenco delle informazioni conseguenti alle azioni di registrazione di protocollo effettuate nell'arco di uno stesso giorno oltre agli eventuali annullamenti effettuati nella giornata, riferiti a registrazioni dei giorni precedenti.

In continuità con le regole tecniche precedenti, le “Linee guida sulla formazione, gestione e conservazione dei documenti informatici” al paragrafo 3.1.6 “Requisiti minimi di sicurezza dei sistemi di protocollo informatico” stabiliscono che il registro giornaliero di protocollo è trasmesso entro la giornata lavorativa successiva al sistema di conservazione, garantendone l'immodificabilità del contenuto.

Le stesse Linee guida AgID al paragrafo 3.1.3 e 3.3.3 sviluppano da un lato il concetto che “la registrazione informatica dei documenti è rappresentata dall'insieme di dati in forma elettronica allegati o connessi al documento informatico al fine dell'identificazione univoca di tutti i documenti prodotti e acquisiti” e dall'altro che nei sistemi informatici i registri sono formati “attraverso la generazione o il raggruppamento anche in via automatica di un insieme di dati o registrazioni, provenienti da una o più banche dati”, ma nello stesso tempo definisce che i registri sono formati “secondo una struttura logica predeterminata e memorizzata in forma statica”, quindi nella forma di un documento informatico secondo quanto previsto dalla lettera d) del paragrafo 2.1.1 delle Linee guida AgID.

I dati delle registrazioni che confluiscono in tale documento informatico provengono dalle basi di dati del sistema documentale dell'ente, pertanto, la soluzione oggi in uso è da annoverare nella forma di conservazione tramite staticizzazione come precedentemente illustrata.

Allo stato attuale le soluzioni per la produzione di un registro di protocollo giornaliero da conservare privilegiano gli elementi in grado di garantire meglio l'immodificabilità piuttosto che il mantenimento delle funzioni essenziali.

¹⁸DPR 445/2000 art. 53. Lo stesso articolo del DPR 445/2000 stabilisce che la registrazione di protocollo per ogni documento ricevuto o spedito richiede la memorizzazione delle seguenti informazioni:

- a) il numero di protocollo del documento generato automaticamente dal sistema;
- b) la data di registrazione di protocollo assegnata automaticamente dal sistema e registrata in forma non modificabile;
- c) il mittente per i documenti ricevuti o, in alternativa, il destinatario o i destinatari per i documenti spediti;
- d) l'oggetto del documento;
- e) la data e il protocollo del documento ricevuto, se disponibili;
- f) l'impronta del documento informatico, se trasmesso per via telematica.

La pratica quotidiana ha portato, nella maggior parte, a produrre il registro giornaliero di protocollo nei formati PDF e PDF/A, ma sarebbe più semplice e in certi casi è già stato fatto, passare ad altre tipologie di formati quali quelli ipertestuali e per dati strutturati. Tali formati *machine readable* offrirebbero chiari vantaggi in termini di efficienza e compattezza delle informazioni memorizzate.

A livello attuale di sviluppo dei sistemi informatici, oltre a quanto previsto dalla normativa vigente, si deve però valutare come mantenere attive le funzioni della registrazione di protocollo, in particolare, relative alla ricerca e reperibilità dei documenti registrati, contemperando gli aspetti di staticità, autenticità, integrità e sicurezza.

Un processo di conservazione dovrebbe prevedere, in prospettiva, non solo di conservare e di produrre copie accurate e autentiche del registro di protocollo, inteso come specifico documento archivistico, ma anche di riprodurre le principali funzionalità del sistema originario, utilizzando il registro di protocollo conservato come elemento di prova della autenticità dei documenti digitali conservati.

Nel caso della conservazione dei registri e repertori realizzata come basi di dati, si conservano non pagine di un registro, intese come un documento informatico, ma le singole linee di scrittura che lo compongono. In tal senso, l'oggetto digitale in conservazione diventa la singola linea/record che sarà memorizzata nella base di dati.

Per soddisfare vincoli di norme o degli specifici processi organizzativi di un ente, si può anche procedere a generare periodicamente documenti informatici, che aggregano l'insieme delle linee giornaliere del registro, da inviare in conservazione.

Tale soluzione, oggi praticabile, potrebbe evolvere a seguito di ulteriori definizioni normative, verso la realizzazione di un registro di protocollo come base di dati qualificata da conservare annualmente.

Infine, è da ricordare la rilevanza dell'analisi sulla produzione e conservazione del registro di protocollo, anche in una visione più ampia, alla luce della possibilità per la Pubblica Amministrazione di estendere alle altre forme di registrazione informatica i requisiti fissati per la registrazione di protocollo secondo quanto previsto nel citato paragrafo 3.1.3 delle Linee guida AgID¹⁹.

¹⁹ L'ultimo capoverso di detto paragrafo recita: "La Pubblica Amministrazione, al fine di dare attuazione alle disposizioni introdotte dal CAD stesso in materia di sistema di gestione informatica dei documenti realizza le funzionalità di gestione dell'archivio corrente, dell'archivio di deposito, dei flussi documentali, automatizzazione dei procedimenti amministrativi sulla base dei propri obiettivi di miglioramento dei servizi e di incremento dell'efficienza operativa, tenuto conto del rapporto costi e benefici, nel rispetto degli articoli 53 e 55 del TUDA e dei requisiti del sistema di gestione informatica dei documenti e dei flussi documentali definiti negli articoli 52, 65 e 67 del TUDA, applicando ove possibile i requisiti fissati per la registrazione di protocollo anche alle altre forme di registrazione informatica dei documenti, fatto salvo quanto disposto per esse da eventuali norme vigenti".

1.12. Conservazione dei verbali di esame

Un altro esempio di conservazione di basi di dati è dato dalla tenuta dei registri d'esame universitari (noti come verbali/camicie di un insegnamento). Un esame universitario relativo ad uno specifico insegnamento interessa, come attori principali, la commissione d'esame presieduta dal titolare del corso (o suo sostituto) e da uno o più membri (il numero è definito dagli specifici regolamenti universitari) e dallo studente che intende sostenere l'esame. Oltre a tali soggetti sono interessati anche vari uffici amministrativi che partecipano al processo di verbalizzazione svolgendo prevalentemente funzioni gestionali e di controllo.

La seduta d'esame è pubblica e può essere descritta dalle seguenti fasi:

- 1 prenotazione dell'esame da parte di uno studente;
- 2 apertura sessione d'esame con autenticazione a due fattori almeno (ad es. SPID) del docente che presiede la sessione;
- 3 operazione di appello da lista dei prenotati ed eliminazione dei candidati assenti;
- 4 avvio seduta esami (sessione autenticata e profilata);
- 5 per ogni studente esaminato deve sempre registrarsi una linea (record) con indicazione dell'esito (positivo/ rinuncia/ bocciatura) e del voto assegnato. La linea sarà marcata con un timestamp di sistema e il check di chiusura del presidente della commissione e che, unitamente all'accesso con identificazione a due fattori (SPID-firma avanzata), consolida la registrazione dell'esame. Una volta chiusa la registrazione, la linea è memorizzata definitivamente nel sistema di conservazione qualificato del registro verbale e non più modificabile;
- 6 tutti gli studenti che sostengono l'esame in una sessione sono verbalizzati ciascuno singolarmente, una chiave di sessione apposta in ciascuna linea d'esame, ne permette l'aggregazione logica al verbale di sessione;
- 7 A fine sessione d'esame il presidente, con la chiusura del verbale della sessione (evento tracciato nel log qualificato di sistema), attiva la trasmissione della notifica (mediante un sistema di messaggistica sicuro interno o esterno all'ente) del report del verbale d'esame, con la descrizione di tutti gli esami effettuati e dei tempi di apertura e chiusura di sessione, a tutti membri della commissione. Parallelamente viene notificata ad ogni studente la relativa linea di verbale;
- 8 La segreteria studenti, mediante un'interfaccia applicativa di propria pertinenza, può generare in qualunque momento un verbale di sessione ed estrarre singole linee d'esame a partire dall'identificatore di sessione d'esame e dalla matricola di uno studente. Eventuali contestazioni da parte di uno studente o di un componente la commissione in merito alla verbalizzazione devono essere notificate agli appositi uffici dell'università e sono gestite da questi al di fuori del processo di verbalizzazione. Eventuali modifiche o correzioni che ne dovessero derivare vanno comunicate al presidente della commissione (o a un soggetto con i poteri di modifica) che provvederà ad inserire una nuova linea, con la stessa chiave di sessione, correttiva della precedente contestata. Tutte le linee resteranno, comunque, memorizzate nella base dati.

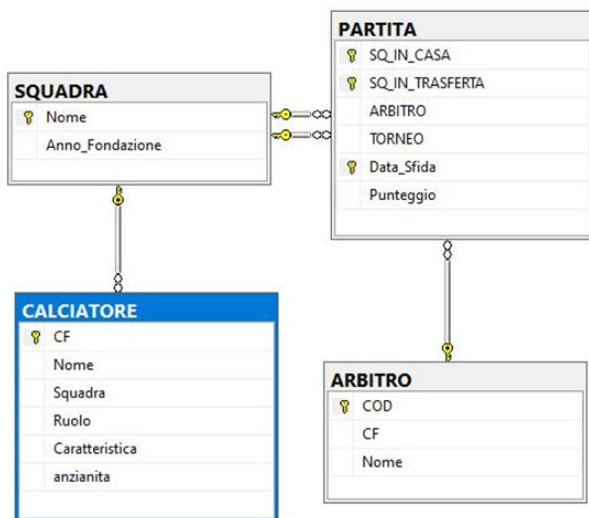
Come per gli esempi precedenti la base dati del registro verbale d'esame deve essere ospitata su di un sistema tecnologico certificato e qualificato secondo i criteri organizzativi e di sicurezza propri di un sistema di conservazione documentale.

Allegato - Sperimentazione di SIARD

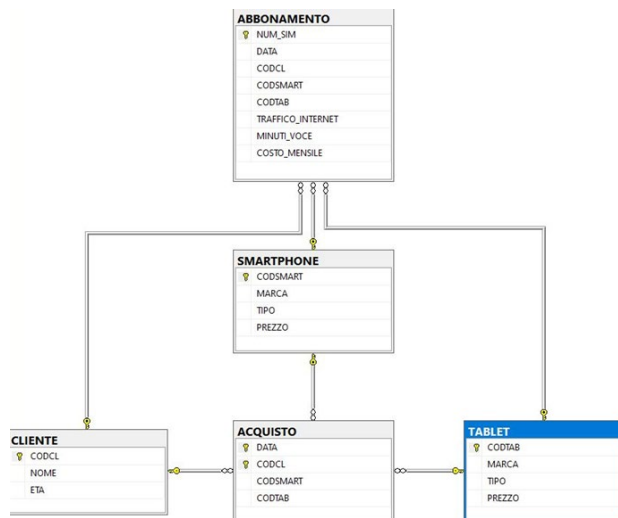
1.13. Basi di dati utilizzate

Di seguito viene riportato lo schema delle basi di dati utilizzate nella sperimentazione. Tutte queste basi di dati risiedevano su un'istanza di SQL Server Express 2019

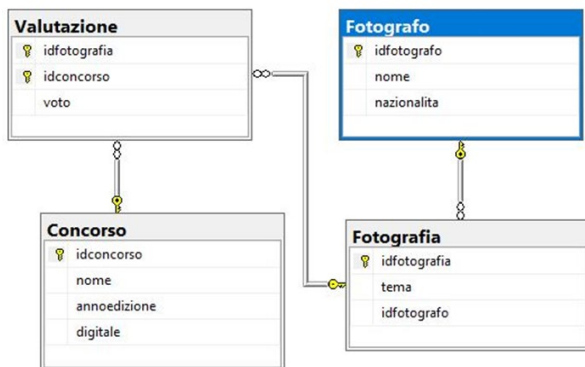
DB CALCIO



DB SMARTPHONE



DB CONCORSO FOTOGRAFICO

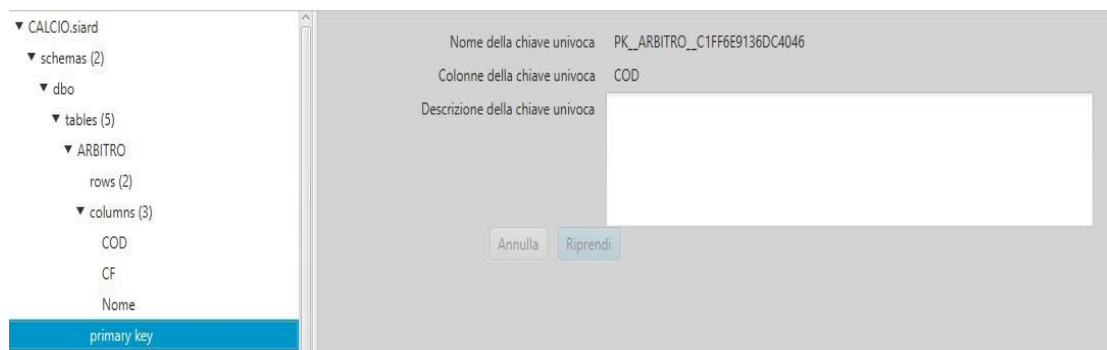


Oltre a queste basi di dati, è stato utilizzato anche una base di dati d'esempio fornito da Microsoft: AdventureWorks2019.

1.14. Test di normalizzazione

Come prima cosa, si è provato a esportare queste basi di dati tramite il tool SiardGui, per valutare quali informazioni vengono mantenute e quali vengono perse o modificate.

- Mantenate tutte le chiavi primarie:



- **Mantenute tutte le chiavi esterne:**

```

▼ CALCIATORE
  ▼ columns (4)
    CF
    Nome
    Squadra
    Ruolo
    primary key
  ▼ foreign keys (1)
    FK__CALCIATOR__Squad__29572725
▼ PARTITA
  rows (2)
  ▶ columns (6)
  primary key
  ▼ foreign keys (3)
    FK__PARTITA__ARBITRO__2E1BDC42
    FK__PARTITA__SQ_IN_C__2C3393D0
    FK__PARTITA__SQ_IN_T__2D27B809
  
```

- **Mantenuti i vincoli di NOT NULL, mentre si perdono i vincoli CHECK:**

Tipo di dato della colonna originale	varchar
E' ammesso un valore non definito per la colonna	false
Valore standard per la colonna	

- **Mantenuti i tipi di dato numerici e di stringa:**

Nome della colonna	Nome	Nome della colonna	CF
Posizione della colonna	3	Posizione della colonna	2
Elenco esterno per la colonna LOB		Elenco esterno per la colonna LOB	
Tipo MIME degli oggetti nella colonna		Tipo MIME degli oggetti nella colonna	
Tipo di dato SQL Datentyp della colonna	<u>VARCHAR(30)</u>	Tipo di dato SQL Datentyp della colonna	<u>VARCHAR(20)</u>

- **Mantenute le viste: colonne di cui sono composte, query che le genera e breve descrizione**

Nome della View	VistaProva
Richiesta originale	SELECT CALCIATORE.CF as cfCalc from CALCIATORE
Richiesta standard SQL	
Numero di record	0
Descrizione della View	

Posizione	Nome della colonna	Tipo della colonna	Cardinalità
1	cfCalc	VARCHAR(20)	

Per verificare invece il comportamento riguardo ai BLOBs, SQL Server è orientato all'uso delle File Tables, quindi, è stata creata una base di dati con una File Table che contenesse qualche file. Con questa soluzione i BLOBs sono file mantenuti all'interno del File system e la base di dati ne conserva solo i metadati trattandoli come oggetti "condivisi" con il sistema operativo.

- **Come prevedibile, essendo questa una funzione specifica di SQL Server, le File Tables sono salvate come tabelle normali, sono mantenuti quindi i metadati relativi ai file e non i file stessi.**

Nome della tabella	DocumentStore
Numero di record	1
Descrizione della tabella	
Annulla	Riprendi

Posizione	Nome della colonna	Tipo della colonna	Cardinalità
1	stream_id	CHAR(36)	
2	file_stream	VARBINARY(2147483647)	
3	name	NCHAR VARYING(255)	
4	path_locator	VARCHAR(892)	
5	parent_path_locator	VARCHAR(892)	
6	file_type	NCHAR VARYING(255)	
7	cached_file_size	BIGINT	
8	creation_time	VARCHAR(64)	
9	last_write_time	VARCHAR(64)	
10	last_access_time	VARCHAR(64)	

1.15. Importazione da file SIARD

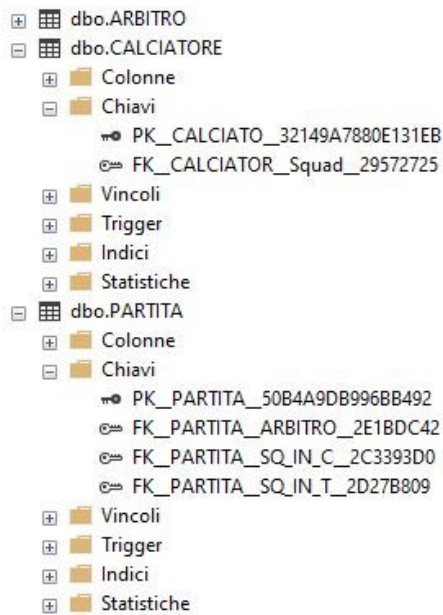
Il Client SiardGui permette di ricreare una base di dati a partire dai file SIARD sui DBMS supportati. I seguenti test sono stati fatti per evidenziare le differenze che la base di dati generata dal file SIARD rispetto alla base di dati originale. Per fare questi test sono stati usati basi di dati provenienti da SQL Server e sono stati esportati su DBMS diversi per tentare di evidenziare gli eventuali problemi di compatibilità che possono sorgere da queste conversioni, prestando attenzione agli elementi elencati prima.

Questi problemi di compatibilità potrebbero anche verificarsi quando si esportano dati da una versione ad un'altra dello stesso DBMS.

Ricordiamo che tutte le basi di dati di origine risiedevano su SQL Server.

1.15.1. Destinazione: SQLServer

- **La base di dati ricreata mantiene le chiavi primarie ed esterne**

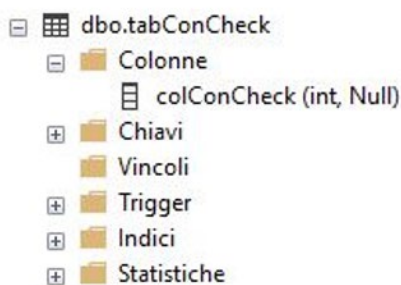


- **Vengono mantenuti i vincoli NOT NULL, ma non i CHECK**

DB originale

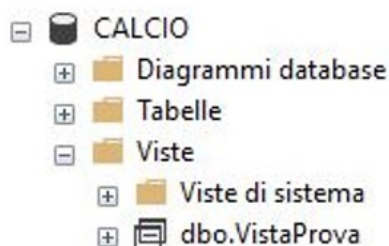


DB rigenerato, cartella Vincoli vuota:

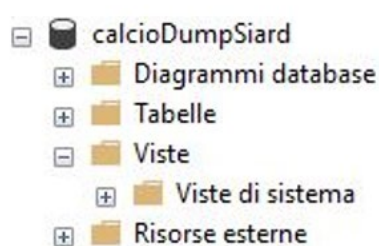


- **Vengono conservati anche i tipi di dato numerici e di stringa**
- **Le viste, invece, non sono presenti nel DB esportato:**

DB originale:



DB dopo l'esportazione, cartella Viste vuota:



- Per quanto riguarda le FileTables, nel DB esportato queste sono trasformate in semplici tabelle che mantengono i metadati dei file che erano presenti al momento in cui è stato effettuato il dump, ma alcune colonne non mantengono il loro tipo originale:

DB-originale:

Colonne	
stream_id	(uniqueidentifier, Non Null)
file_stream	(varbinary(max), Null)
name	(nvarchar(255), Non Null)
path_locator	(PK, hierarchyid, Non Null)
parent_path_locator	(FK, Calcolata, hierarchyid, Null)
file_type	(Calcolata, nvarchar(255), Null)
cached_file_size	(Calcolata, bigint, Null)
creation_time	(datetimeoffset(7), Non Null)
last_write_time	(datetimeoffset(7), Non Null)
last_access_time	(datetimeoffset(7), Null)
is_directory	(bit, Non Null)
is_offline	(bit, Non Null)
is_hidden	(bit, Non Null)
is_readonly	(bit, Non Null)
is_archive	(bit, Non Null)
is_system	(bit, Non Null)
is_temporary	(bit, Non Null)

DB-dopo-l'esportazione: colonne che hanno un tipo non "standard" non mantengono il tipo originale:

Colonne	
stream_id	(char(36), Non Null)
file_stream	(varbinary(max), Null)
name	(nvarchar(255), Non Null)
path_locator	(PK, varchar(892), Non Null)
parent_path_locator	(FK, varchar(892), Null)
file_type	(nvarchar(255), Null)
cached_file_size	(bigint, Null)
creation_time	(varchar(64), Non Null)
last_write_time	(varchar(64), Non Null)
last_access_time	(varchar(64), Null)
is_directory	(bit, Non Null)
is_offline	(bit, Non Null)
is_hidden	(bit, Non Null)
is_readonly	(bit, Non Null)
is_archive	(bit, Non Null)
is_system	(bit, Non Null)
is_temporary	(bit, Non Null)

Database AdventureWorks2019:

Durante l'importazione la base di dati di prova AdventureWorks2019 genera un errore su di un campo che nella base di dati originale può assumere valori NULL, questo risulta anche nel file SIARD, però quando la tabella viene ricreata il campo viene settato come NOT NULL creando problemi durante il ripristino delle istanze. Abbiamo visto che è possibile intervenire manualmente per correggere il problema sulla base di dati per poi continuare con l'importazione dei dati.

dbo.DatabaseLog	
Colonne	
DatabaseLogID	(PK, int, Non Null)
PostTime	(datetime, Non Null)
DatabaseUser	(sysname(nvarchar(128)), Non Null)
Event	(sysname(nvarchar(128)), Non Null)
Schema	(sysname(nvarchar(128)), Null)
Object	(sysname(nvarchar(128)), Null)
TSQL	(nvarchar(max), Non Null)
XmlEvent	(XML(.), Non Null)

Nome della colonna	Schema
Posizione della colonna	5
Elenco esterno per la colonna LOB	
Tipo MIME degli oggetti nella colonna	
Tipo di dato SQL Datentyp della colonna	
Schema del tipo UDT di colonna	sys
Nome del tipo UDT di colonna	sysname
Tipo di dato della colonna originale	
E' ammesso un valore non definito per la colonna	true
Valore standard per la colonna	
Cardinalità della colonna ARRAY	

Ricapitolando quanto detto finora:

Operazione	Dopo la normalizzazione	Dopo la de-normalizzazione
Primary Keys e Foreign Keys	Mantenute	Mantenute
Vincoli Not-Null	Mantenuti	Mantenuti
Tipi di dato "string"	Mantenuti	Mantenuti
Tipi di dato "numerici"	Mantenuti	Mantenuti
Viste	Mantenute le query che le generano, colonne di cui sono composte e breve descrizione	NON Mantenute
BLOBs in FileTables	Mantenuti solo metadati dei file	Mantenuti solo metadati dei file, non più riconosciuta la tabella come File Table
Vincoli Check	NON Mantenuti	NON Mantenuti

1.15.2. Destinazione: MySQL:

- Sono mantenute chiavi primarie ed esterne
- Vincoli NOT NULL mantenuti, vincoli CHECK non presenti e viste non mantenute:

Host: 127.0.0.1

Database: dbo1

Tabella: calciatore

Dati

Query*

Query #2* X

Base

Opzioni

Indici (2)

Chiavi esterne (1)

Verifica vincoli (0)

Partizioni

Codice CREATE

Aggiungi

Rimuovi

Pulisci

Nome chiave	Colonne	Tabella della c...	Colonne es...	Su aggiorn...	Su elimina...
FK_CALCIIATOR_Squad_29572725	Squadra	dbo1.squadra	Nome	NO ACTION	NO ACTION

Colonne:

Aggiungi

Rimuovi

Su

Giù

#	Nome	Tipo di dati	Lunghezza/set	Senza s...	Permett...	Riem...	Predefinito	Comn
1	CF	VARCHAR	20	☐	☐	☐	Nessun valore pr...	
2	Nome	VARCHAR	30	☐	☒	☐	NULL	
3	Squadra	VARCHAR	30	☐	☒	☐	NULL	
4	Ruolo	VARCHAR	10	☐	☒	☐	NULL	
5	Caratteristica	VARCHAR	100	☐	☒	☐	NULL	
6	anzianita	INT	10	☐	☒	☐	NULL	

- **Riscontrato problema di compatibilità riguardo i tipi VARCHAR e NVARCHAR:**

In MySQL, il limite per la lunghezza dei dati di tipo VARCHAR è di 256 caratteri, essendo questo limite posto ad 8000 caratteri in SQL Server, insorgono problemi di compatibilità quando nel DB sorgente ci sono colonne di tipo VARCHAR con lunghezza massima maggiore di 256 caratteri. Una tale colonna viene convertita in un campo TEXT, che però non può essere usato come Primary Key in quanto di lunghezza variabile. In questo caso viene generato un errore e non è possibile procedere con l'inserimento.

1.15.3. Destinazione Oracle:

- **Problemi nell'esportazione di Foreign Keys:**

Riscontrato comportamento anomalo riguardante le Foreign Keys: alcune di queste sono mantenute nel DB rigenerato mentre altre non sono presenti.

- **DB CALCIO:**

Delle 3 Foreign Keys presenti nella base di dati originale, soltanto una è mantenuta:

DB ORIGINALE:

dbo.CALCIATORE

- Colonne
- Chiavi
 - PK_CALCIATO_32149A78521B43E2
 - FK_CALCIATOR_Squad_29572725
- Vincoli
- Trigger
- Indici
- Statistiche

dbo.PARTITA

- Colonne
- Chiavi
 - PK_PARTITA_50B4A9DBAC663E72
 - FK_PARTITA_ARBITRO_2E1BDC42
 - FK_PARTITA_SQ_IN_C_2C3393D0
 - FK_PARTITA_SQ_IN_T_2D27B809

DB ESPORTATO SU ORACLE:

CONSTRAINT_NAME	CONSTRAINT_TYPE	SEARCH_CONDITION
1 FK_PARTITA_ARBITRO_2E1BDC42	Foreign_Key	(null)
2 SYS_C008274	Check	"SQ_IN_CASA" IS NOT NULL
3 SYS_C008275	Check	"SQ_IN_TRASFERTA" IS NOT NULL
4 SYS_C008276	Check	"Data_Sfida" IS NOT NULL
5 SYS_C008277	Primary_Key	(null)

CONSTRAINT_NAME	CONSTRAINT_TYPE	SEARCH_CONDITION	R_OWNER
1 SYS_C008272	Check	"CF" IS NOT NULL	(null)
2 SYS_C008273	Primary_Key	(null)	(null)

Da qui si nota come sia presente solo una delle 3 Foreign Keys nella tabella PARTITA, mentre nessuna è stata mantenuta nella tabella CALCIATORE.

- **DB SMARTPHONE:**

Questa volta tutte le Foreign Keys sono mantenute sul DB esportato su Oracle.

DB ORIGINALE:

dbo.ABBONAMENTO
Colonne
Chiavi
PK_ABBONAME_D62994E50E0459A4
FK_ABBONAMEN_CODCL_47DBAE45
FK_ABBONAMEN_CODSM_300424B4
FK_ABBONAMEN_CODTA_30F848ED
Vincoli
Trigger
Indici
Statistiche
dbo.ACQUISTO
Colonne
Chiavi
PK_ACQUISTO_63A1FC7297142477
FK_ACQUISTO_CODCL_2A4B4B5E
FK_ACQUISTO_CODSMA_2B3F6F97
FK_ACQUISTO_CODTAB_2C3393D0

DB ESPORTATO SU ORACLE

CONSTRAINT_NAME	CONSTRAINT_TYPE	SEARCH_CONDITION
1 FK_ABBONAMEN_CODCL_2F10007B	Foreign_Key	(null)
2 FK_ABBONAMEN_CODSM_300424B4	Foreign_Key	(null)
3 FK_ABBONAMEN_CODTA_30F848ED	Foreign_Key	(null)
4 SYS_C008390	Check	"NUM_SIM" IS NOT NULL
5 SYS_C008391	Check	"COSTO_MENSILE" IS NOT NULL
6 SYS_C008392	Primary_Key	(null)

CONSTRAINT_NAME	CONSTRAINT_TYPE	SEARCH_CONDITION
1 FK_ACQUISTO_CODCL_2A4B4B5E	Foreign_Key	(null)
2 FK_ACQUISTO_CODSMA_2B3F6F97	Foreign_Key	(null)
3 FK_ACQUISTO_CODTAB_2C3393D0	Foreign_Key	(null)
4 SYS_C008393	Check	"DATA" IS NOT NULL
5 SYS_C008394	Check	"CODCL" IS NOT NULL
6 SYS_C008395	Primary_Key	(null)

Sono mantenute tutte le FK del Database originale.

- **DB CONCORSO FOTOGRAFICO:**

In questo caso nessuna delle Foreign Keys viene mantenuta nel DB esportato.

DB ORIGINALE:

dbo.Fotografia
Colonne
Chiavi
PK_Fotograf_20DE9073D2D710FC
FK_Fotografi_idfot_267ABA7A
Vincoli
Trigger
Indici
Statistiche
dbo.Fotografo
dbo.Valutazione
Colonne
Chiavi
PK_Valutazi_BC89791F1974E73F
FK_Valutazio_idcon_2D27B809
FK_Valutazio_idfot_2C3393D0

DB ESPORTATO SU ORACLE:

CONSTRAINT_NAME	CONSTRAINT_TYPE	SEARCH_CONDITION
1 SYS_C008386	Check	"idfotografia" IS NOT NULL
2 SYS_C008387	Check	"idconcorso" IS NOT NULL
3 SYS_C008388	Check	"voto" IS NOT NULL
4 SYS_C008389	Primary_Key	(null)

CONSTRAINT_NAME	CONSTRAINT_TYPE	SEARCH_CONDITION
1 SYS_C008375	Check	"idfotografia" IS NOT NULL
2 SYS_C008376	Check	"tema" IS NOT NULL
3 SYS_C008377	Check	"idfotografo" IS NOT NULL
4 SYS_C008378	Primary_Key	(null)

Non è presente nessuna Foreign Key fra i vincoli del DB esportato su Oracle.

- **Anche qui problemi per quanto riguarda tipi VARCHAR ed NVARCHAR:**

Riscontrato problema analogo a quanto avveniva su MySQL, infatti anche qui sono diversi i limiti di lunghezza dei tipi VARCHAR ed NVARCHAR rispetto ad SQL Server. In particolare, questi limiti sono di 4000 caratteri per i VARCHAR2 (tipo analogo a VARCHAR di SQLServer) e 2000 caratteri per il tipo NVARCHAR2 (analogo a NVARCHAR in SQL Server). Se nel DB sorgente sono presenti colonne che violano questi limiti la rigenerazione della base di dati non va a buon fine.

Inoltre:

- **Non mantenute le viste ed i vincoli CHECK**
- **Mantenute Primary Keys, vincoli di NOT NULL e differenze fra i tipi di dato numerici**

1.15.4. Destinazione PostgreSQL:

- **Anche qui problemi riguardo le Foreign Keys:**

Riscontrato problema analogo a quanto avviene in Oracle: non tutte le Foreign Keys sono mantenute nelle basi di dati esportate. In particolare, però, le chiavi esterne mantenute sono diverse da quelle mantenute in Oracle:

- Non mantenute FK del DB CALCIO
- Mantenute le FK del DB SMARTPHONE e del DB CONCORSO FOTOGRAFICO

- **Nessun problema riguardo VARCHAR ed NVARCHAR:**

Su PostgreSQL, sia i tipi di dato VARCHAR che NVARCHAR vengono mappati nel tipo VARCHAR, che, non avendo un limite di massima lunghezza, non provoca incompatibilità nella fase di esportazione.

- **Mantenute le Primary Keys, i vincoli NOT NULL e le differenze fra i tipi di dato numerici**
- **Non mantenute viste e vincoli CHECK**

In sintesi, i risultati ottenuti esportando da SQL Server ed eseguendo l'import sui vari DBMS si hanno i seguenti risultati:

Operazione	SQL Server	MySQL	Oracle	PostgreSQL
1. Primary Keys	Mantenute	Mantenute	Mantenute	Mantenute
2. Foreign Keys	Mantenute	Mantenute	Mantenute in parte	Mantenute in parte
3. Vincoli NOT NULL	Mantenuti	Mantenuti	Mantenuti	Mantenuti
5. Tipi di dato "string"	Mantenuti	Problemi con lunghezze massime dei dati VARCHAR	Problemi con lunghezze massime dei dati VARCHAR	Mantenuti, ma non le differenze fra VARCHAR e NVARCHAR
6. Tipi di dato "numerici"	Mantenuti	Mantenuti	Mantenuti	Mantenuti
7. Vincoli CHECK	NON Mantenuti	NON mantenuti	NON mantenuti	NON mantenuti
8. Viste	NON mantenute	NON mantenute	NON mantenute	NON mantenute
9. File Tables (BLOB)	Mantenuti solo metadati dei file, non più riconosciuta la tabella come File Table	NON mantenute	NON mantenute	Mantenuti solo metadati dei file
10. Stored procedure e Trigger	NON Mantenuti	NON mantenuti	NON mantenuti	NON mantenuti

1. **Primary key:** le primary key sono sempre state mantenute in tutte le basi di dati analizzate;
2. le **foreign key** non sempre sono state ricreate, ma questo è scritto anche nella documentazione di SIARD, in quanto si vuole preservare la lettura;
3. i vincoli di **NOT NULL/NULL** sono mantenuti in generale, però c'è un problema con la base di dati di test di Microsoft "AdventureWorks2019". Nella colonna "schema" della tabella "dbo.DatabaseLog" nella base di dati originaria ammette valori NULL, nel file che crea SIARD sono ammessi valori nulli, ma quando SIARD ricrea la base di dati impone il campo come NOT NULL. Questo impedisce l'inserimento dei record originali bloccando l'importazione dell'intera base di dati. È possibile intervenire manualmente modificando le impostazioni sulla base di dati;
4. i **tipi di dato stringa** creano problemi nell'importare la base di dati su DBMS diversi. Ad esempio, MySQL ammette massimo 255 caratteri per il tipo VARCHAR, mentre SQL server ne ammette 8000. Su MySQL bisognerebbe usare il tipo TEXT, su cui però non si possono creare chiavi e questo crea altri problemi di incompatibilità;
5. i **tipi di dato numerici** non creano alcun problema;
6. le **viste** non vengono ricreate, ma è possibile leggere dal file creato da SIARD la query che è stata utilizzata per la loro creazione e quindi ricrearle manualmente;

7. le tabelle con tipo di dato **BLOB** presentano gli stessi problemi sulla lunghezza dei caratteri dei campi di tipo CHAR. In ogni caso, vengono mantenuti solo i path dei file e non i file stessi in quanto memorizzati su filesystem all'esterno della base di dati;
8. le **Stored procedure** e i **Trigger** non vengono mantenuti e non vengono riportate nel documento SIARD le istruzioni necessarie alla loro creazione come avviene per le viste.