Poli di Conservazione - Gruppo 2
La conservazione delle basi di dati

Prof. **Antonino Mazzeo**
Dip. di Ingegneria Elettrica e delle Tecnologie dell' Informazione
Università di Napoli Federico II
mazzeo@unina.it

Forum della Conservazione AgID
24-11-2022

Programma dell'evento

- **9.30/9.45 - Apertura dei lavori** (AgID)

- **9.45/11.00 – I gruppi di lavoro**

- Gruppo 1: Metadati, segnatura di protocollo e interoperabilità - Allegato 5 alle Linee Guida AGID – **Valeria Pavone** (Comune di Padova) *
- Gruppo 1: Metadati, segnatura di protocollo e interoperabilità - Allegato 6 alle Linee Guida AGID – **Raffaele Gonnella** (Centro di Dematerializzazione e Conservazione Unico della Difesa)
- Gruppo 2: Conservazione di basi di dati – La conservazione delle basi di dati - **Antonino Mazzeo** (Dip. di Ingegneria Elettrica e delle Tecnologie dell' Informazione - Università di Napoli Federico II)
- Gruppo 3 - Interoperabilità tra erogatori di servizi di conservazione - Gruppo 3 - UNI SInCRO vs Allegato 5 –problematiche e soluzioni per l'interoperabilità - **Gabriele Bezzi** (Responsabile della funzione archivistica di Conservazione Polo archivistico Regione Emilia Romagna (ParER))

- **11.00/12.00 – Best practices**

- Notariato – Evoluzione sistema di conservazione e conservazione repertori – **Luigi D'Ardia** (Notariato) *
- Conservazione Sogei –
- Monitoraggio e analisi di qualità del processo di conservazione - **Francesca Quai** (Responsabile della funzione archivistica - Sogei) *

- **12.00/12.30 - Dibattito e Conclusioni**

La conservazione delle informazioni

- La conservazione digitale a lungo termine mira a garantire l'accessibilità, l'autenticità, l'intelligibilità e l'integrità degli oggetti digitali per lunghi periodi che possono essere illimitati. È una grande sfida per le istituzioni che cercano di preservare le loro informazioni sensibili, patrimoniali o scientifiche, tra cui la comunità di biblioteche digitali, i data center e gli archivi

- informazioni digitali: varie forme di rappresentazione
  - documenti informatici
  - dei database
  - dei siti web
  - delle email e della messaggistica
  - dei social
  - ...
- Analisi degli specifici domini di informazione degli Enti e identificazione dei requisiti di conservazione per creare concreti use case da proporre

Normativa e standard di riferimento

- eIDAS, CAD, Linee Guida e norme/regolamenti su protocollo, gestione documentale, etc.
- Standard OAIS, Premis, …

La conservazione delle informazioni insita nel CAD e in e-IDAS

- Il Codice dell'Amministrazione Digitale (CAD-DLgs 82/2005) definisce il documento informatico come "rappresentazione informatica di atti, fatti o dati giuridicamente rilevanti", in contrapposizione al documento analogico "rappresentazione non informatica di atti, fatti o dati giuridicamente rilevanti".
- eIDAS: «**documento elettronico**», qualsiasi contenuto conservato in forma elettronica, in particolare testo o registrazione sonora, visiva o audiovisiva;
- eIDAS: i documenti elettronici sono importanti per l'evoluzione futura delle transazioni elettroniche trasfrontaliere nel mercato interno. Il presente regolamento dovrebbe stabilire il principio secondo cui a un documento elettronico non dovrebbero essere negati gli effetti giuridici per il motivo nella sua forma elettronica al fine di assicurare che una transazione elettronica non possa essere respinta per il solo motivo che un documento è in forma elettronica.
- eIDAS art.46°. Effetti giuridici dei documenti elettronici
  A un documento elettronico non sono negati gli effetti giuridici e l'ammissibilità come prova in procedimenti giudiziali per il solo motivo della sua forma elettronica.

Standard OAIS

- Anche le informazioni che costituiscono parte costitutiva dell’oggetto/contenuto informativo da conservare sono a loro volta oggetto di conservazione (non basta inserirle in un db descrittivo, ma vanno gestite in forme appropriate
- Non si deve distinguere e contrapporre (come fatto dal legislatore italiano) tra informazione e documento:
- Un’informazione, giuridicamente rilevante e che dispone di tutte le componenti costitutive necessarie (provenienza, integrità, data), sia pure diversamente accertabili (anche a livello di sistema), può avere, anzi deve avere natura documentale.

Linee guida agid 2020 e All.2

- 2.3 Dati strutturati

1. In questa sezione si descrivono brevemente alcuni formati dedicati al trasporto di dati strutturati, intendendo con questa accezione riferirsi a formati ove la tipologica di contenuto non è predeterminata a propri. Esempi di applicazioni che fanno uso di dati strutturati sono le basi di dati (per le quali rappresentano qui i formati SQL e quelli relativi all'applicativo Microsoft ® Access ®). Si sottolinea che l'adeguatezza di una base di dati alla normativa vigente in materia di protezione dei dati personali (p.es. pseudonimìa) e privacy può essere indipendente dal formato di file adottato, mentre è fortemente caratterizzata dai criteri architetturali adottati durante la fase progettuale.

Linee guida agid 2020 e All.2

- 2. I file SQL servono a contenere configurazioni e tabelle per basi di dati relazionali, complete o parziali, sotto forma del loro linguaggio di programmazione comune.
- Ogni file SQL descrive la formazione della base di dati, “da zero” o a partire da una base supposta già esistente: fornendo un tale file ad un gestore di basi di dati vengono perciò costituite le sue tabelle. Vice versa, una o più tabelle possono essere archiviate effettuandone uno “scarico” (dump in inglese) in un file SQL che descrive come il contenuto dello scarico può essere formato, da zero, in un nuovo gestore di basi di dati che interpreti il medesimo linguaggio. A tale scopo bisogna dunque specificare che SQL è in realtà un ceppo linguistico, da cui sono derivati molteplici dialetti di SQL, differenziati a seconda dell’applicativo –commerciale o meno– che funge da gestore di basi di dati. Si raccomanda quindi, per l’archiviazione a lungo termine e l’interscambio, di utilizzare sempre il formato SQL standardizzato dalla ISO (e riportato in tabella).
- 3. Ove non possibile, va sempre indicata la versione esatta del linguaggio SQL adottato, incluso preferenzialmente il nome e la versione completa dell’applicativo di gestione (MySQL, Microsoft® SQL™, ecc.).

Il documento elettronico (eIDAS) e informatico (CAD)

- Le definizioni presenti in eIDAS e nel CAD non si limitano al tradizionale concetto di documento digitale (inteso come stringa di bit) da far corrispondere all'equivalente rappresentazione cartacea, ma includono elementi di informazione riportabili, ad esempio, a quelle presenti in un db organizzate in record e tabelle.
- Non esiste una definizione univoca e semplice di documento, almeno per quanto riguarda l'archivistica e la gestione documentale: l'elemento distintivo riguarda la funzione giuridica che l'oggetto identificato come documento (qualunque sia la sua natura e il suo supporto) svolge e che ha come conseguenza la necessità di individuarne le componenti costitutive (diversamente trattabili naturalmente in base alla specificità dell'oggetto stesso: internamente o esternamente, in forme granulari o a livello di sistema)
- In OAIS non si parla mai di documenti, ma di informazioni e si distingue con molta nettezza tra informazioni e dati:
  - nessun dato è conservabile (merita di essere conservato) senza le informazioni relative (informazioni di rappresentazione - RepInfo e informazioni di conservazione PDI);
- La conservazione di tali informazioni non prettamente rappresentazioni documentali, a tutti gli effetti documenti informatici, deve, pertanto, essere oggetto di regolamentazione e di pianificazione negli enti

La conservazione delle Basi di Dati a lungo termine: metodi utilizzabili

- **Emulation**: si usa un emulatore del sistema originale, per un database relazionale si intende preservare il DBMS originale (software) e i dati;
- **Migration**: migrare il software vecchio a quello nuovo. Si esegue un dump del database e poi lo si ripristina su un nuovo DBMS;
- **Normalization:** i dati sono convertiti in un formato standard che tutti sanno leggere.
- Queste strategie presentano dei limiti:
- **Emulation**: aumento dei costi perché bisogna gestire vari database ed è necessario avere più amministratori di database (DBA);
- **Migration**: vi possono essere problemi di incompatibilità tra le varie versioni dei database, è necessario ricontrollare approfonditamente i dati importati;
- **Normalization:** non presenta alcun problema perché i dati sono in un formato standard e vengono migrati sul nuovo sistema.

Migrazione/normalizzazione

- L'approccio di migrazione/normalizzazione funziona esportando alcune proprietà del database originale dal suo DBMS in un altro DBMS o formato di file più adeguato per la 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 e dovrebbe supportare i requisiti di conservazione e i futuri utilizzi stabiliti durante la pianificazione della conservazione.
- Vincoli
  - La compatibilità con le versioni precedenti di un DB comporta l'utilizzo di versioni software e/o hardware più recenti per aprire, accedere e leggere un documento creato 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.

Migrazione/normalizzazione

- Tuttavia, l'approccio di normalizzazione rappresenta l'attuale migliore pratica quando si tratta di preservare i database.
- Il vantaggio principale della migrazione è la capacità di diffondere le risorse del database in un modo che sia facile da comprendere e riutilizzare per i suoi futuri utenti.
- Lo svantaggio principale è la potenziale perdita di informazioni a causa di formati di conservazione o di software di migrazione non inadeguati.
- Programmi di conservazione di database su larga scala per Enti pubblici sono in vigore da oltre 10 anni (UK, Svizzera, Danimarca, Svezia, etc.).
- Istantanee dell'esecuzione di database sono effettuate all'incirca ogni cinque anni.

Incapsulamento

- L'incapsulamento comporta la raccolta di documentazione sull'ambiente tecnologico di un database. Questa documentazione può includere manuali del Database Management System (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. Per questo motivo, l'incapsulamento è raramente usato da solo, ma serve piuttosto come attività fondamentale o un supplemento ad altri approcci, come l'emulazione o la migrazione (Digital Preservation Testbed, 2001; Faria, 2015, Ferreira, 2006).

Emulazione

- L'emulazione comporta la sostituzione di componenti software e/o hardware dello stack tecnologico del database 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 del database, uno dei vantaggi dell'utilizzo Emulazione - ma potrebbe anche essere uno svantaggio in quanto sia il tempo che le differenze tecnologiche 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 (Thibodeau, 2002; Waugh, Wilkinson, Hills e Dell'oro, 2000).
- L'emulazione può anche introdurre problemi per quanto riguarda le licenze software e i 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 i database sono distribuiti su una rete. In questi casi, l'intera configurazione di rete deve essere preservata dall'ambiente di emulazione.

Staticizzazione Ultima spiaggia…😢

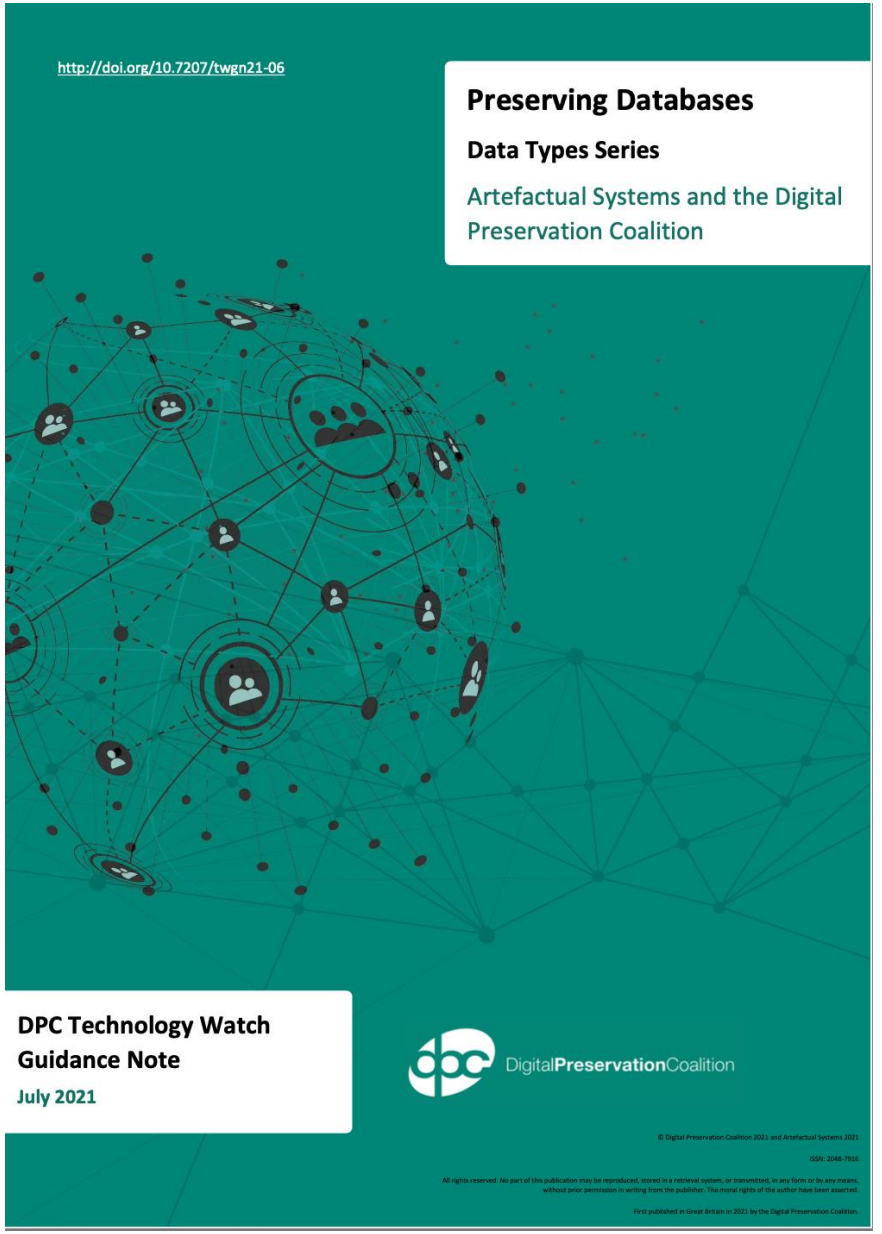
- La staticizzazione delle basi di dati con un piano di generazione di documenti informatici (in formati documentali PDF, PDF/A, …) analoghi ad un equivalente cartaceo
- Definire, sin dalla progettazione del sistema informatico, i documenti informatici da generare e da conservare e i lor formati
- Pianificare i punti di sincronizzazione periodica in cui operare con l'estrazione dai dati e la loro aggregazione in documenti informatici
- Definire i meccanismi per la generazione anche dei metadati necessari alla conservazione dei documenti informatici generati

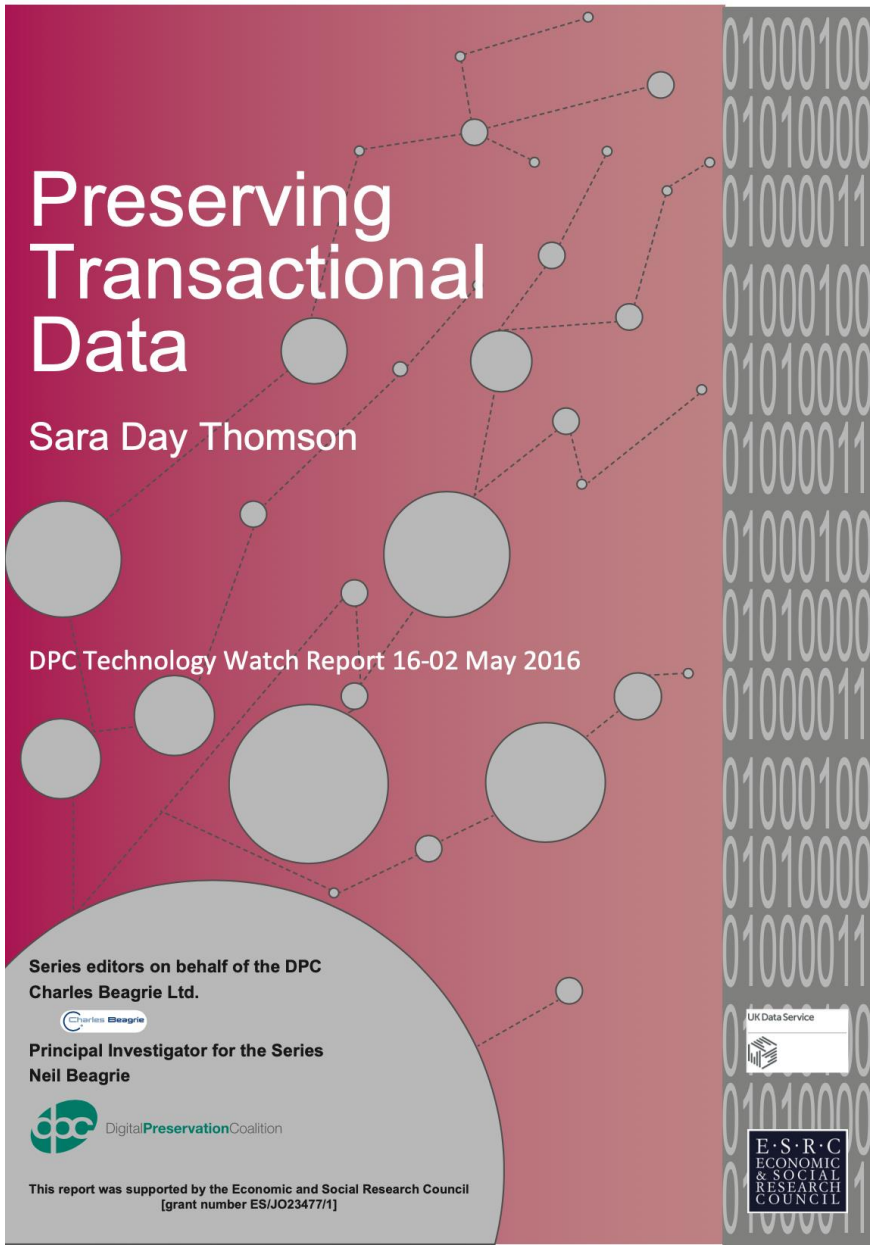
Alcuni progetti di database preservation
Analisi del contesto internazionale

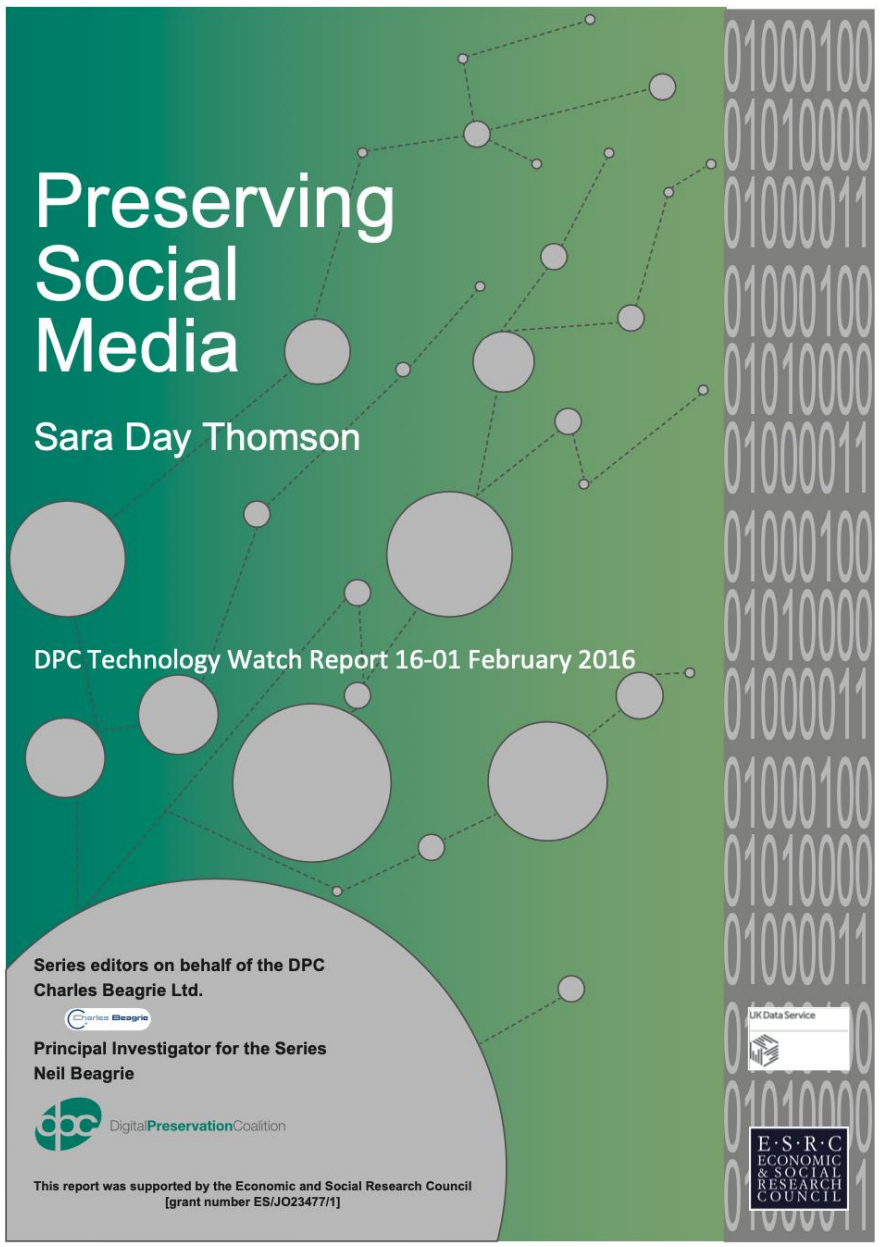
- Software independent archival of relational databases (SIARD)
- Software Database Preservation Toolkit (open-source, supports SIARD 2.0)
- Repository of Authentic Digital Objects (RODA)
- Digital Preservation Testbed
- Lots of Copies Keep Stuff Safe (LOCKSS) project was led by libraries at Stanford University
-

DigitalPreservationCoalition

- https://www.dpconline.org/about/members




Preserving Transactional Data
Sara Day Thomson
DPC Technology Watch Report 16-02 May 2016
Series editors on behalf of the DPC
Charles Beagrie Ltd.
Principal Investigator for the Series
Neil Beagrie
DigitalPreservationCoalition
This report was supported by the Economic and Social Research Council
[grant number ES/JO23477/1]
UK Data Service
E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL


Preserving Social Media
Sara Day Thomson
DPC Technology Watch Report 16-01 February 2016
Series editors on behalf of the DPC
Charles Beagrie Ltd.
Principal Investigator for the Series
Neil Beagrie
DigitalPreservationCoalition
This report was supported by the Economic and Social Research Council
[grant number ES/JO23477/1]
UK Data Service
E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

SIARD

- L’Archivio Federale Svizzero (AFS) ha fatto la scelta strategica di archiviare solo basi di dati relazionali. Come parte del progetto ARELDA, venne ideato e sviluppato un nuovo formato per l’archiviazione di basi di dati relazionali:
- il formato SIARD (Software-Independent Archiving of Relational Databases) fu presentato nel 2004 ed è stato ulteriormente sviluppato in seno nell’ambito del progetto PLANETS.
- Verso fine 2008 AFS pubblica la versione definitiva del formato SIARD e di un ambiente software a support del porting di una base di dati
- SIARD propone uno standard di **Normalization** per conservare i dati di un database per periodo di 50-100 anni
- è un formato unificato per eseguire il porting di un db in un formato intermedio standard e di ricorrere ad una singola applicazione per eseguire le attività di normalization/denormalization di un database, per far si che solo questa applicazione deve essere mantenuta aggiornata per consentire il collegamento a nuovi DBMS quando ne variano le specifiche nel tempo.
- è concepito con l’idea che un database è preservato se una query di selezione eseguita sul database originale e su quello ripristinato danno lo stesso risultato
- Uso di solo query di SELECT e non le altre forme, perché interessa solo preservare i dati**.**

SIARD-Specifiche/Informazioni:

- Memorizza un database 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
- 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

SIARD

- SIARD converte basi di dati proprietarie (MS Access, MS SQL, Oracle, etc.) nel formato non proprietario SIARD. L'archivio SIARD (con estensione di file .siard) rappresenta la base di dati nella sua struttura logica, mantenendo non solo i dati primari e i metadati, ma soprattutto le relazioni

- 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 preservazione.

- 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.

- Il gruppo di lavoro Agid ha fatto un attenta valutazione dei pro-contro per testare la concretta applicabilità di SIARD a varie tipologie di basi di dati. I risultati di tale analisi saranno allegati al documento che sarà prodotto

Repository of Authentic Digital Objects (RODA)

- un progetto avviato in Portogallo nel 2006 dagli Archivi nazionali portoghesi, al fine di preservare quegli oggetti digitali prodotti dalle istituzioni governative portoghesi. Il progetto mirava a combinare diversi tipi di oggetti digitali in un unico repository, compresi i database relazionali.
- Come repository singolare di molti diversi tipi di oggetti digitali, RODA mira a normalizzare tutti gli oggetti ingeriti, cioè a ridurre al minimo i tipi di formato utilizzati per archiviare documenti e conservare documenti simili in formati simili.
- Il progetto RODA ha sottolineato la creazione di un metodo standardizzato per preservare i database come oggetti digitali. La conservazione del database rappresenta una sfida unica in quanto il processo di conservazione è suddiviso in tre livelli: dati, struttura (logica) e semantica (interfaccia).
- dati delle banche dati, così come la loro struttura e semantica, devono essere conservati. Al fine di preservare tutti e tre questi elementi, il progetto RODA ha sviluppato il Database Preservation Toolkit.

La conservazione dei registri e repertori un utile e interessante esempio concreto di conservazione di una semplice base di dati

- Casi d’uso di interesse
  - I registri di protocollo
  - conservazione dei repertori notarili digitali
  - I registri di protocollo
  - La digitalizzazione dei verbali d’esame

Esempio Verbalizzazione esami universitari

- Creazione di un registro verbale digitale conservato su di un sistema qualificato
- **Macro fasi del processo:**
- prenotazione
- Apertura sessione d'esame con autenticazione forte (SPID) docente presidente sessione
  - Appello da lista prenotati ed eliminazione assenti
  - Avvio esami (sessione autenticata e profilata)
  - per ogni studente si registra sempre una linea /record con indicazione esito (positivo/ rinuncia/ bocciatura). Ogni linea è messa in un'area di stage qualificata e trasferita definitivamente nel sistema di conservazione qualificato allo scadere del timout (2-3gg) per eventuali ricorsi dopo notifica.
  - La linea avrà un TS di sistema e il check del presidente che, unitamente all'accesso spid (firma avanzata), suggella l'esame.
- Tutti gli studenti della sessione sono memorizzati come linee indipendenti, accorpate in un verbale di sessione logica a partire dalla ID della sessione d'esame e dai ts.
- A fine esame il Professore chiude la sessione e avviene la notifica del verbale logico ai membri della commissione e agli studenti (per la linea di loro pertinenza)
- La segreteria può estrarre il verbale logico intero o singole linee d'esame a partire dall' ID di sessione del professore E matricola studente

Conclusioni

Proposta di indice del report che il gdl AGID sta scrivendo

- Metodi per la conservazione delle basi di dati
- Analisi del contesto internazionale (progetti e ricerche scientifiche) relativamente alla conservazione delle informazioni e, in particolare, a quella della conservazione dei DB:
- Razionale per la valutazione della complessità del DB origine per selezionare una soluzione per la conservazione nel tempo e vincoli di cui tenere conto
- Documentazione della base di dati di interesse
- Aspetti tecnologici della base di dati di interesse
- Modalità di utilizzo futuro della base di dati storicizzata
- Valutazione della qualità dello schema e dei dati delle basi di dati da storicizzare

Punti rilevanti che saranno trattati nella relazione

- Aspetti archivistici e giuridici della conservazione delle basi di dati
- Normativa sulla conservazione dei documenti informatici
- Vincoli di privacy
- Aspetti giuridici (valore probatorio dei dB) e di privacy (GDPR)
- Criteri e suggerimenti per la messa a punto di un piano di conservazione delle informazioni di un ente a seconda del dominio delle informazioni e dei metodi scelti per conservarle
- Rassegna delle varie tipologie di DB presenti nelle PP:AA. e relative necessità di conservazione
  - Complessità (n.ro tabelle, relazioni, etc.)
  - Criticità delle informazioni e aspetti di sicurezza
  - …

In Italia che si è fatto e si sta facendo?

- Gruppi di ricerca attivi nelle università e centri nazionali di Archivistica e Informatica svolgono attività e partecipano a progetti europei e internazionali sul tema o seguono attività in specifiche preservation coalition
- Poco a livello istituzionale
- Scarsa o assenza di pianificazione delle azioni miranti alla preservazione del patrimonio dati e assenza di norme e regoamenti specifici
- Necessità di costituire gruppi di competenza sulla preservazione dell'Informazione e, in modo particolare, sulla conservazione dei DB relazionali multidisciplinari
- AGID?
  - In seno al GdL sulla rete dei Poli di Conservazione ha attivato un gdl su tale tema di cui alla presente relazione
  - Cosa fare dopo?
- Indispensabile una forte azione di sensibilizzazione di tutti gli stackolder