JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

JLIS.it

Metadata profiles for interoperability. The E-ARK specifications for e-archiving*

Pasqualina Adele Marzotti

Contact: Pasqualina Adele Marzotti, linamarzotti@gmail.com
Received: 9 March 2021; **Accepted:** 24 April 2021; **First Published:** 15 September 2021

ABSTRACT

The paper is a reflection on the importance of solutions for interoperability in European contest and compares the AGID guidelines with the DILCIS Board e-archiving building block. It describes E-ARK eco-system, analysing the E-ARK Specifications for Common Information Package. The other E-ARK specifications for SIP, AIP and DIP are quickly presented, as is the Content Information Type Specification for SIARD (CITS SIARD). Finally, the author shows the success stories on implementation of CITS SIARD; among these study cases we can recently find the Italy's Central States Archives project. The aim is to participate to the discussion on Italian rules and guidelines for record management and digital preservation, imagining an Italy's open digital preservation model.

KEYWORDS

E-ARK; E-archiving; Metadata; Digital preservation; Interoperability; Sustainability; Archivio centrale dello Stato.

CITATION

Marzotti, P.A. "Metadata profiles for interoperability. The E-ARK specifications for e-archiving." *JLIS.it* 12, 3 (September 2021): 105–118. DOI: 10.4403/jlis.it-12714.

* Il contributo costituisce una sintesi aggiornata della tesi discussa dall'autrice a conclusione del master in Formazione gestione e conservazione degli archivi digitali in ambito pubblico e privato, AA. 2018-2019.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

Dati, informazioni, conoscenza[1]

In un contesto globale, dati e informazioni rivestono un ruolo di primo piano per l'economia, per la sanità, per la sicurezza pubblica e più in generale per la condivisione della conoscenza. In una dimensione globale, infatti, la tutela di dati integri e informazioni autentiche costituisce oggi un irrinunciabile strumento di e-government e democrazia. Come potremmo definire il complesso di dati e informazioni di questo tipo, scaturiti dall'attività di organizzazioni diverse, se non archivi digitali?
La conservazione dei dati e delle informazioni con valore documentale rappresenta senza dubbio la sfida più grande per gli Archivi del XXI secolo, che in virtù delle loro finalità istitutive sono responsabili della custodia e garanti del diritto di accesso dei cittadini agli archivi digitali (Vitali 2010, 36-61). Come garantire l'accesso alle informazioni nel tempo e al tempo stesso mantenere l'integrità e l'autenticità dei dati? Le soluzioni tecniche implementate sino ad oggi determinano la necessità di trasferire i dati di continuo sia da un formato a un altro, sia da un sistema a un altro, attraverso processi di migrazione più o meno complessi. Sfuggire a questa necessità è impossibile: l'obsolescenza tecnologica di formati e software, nonché i processi di conversione, riversamento o migrazione rappresentano un rischio concreto per l'integrità delle informazioni.
Dal punto di vista metodologico, la soluzione al problema sta nella progettazione di sistemi, procedure e processi affidabili e nella gestione dei dati all'interno di file-contenitori corredati da metadati che li rendano autoconsistenti e indipendenti dall'ambiente di produzione. Da questa prospettiva, il momento centrale del ciclo di vita delle informazioni è quello della loro conservazione in un deposito digitale: le informazioni vengono immesse e vengono diffuse a partire dalla memoria digitale di un *repository* e di conseguenza il metodo di gestione dei dati all'interno di questa memoria condiziona tutte le fasi del processo. Il modello generalmente adottato prevede quindi la formazione di *information package* (IP) comprendenti dati e informazioni gestionali e descrittive sui dati (OAIS, 2012).
Il biennio appena trascorso ha portato con sé importanti novità per chi si occupa di gestione informatica dei documenti e conservazione digitale. Sia in Italia che in Europa, infatti, sono stati completati i processi di consultazione pubblica e revisione preliminari alla pubblicazione di alcuni documenti di riferimento per il settore: si tratta delle nuove *Linee guida sulla formazione, gestione e conservazione dei documenti informatici* di AGID e delle specifiche E-ARK per l'interoperabilità dei pacchetti informativi definite dal DILCIS Board.
La conservazione digitale va intesa come una gestione attiva del patrimonio documentario digitale al fine di garantirne l'autenticità, l'affidabilità e l'accesso nel tempo. Per questo motivo sia AGID che il DILCIS-Board perseguono l'obiettivo di fornire alle PA, alle imprese e alla cittadinanza strumenti utili

[1] I termini "dati" e "informazioni", utilizzati in questo contributo singolarmente o in congiunzione, non sono da considerarsi sinonimi. In informatica, un dato è una unità elementare, che nell'ambito di un sistema informativo potremmo definire come una entità memorizzabile (un *bit* è un dato, un *bit set* è un insieme di dati); un'informazione è l'insieme di alcuni dati memorizzati sul sistema e delle relazioni e inferenze esistenti tra questi dati. La conoscenza si colloca ad un livello di complessità superiore, che comporta la capacità di interpretare informazioni in relazione al loro utilizzo e al contesto (o ai contesti). Il tema è complesso e supera i confini della disciplina, per questo la bibliografia è davvero estesa e abbraccia contributi scientifici prodotti in oltre un secolo in tutti i paesi industrializzati. Senza scomodare Bertalanffy, Shannon, Ashby e molti altri, in questa sede è interessante fare riferimento a Wessels *et aliii* (2017, 25–44).

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

a realizzare una completa interoperabilità tra sistemi e contenuti digitali, tutelando al tempo stesso l'integrità, l'autenticità e dunque l'affidabilità dei dati oggetto di conservazione.
Le soluzioni individuate da queste due istituzioni, pur basandosi sugli stessi modelli, derivano da strategie diverse e da diversi livelli di applicazione: quella italiana è incentrata sulla definizione di procedure e processi documentati e certificati, oltre che sull'assunzione di precise responsabilità da parte dei soggetti coinvolti, mentre quella europea si concentra sulla definizione di un modello in grado di armonizzare standard di metadati già esistenti e ampiamente diffusi in ambito internazionale, dando per acquisita la capacità delle organizzazioni coinvolte di implementare procedure e processi affidabili. A mio avviso, ci troviamo di fronte a soluzioni complementari.

Italia - Europa

In riferimento al problema dell'interoperabilità dei pacchetti informativi contenenti documenti informatici,[2] sin dal 2013 l'AGID sostiene lo sviluppo e l'applicazione di un profilo di metadati per l'interscambio dei pacchetti informativi[3] all'interno di un archivio OAIS. A settembre 2020, dopo una lunga consultazione pubblica avviata il 14 dicembre 2019, sono state approvate le nuove *Linee guida* con cui AGID prescrive l'utilizzo di IP conformi allo standard UNI 11386:2020 SInCRO per il trasferimento di oggetti digitali dal sistema di gestione documentale a quello di conservazione. L'interoperabilità degli oggetti digitali nel sistema di gestione documentale (documenti o aggregazioni documentali) è realizzata invece attraverso lo schema di metadati definiti dall'*Allegato 5* (AGID 2020). Si tratta di indicazioni che in parte confermano quelle contenute nelle precedenti *Regole tecniche*[4] e che convalidano soluzioni applicative implementate già da tempo da molte pubbliche amministrazioni e da molti conservatori accreditati (Pigliapoco 2019).
Una delle novità delle *Linee guida* è rappresentata dall'introduzione di un processo di "valutazione di interoperabilità" basata su procedure da definirsi nel manuale di gestione che prevedano sia l'analisi periodica dei formati digitali utilizzati, con l'attribuzione di un "indice di interoperabilità" a ciascun formato, sia il riversamento dei file originali in formati più adatti.
Riassumendo, il problema dell'interoperabilità è considerato dall'AGID a tre diversi livelli: quello logico dei metadati, quello fisico dei formati file utilizzati e quello generale (logico e fisico) dei pacchetti informativi. Le diverse strategie di intervento devono essere armonizzate dalle procedure descritte nel manuale di gestione e in quello di conservazione e le responsabilità ripartite tra le varie funzioni organizzative (responsabile della gestione, della conservazione, della privacy, etc.).
In sostanza, l'enfasi è posta sui processi, più che sui prodotti: l'integrità e l'autenticità delle informazioni dipendono prima di tutto dalla piena realizzazione dei processi definiti nel manuale di gestione e nel manuale di conservazione.
Quanto ai "prodotti", se nel nostro Paese è avvertita l'urgenza di definire raccomandazioni per la formazione di pacchetti normalizzati in grado di veicolare la conservazione di documenti digitali come, ad esempio, referti, reperti e altri documenti formati in ambito sanitario, fatture elettroniche,

[2] Per la definizione di documento informatico, cfr. D.Lgs. 82/2005, art. 1, co. 1, lett. p.
[3] Cfr. Circolare 60/2013, oggi abrogata.
[4] Cfr. il DPCM 3 dicembre 2013 "Regole tecniche in materia di sistema di conservazione", all. 4.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

documenti protocollati, atti e provvedimenti, registri e repertori (UNI 11386:2020, 37–73), non sembra invece così chiaro come IP di tal fatta possano considerarsi interoperabili a livello transnazionale, anche allo scopo di condividere dati e informazioni specifiche nell'ambito di programmi comunitari. In proposito si pensi, ad esempio, al problema attualissimo della disponibilità di dati clinici e statistici utili alla ricerca e alla programmazione strategica per fronteggiare a livello europeo l'emergenza pandemica.

Guardando all'orizzonte europeo, l'attività di ricerca e sviluppo per lo scambio e l'archiviazione di dati autentici è affidata al Digital Information Life Cycle Interoperability Standards Board (DILCIS Board), un organismo indipendente che ha per obiettivo lo sviluppo e il mantenimento di specifiche tecniche a supporto dell'interoperabilità: specifiche pubbliche e disponibili anche per implementazioni personalizzate, secondo un modello di ricerca scientifica basato sulla condivisione dei saperi.

Il progetto E-ARK4ALL diretto dal DILCIS Board 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 nel raggiungimento di una ampia interoperabilità, da realizzare partendo dal livello generale e astratto di un *common information package* da specificare in base alle esigenze contingenti, che possono essere determinate sia dal diverso tipo di contenuto incluso in un IP, sia dalle caratteristiche del processo in cui un IP è coinvolto partecipando ad una o all'altra fase del ciclo di vita delle informazioni.

I progetti E-ARK4ALL ed E-ARK3

Il progetto *European Archival Records and Knowledge Preservation* (E-ARK), avviato nel 2014 dalla Commissione europea nell'ambito dell'ICT Policy Support Programme (PSP) e del Competitiveness and Innovation Framework Programme (CIP), si è concluso nel 2017 raggiungendo alcuni traguardi importanti per la definizione di una metodologia condivisa per la conservazione digitale, basata sugli standard internazionali e nazionali maggiormente diffusi e utilizzati, oltre che su alcune buone pratiche attuate nel contesto europeo (Bredenberg *et alii* 2018). E-ARK realizza i principi di sviluppo digitale, condivisione della conoscenza e riuso dettati da Horizon2020. La necessità di mantenere nel tempo e di diffondere tali risultati, nell'ottica di un ciclo di miglioramento continuo, ha portato all'inclusione del progetto all'interno delle attività del Connecting Europe Facility (CEF) Programme, finanziato dall'Innovation and Networks Executive Agency (INEA).[5] I risultati progettuali di E-ARK (linee guida, modelli, metadati, tool, etc.), implementati e perfezionati nell'ambito del progetto E-ARK4ALL sviluppato nel triennio 2017-2020 in continuità con il progetto precedente, costituiscono l'*eArchiving Building Block*, un insieme di strumenti (*Digital Service Infrastructures*) da utilizzare per la realizzazione di servizi pubblici digitali.

In questa cornice istituzionale, la ricerca di E-ARK si sviluppa a partire dalla consapevolezza dell'importanza che riveste la tutela dell'autenticità, della garanzia di accesso e della possibilità di riuso delle informazioni nel lungo periodo. Per realizzare tutto questo, E-ARK definisce un set di specifiche tecniche a supporto dell'interoperabilità dei pacchetti informativi, ossia raccomandazioni

[5] European Climate, Infrastructure and Environment Executive Agency (CINEA) dal 1° aprile 2021.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

utili a mantenere l'integrità e l'autenticità dei dati realizzando la loro migrazione tra diversi sistemi di produzione e di gestione, il loro trasferimento in un *digital repository* e la loro conservazione digitale, l'accesso e il loro riuso nel lungo periodo. Il modello è progettato con l'obiettivo di superare sia i limiti dettati dalla tecnologia di produzione dei dati, sia quelli scaturiti dai processi di produzione, gestione e accesso ai dati. In quest'ottica, le *Common Specification for Information Packages* (CSIP), basate sul *Metadata Encodings and Transmission Standard* (METS), rappresentano l'archetipo da cui derivano i profili di metadati per la formazione dei pacchetti di versamento, di archiviazione e di distribuzione.

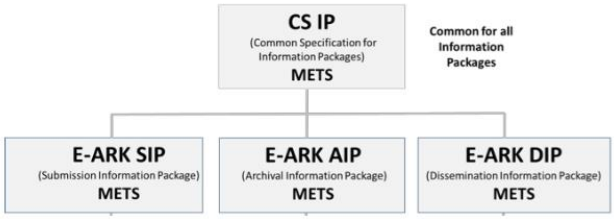

Fig. 1. E-ARK e METS (DILCIS 2019a, 3)

Su questo archetipo si fonda l'interoperabilità tra produttori di dati, loro conservatori e utilizzatori, produttori di soluzioni software per la formazione, la conservazione e il riuso dei dati, ovvero tra chi ha la necessità di tutelare l'autenticità dei dati nel lungo periodo, chi assume la responsabilità di questa tutela e chi è in grado di fornire soluzioni tecnologiche adeguate a realizzarla.
L'obiettivo è la definizione e il mantenimento di un set di specifiche tecniche a supporto dell'interoperabilità degli IP, che integrino e definiscano con un elevato grado di dettaglio i due standard più largamente utilizzati nell'ambito dell'intero ciclo della conservazione digitale: METS e PREMIS (*PREservation Metadata Implementation Strategies*). Altro obiettivo del DCLIS Board, non meno importante è quello di sviluppare linee guida e raccomandazioni per includere negli IP specifici tipi di contenuto, come ERMS, GIS, database o altro (chiaramente, adottando soluzioni scalabili). La definizione di questo *set* di specifiche rappresenta uno degli aspetti più caratteristici del progetto E-ARK4ALL.
I risultati finora raggiunti sono il punto di partenza per il nuovo progetto E-ARK3, nell'ambito del quale il DILCIS Board coordina alcuni *interest group* formati da soggetti interessati allo sviluppo delle specifiche E-ARK e impegnati nella loro applicazione.

Common Specification for Information Packages (CSIP)

L'ecosistema del progetto E-ARK4ALL si sviluppa in tre livelli: le CSIP, le raccomandazioni E-ARK per SIP, AIP e DIP e le *Content Information Type Specification* (CITS). Le specifiche per la formazione di E-ARK SIP, AIP e DIP sono una estensione delle CSIP, mentre le CITS, potenzialmente numerose quanto numerosi possono essere i tipi di contenuti da trattare nell'ambito di un processo di

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

conservazione digitale, costituiscono un approfondimento di dettaglio relativo al contenuto incluso in un IP.

L'orizzonte di riferimento di E-ARK è quello rappresentato da ciò che potremmo definire come il ciclo di vita delle informazioni, originato dalla creazione (o ricezione) di dati e informazioni e scandito nel corso del tempo da diversi eventi, legati ad attività di registrazione, uso, riuso, migrazione, versamento e accesso, che determinano il trasferimento dei dati dai sistemi di gestione a quelli di conservazione e viceversa. In questo scenario è importante che dati e informazioni possano esser trasferiti da un sistema all'altro senza comprometterne l'integrità e l'autenticità e senza che ne siano limitate le possibilità di ricerca e di riutilizzo.

La soluzione per garantire tutto questo, conformemente al modello OAIS, è quella di documentare le attività di gestione, trasmissione e conservazione digitale attraverso l'uso di metadati specifici per ogni fase del processo: trasmissione degli IP tra diversi sistemi di gestione documentale o a un deposito per la conservazione digitale, gestione degli IP all'interno del sistema di conservazione e loro eventuale trasmissione al di fuori del sistema in risposta ad una richiesta di accesso specifica. L'elemento di novità introdotto da E-ARK4ALL è l'implementazione di un profilo di metadati comune a tutte le "fasi di vita" di un IP, comprendente elementi e attributi di base che devono o possono essere compresi da SIP, AIP o DIP. L'obiettivo è quello di realizzare una base comune, proponendo delle "*universal common specification*" che possano essere implementate in qualsiasi sistema per supportare il modo in cui dati e metadati devono essere strutturati negli IP trasmessi, custoditi, usati o riutilizzati all'interno di sistemi informativi diversi.

Le CSIP, a partire dai principi generali,[6] veicolano a livello teorico una completa interoperabilità tra ambiente di produzione, ambiente di conservazione digitale e ambiente di ricerca e accesso. Un E-ARK IP può includere qualsiasi tipo di dato e di metadato appartenenti a qualsiasi ambito di applicazione, non pone restrizioni a strumenti, metodi e applicativi di scambio né definisce uno specifico metodo di conservazione, è sia *machine-readable* che *human-readable*, può essere scalabile, deve essere identificato in modo univoco sia dal punto di vista del processo (SIP, AIP, DIP) che dal punto di vista del contenuto (*content type*), deve essere contraddistinto da identificativi univoci (per l'IP, validi in un *repository* e/o in una rete di sistemi, e per i suoi componenti, validi nel *repository*), deve avere una struttura logica e una struttura fisica determinate e deve essere corredato da metadati standard sottoposti a processi di analisi, disambiguazione e validazione.

Una delle parti più rilevanti delle CSIP è quella dedicata ai requisiti per la definizione della struttura di un IP, che prevede la separazione di dati e metadati sia dal punto di vista logico che dal punto di vista fisico e che si realizza distribuendo gli uni e gli altri in altrettante cartelle fisiche. Questa soluzione consente di ridurre l'impegno necessario a identificare e validare il contenuto di un IP e quindi di semplificarne la conservazione a lungo termine, anche attraverso il trattamento separato di porzioni selezionate di dati e metadati.

Questo "principio di separazione", adottato ad ogni livello della struttura, serve a gestire in modo più efficace gli IP. Infatti, oltre a separare e raggruppare in modo adeguato i file di dati e metadati all'interno di un IP, in questo modo è possibile sia supportare operazioni di *splitting* di IP di grandi

[6] Le E-ARK CSIP definiscono 21 caratteristiche di base: principi generali (*Principles* 1.1-1.7), identificazione IP (*Principles* 2.1-2.5), struttura IP (*Principles* 3.1-3.6), metadati (*Principles* 4.1-4.3) (DILCIS 2019a, 17–25).

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

dimensioni, sia includere (almeno da un punto di vista logico) all'interno dello stesso IP diverse rappresentazioni degli stessi dati.

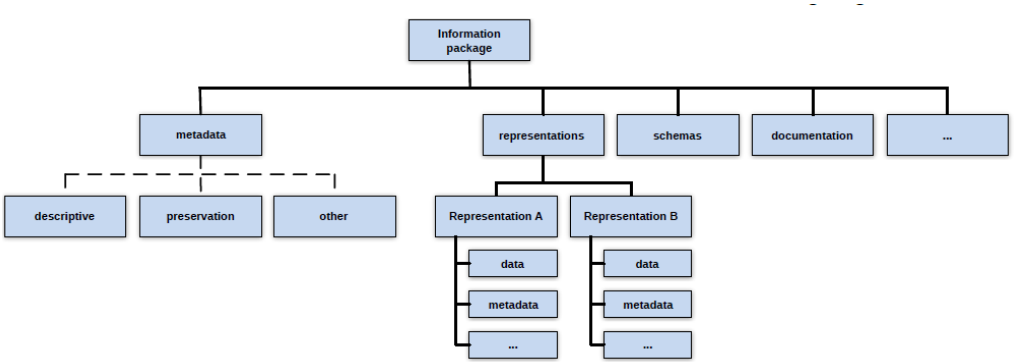

Fig. 2. Struttura concettuale delle *Common Specification* (DILCIS 2019a, 23)

Un altro aspetto cardine del modello concettuale progettato da E-ARK è quello che potremmo definire "principio di interoperabilità semantica", che informa ogni livello della struttura concettuale di un E-ARK IP. Si tratta dell'utilizzo di metadati standard, *schema* e vocabolari necessari a raccogliere tutte le informazioni descrittive e quelle di contesto e di processo in grado di documentare la "storia" delle informazioni contenute in un IP e dell'IP stesso: metadati di trasmissione implementati secondo lo standard METS,[7] metadati di conservazione implementati in base al modello proposto da PREMIS,[8] metadati descrittivi implementati in base allo standard EAD, altri metadati individuati da ogni singola istituzione. Tali metadati costituiscono una componente logica essenziale: devono essere compresi nel pacchetto sia a livello generale (*root folder*) sia all'interno delle sezioni che contengono i dati separati in base alla loro specifica rappresentazione. Le E-ARK CSIP individuano gli elementi di METS, PREMIS ed EAD funzionali al modello (*crucial metadata*, DILCIS 2019a, 24), ma non escludono la possibilità di ricorrere ad ulteriori metadati di settore.[9] Questa ricchezza di informazioni sui processi e sui contesti in cui sono coinvolti gli IP, nonché sulla loro struttura e sulla loro realizzazione tecnica, va quindi governata attraverso il ricorso a standard internazionali valutati e validati in modo da non dar luogo ad ambiguità. La corretta distribuzione all'interno degli IP di metadati cruciali, ad un livello strutturale elevato, e degli altri eventuali metadati descrittivi specifici

[7] Le E-ARK CSIP comprendono 114 requisiti per l'implementazione di *core metadata* basati sul METS *schema* v. 1.2; in un E-ARK IP sono utilizzati alcuni elementi cardine di METS: 'mets', 'metsHdr', 'dmdSec', 'amdSec', 'fileSec' e 'structMap'. L'elemento 'dmdSec' serve per i metadati descrittivi, mentre l'elemento 'amdSec' e i sub elementi 'digiprovMD' e 'rightsMD' servono per i metadati gestionali. L'elemento 'fileSec' serve per elencare i singoli file contenuti nel pacchetto, identificandoli, validandoli, localizzandoli ed eventualmente raggruppandoli in maniera opportuna. (DILCIS 2019a, 32–59).

[8] Le E-ARK CSIP raccomandano l'utilizzo di PREMIS 3.0, *PREMIS Data Dictionary e Standard Identifiers Scheme* per i metadati di conservazione. I metadati PREMIS possono essere allocati in una cartella specifica e referenziati con l'elemento 'amdSec/digiprovMD' nel 'package METS.xml' o nel 'representation METS.xml' più pertinente. PREMIS può essere usato per registrare metadati tecnici, informazioni su *agent* ed *event*, diritti specifici per gli oggetti digitali conservati e sui formati file (DILCIS 2019a, 59–60).

[9] Metadati di diverso tipo possono essere implementati in base a esigenze specifiche, purché siano documentati inserendo i relativi tracciati nella cartella 'schema' (DILCIS 2019a, *ibid.*).

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

per il dominio e il tipo di contenuto, nei livelli strutturali più interni (in modo da evitare ridondanze), è in grado di dare sufficienti garanzie per l'interoperabilità degli IP e l'automazione dei processi di trasmissione, conservazione e accesso.

Morfologia di un E-ARK IP

Le CSIP propongono quindi una struttura concettuale per l'implementazione tecnica di E-ARK IP: il pacchetto, articolato in più livelli, dal generale al particolare, deve contenere a livello generale i metadati rilevanti per l'intero pacchetto e dovrebbe essere articolato in cartelle e sottocartelle che rendano manifesta la struttura logica generale, raggruppando i metadati secondo le tipologie impiegate e separandoli dalle rappresentazioni dei dati in tante componenti strutturali quante siano necessarie. Tale struttura ripropone in maniera ricorsiva la separazione tra metadati (EAD, METS, PREMIS, etc.) e dati, che sono gestiti in base alla loro rappresentazione (*representation*), e prevede anche la gestione di *schema*, documentazione aggiuntiva ed eventuali altri dati di corredo e di contestualizzazione definiti in base a particolari esigenze di conservazione o di accesso.
Il risultato dell'implementazione delle CSIP è un IP complesso, ma strutturato in modo molto semplice. Solo due componenti sono obbligatorie: la *root folder* e il file di metadati *Package METS*, che serve a referenziare tutti gli elementi contenuti nel pacchetto. Tuttavia, possono verificarsi situazioni in cui risulti più pratico definire più dettagliatamente un IP di grandi dimensioni o contenente un elevato numero di *digital obiect*, separando i dati in base alla loro rappresentazione e i metadati relativi a ogni rappresentazione in diversi file METS.xml (*Representation METS*) inclusi nelle sottocartelle annidate nella *root folder* e richiamati in modo opportuno all'interno del METS.xml principale. In pratica i metadati possono essere registrati sia in corrispondenza della radice della struttura di un pacchetto, sia all'interno delle sezioni contenenti le diverse rappresentazioni. È quindi possibile produrre un file *Package METS.xml* e tanti file *Representation METS.xml* quante sono le *representation* all'interno di un IP.[10]

[10] La struttura dell'IP deve essere descritta nel file *Package METS* utilizzando l'elemento obbligatorio 'structMap'. Le E-ARK CSIP definiscono 16 requisiti di struttura (DILCIS 2019a, 25, 51–59).

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

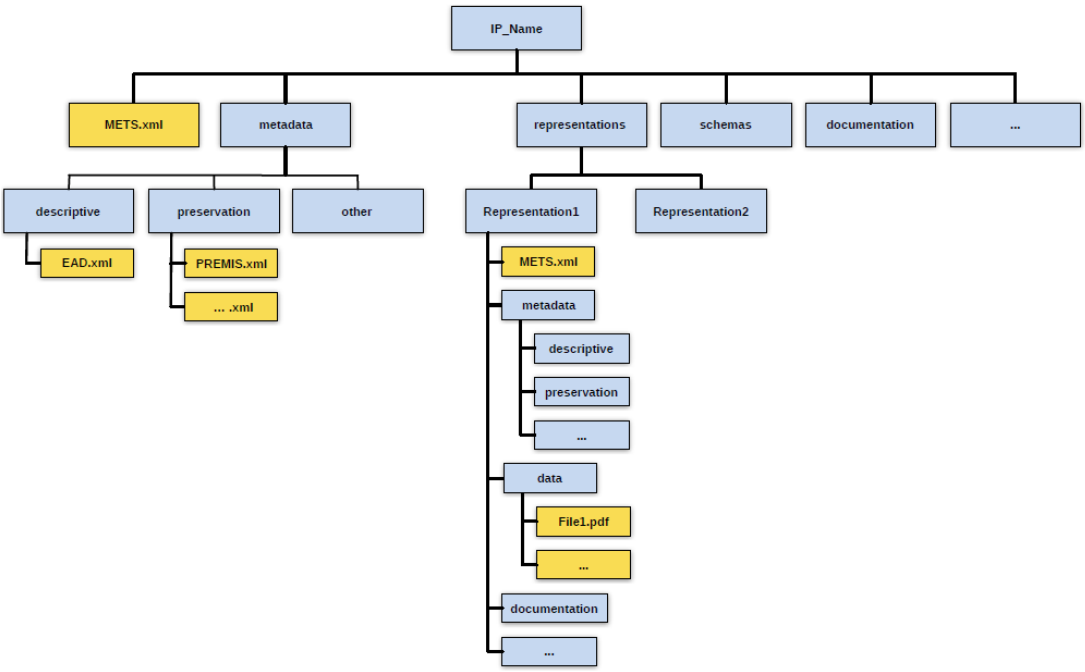

Fig. 3. Struttura fisica di un IP CSIP (in giallo gli elementi minimi, DILCIS 2019a, 26)

L'obiettivo delle CSIP è supportare l'interoperabilità a livello tecnico, controllando i metadati necessari ad un agente (applicativo o utente) per navigare tra le componenti (dati e metadati) di un IP, verificare l'integrità dei dati e dei metadati sottoposti a trasferimento o conservazione, conoscere i processi di produzione, migrazione e conservazione che li coinvolgono, conoscere le modalità di accesso ai dati. I metadati previsti dalle CSIP devono quindi essere in grado di supportare tecnicamente qualità come l'autenticità dei dati e attività come la ricerca e l'accesso agli stessi.
L'utilizzo di una struttura fisica formata da cartelle e sottocartelle e ben documentata, come quella proposta a modello, dà l'opportunità di raggiungere un elevato grado di efficienza e soluzioni tecniche sostenibili. Ad esempio, attività di analisi specifiche possono essere realizzate solo sui file contenuti in determinate cartelle di un IP, mentre attività come il riversamento dei file in formati più adatti alla conservazione può essere realizzata con procedure massive solo sui file contenuti nella sezione *representation* della struttura (o, eventualmente, anche sui file contenuti nella sezione *documentation*) e prevedere la memorizzazione dei file prodotti al termine del processo di riversamento all'interno di una nuova sezione *representation*.

Specifiche tecniche derivate dalle CSIP

Il profilo di metadati disegnato dalle CSIP possiede tre estensioni, che specificano i metadati necessari a supportare le fasi di *submission*, *ingest / preservation* e *dissemination*. Elemento comune a tutti i profili è la presenza di un file *Package METS* che contiene metadati descrittivi (all'interno dell'elemento 'dmdSec') e amministrativi (nell'elemento 'amdSec'), o il rinvio a specifici file di metadati contenuti nell'IP (cartelle 'metadata/descriptive', 'metadata/preservation', etc.). Nel *Package*

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

METS, inoltre, si trova l'indice di tutti i file contenuti in un E-ARK IP (elemento 'fileSec'), di cui possono essere specificati nome, formato e versione, PUID.
Le *E-ARK Specification for SIP*[11] supportano l'interoperabilità dei pacchetti nella fase di trasmissione da un soggetto produttore a un soggetto conservatore e contengono i metadati e gli attributi previsti dalle CSIP per la fase di versamento degli IP. Il *Package METS* di un E-ARK SIP contiene nell'elemento *header* informazioni relative agli accordi di versamento, agli *agent* coinvolti nel processo e allo stato del pacchetto, in modo da chiarire se i dati siano trasmessi per la prima volta o in aggiornamento a dati trasmessi precedentemente.
Le *E-ARK Specification for AIP*[12] definiscono la struttura e i metadati per gli IP trasferiti in un *repository*, dalla fase di acquisizione (*ingest* di un SIP) alla fase di conservazione digitale (*preservation* di un AIP). Un E-ARK AIP si caratterizza per la capacità di governare la sedimentazione dei dati selezionati per la conservazione digitale, grazie alla possibilità di memorizzare all'interno dello stesso AIP n-*representation* degli stessi dati, conservando integri sia i dati acquisiti con un SIP (che viene memorizzato integralmente), sia i dati sottoposti a migrazione o conversione, sia i dati eventualmente elaborati in formati specifici per l'accesso. Un E-ARK AIP, sfruttando i meccanismi previsti dalle CSIP per lo *splitting* e la completa referenziazione di dati e metadati, è quindi in grado di documentare la "storia" dei dati con riguardo sia alla cosiddetta fase attiva del loro ciclo di vita, sia alle fasi di conservazione e accesso.
L'accesso alle informazioni conservate in un *repository* è supportato dalle *E-ARK Specification for DIP*,[13] con cui è possibile implementare IP contenenti una *representation* dei dati determinata dalle caratteristiche del sistema di accesso, corredati da tutti i metadati PREMIS sugli *agent* e sui *rights* che regolano l'accesso alle informazioni.
Il profilo di metadati generale delineato dalle CSIP può essere utilizzato per implementare IP contenenti documenti, immagini, audio, video, ERMS record, GIS record, GIS, ERMS, database, etc. Come anticipato, per gestire i diversi tipi di informazione in modo omogeneo, E-ARK introduce il concetto di *Content Information Type Specification* (CITS), raccomandazioni per la gestione di tipologie particolari di dati all'interno di un IP, insieme alla documentazione e ad altre componenti binarie (software, emulatori, etc.) necessarie per riprodurre i contenuti archiviati. Di conseguenza, implementare un E-ARK IP comporta la creazione di pacchetti conformi sia ai requisiti individuati dalle CSIP, sia a quelli individuati dalle CITS, profili di metadati specifici per gestire all'interno degli IP particolari *Content Data Object*.
Le CITS possono essere definite sia in relazione alla tipologia di contenuto (database, dati scientifici, mappe digitali, etc.) sia al dominio (archivistico, bibliotecario, medico, economico, etc.). Come detto, esistono delle tipologie di dati che rivestono una importanza strategica per il governo dell'Unione, per via del loro impiego in settori governativi e della loro diffusione. La definizione delle specifiche per

[11] Le E-ARK SIP comprendono 38 requisiti che permettono di estendere i file METS.xml utilizzando particolari attributi negli elementi 'root' e 'header' e di utilizzare gli elementi 'dmdSec', 'amdSec' e 'structMap' per informazioni pertinenti alla formazione di DIP (DILCIS 2019b).
[12] Le E-ARK AIP comprendono 47 requisiti per AIP che consentono di gestire il *versioning* del SIP, rappresentazioni multiple degli stessi dati e LOB (DILCIS 2019c).
[13] Le E-ARK DIP comprendono 9 raccomandazioni. Il profilo di un E-ARK DIP si caratterizza per le informazioni su *Package identifier*, *METS Profile*, *OAIS Package type information* e *Status of descriptive metadata* (DILCIS 2019d).

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

queste tipologie rientra tra le attività del DILCIS Board che, valorizzando l'esperienza dei soggetti partner, fino a questo momento si è occupato dello sviluppo di raccomandazioni per tre diversi *Content Information Type*: ERMS,[14] GIS e geodata,[15] database (mentre sono ancora allo studio le raccomandazioni per contenuti informativi appartenenti alla sfera dell'e-Health).
Per ovvie ragioni, come ad esempio il problema di garantire la sostenibilità del progetto, lo sviluppo delle CITS non può essere mantenuto sotto la responsabilità esclusiva del DILCIS, che tuttavia può svolgere funzioni di coordinamento, raccogliendo profili di metadati proposti da altri soggetti in base alle indicazioni e ai requisiti definiti nell'ambito del progetto E-ARK, in modo tale da armonizzare tra loro le specifiche tecniche sviluppate direttamente dal DILCIS Board con quelle sviluppate eventualmente da altri soggetti a livello nazionale. È questo il caso, ad esempio, di SIARD e delle *Content Information Type Specification for Relational Databases using SIARD* (DILCIS 2020).
Gran parte di informazioni, dati e metadati prodotti e gestiti con ERMS o GIS sono trattati attraverso database relazionali (RDB) ed è quindi possibile, oltre a conservare i singoli oggetti digitali gestiti all'interno del sistema (documenti digitali o altro), conservare digitalmente sia i record sia i dati prodotti all'interno di sistemi informativi simili prima trasferendoli in un RDB e poi gestendo il processo di conservazione del database. Possiamo quindi affermare che la chiave di volta del complesso di raccomandazioni E-ARK elaborate per il trattamento e l'inclusione di specifici *content information type* all'interno di un IP è rappresentata dalle CITS SIARD.

Success stories (casi di studio)

Nell'ambito del nuovo progetto E-ARK3 il gruppo di lavoro impegnato nella sperimentazione di SIARD e delle CITS SIARD è arrivato recentemente alla pubblicazione dei primi risultati dell'implementazione (E-ARK3 2021a e 2021b). I casi di studio proposti descrivono i passaggi fondamentali dei processi di conservazione che utilizzano le CITS SIARD, implementati da National Archives of Norway, National Archives of Estonia, National Archives of Finland, Danish National Archives e Archivio federale svizzero. Queste CITS sono incentrate sull'utilizzo all'interno di un E-ARK IP del formato *Software-Independent Archival of Relational Database* (SIARD) sviluppato dall'Archivio federale svizzero tra il 2008 e il 2019. Nel 2013 SIARD è stato dichiarato un e-Government Standard (eCH-0165) e successivamente il progetto è stato incluso tra le attività di E-ARK, coinvolgendo una platea più vasta di ricercatori, sviluppatori e archivisti di diversi paesi europei[16] che hanno contribuito allo sviluppo della versione 2.1.1.
SIARD supporta la normalizzazione dei contenuti di un database di qualsiasi tipo in un database relazionale conforme al formato standard SQL:2008, utilizzando diverse tabelle per rappresentare dati, campi, e relazioni del database originario; inoltre, SIARD supporta anche l'utilizzo di regole di

[14] Le E-ARK for ERMS comprendono una mappatura con le entità definite da MoReq2010, raccomandazioni sui metadati da utilizzare per il trasferimento di *Content Information file* di tipo ERMS, uno XML *schema* e uno *schematron*, che definiscono elementi, *complex type*, attributi e regole per la gestione di IP contenenti ERMS (DILCIS 2019e).
[15] Le E-ARK for GEO IP definiscono gli elementi contenuti in un Geodata AIP, elaborano alcuni esempi e propongono due tracciati di interoperabilità tra i metadati INSPIRE e i metadati ISAD(G) ed EAD3 (DILCIS 2019f).
[16] Il gruppo di lavoro è formato da membri di organizzazioni pubbliche e private di Svizzera, Danimarca, Estonia, Inghilterra, Norvegia, Portogallo, Slovenia, Svezia e Ungheria.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

validazione per struttura e dati in file XML, è in grado di gestire i *Large OBject* (LOB) inclusi nel database, supporta il metodo di compressione *deflate* e può essere aperto con qualsiasi applicazione adatta al formato ZIP64. Utilizzare SIARD comporta altri vantaggi, come la possibilità di utilizzare gli strumenti *open source* messi a disposizione dall'Archivio federale svizzero: una *SIARD Suite*, che comprende un *toolkit* per la conversione di vari DBMS nel formato SIARD o il ripristino di database conservati su DBMS (Database Preservation Toolkit - DBPTK) e un *toolkit* per visualizzare i database basati su SIARD (db-visualization-toolkit).[17]

Il *Relational Database Archiving Interest Group* ha definito a livello generale un processo di conservazione che si sviluppa in quattro fasi (*appraisal*, *predelivery*, *ingest* e *access*), che comprendono ogni attività intrapresa per il trasferimento dei dati dal soggetto produttore al soggetto conservatore: analisi preliminare e definizione di un accordo di versamento, selezione dei dati, formazione del SIP e sua trasmissione, analisi e validazione del SIP da parte del conservatore, conversione del SIP in AIP, accesso ai dati tramite strumenti condivisi. Per ognuna di queste attività, sono poi riassunte le diverse azioni messe in atto, realizzate dal punto di vista tecnico sia grazie ai *tool* dell'Archivio federale svizzero, sia grazie ad altri applicativi proprietari realizzati da specialisti IT coinvolti dagli altri archivi nazionali.

I risultati raggiunti sono diversi e mettono in evidenza quanto sia importante l'esperienza maturata nell'ambito di progetti di cooperazione internazionale, tanto per lo sviluppo di buone pratiche e linee guida da parte della pubblica amministrazione, quanto per lo sviluppo di soluzioni *in house*, tanto per la formazione di professionalità adeguate, quanto per lo sviluppo di soluzioni tecniche anche da parte di privati.

L'esperienza dell'archivio nazionale estone che implementa SIARD da appena quattro anni, ad esempio, è particolarmente interessane sia per le tipologie di database trattate, sia per le soluzioni per l'accesso alla documentazione che sono state implementate. L'attività di valutazione e selezione messa in atto dagli archivisti dei soggetti produttori e da quelli dell'archivio nazionale è spinta sino alla definizione di *viste* che rappresentano una estrazione dei dati di un database originale giudicati interessanti per la ricerca e immediatamente disponibili per la consultazione (perché liberi da obblighi di riservatezza). Ogni database è considerato un fondo archivistico e la descrizione archivistica necessaria alla sua contestualizzazione storico-istituzionale è arricchita da documentazione video che mostra il funzionamento del database nella sua fase di gestione attiva (E-ARK3 2021, 10).

Uno sguardo al domani

La lettura dei casi di studio presentati dal *Relational Database Archiving Interest Group* è interessante e suggestiva sotto diversi punti di vista. Emerge chiaramente come l'attività di analisi, comprensione e descrizione dei contesti di produzione e uso dei database oggetto di conservazione sia affidata alla competenza degli archivisti, chiamati ad intervenire in *team* con gli altri professionisti coinvolti sin dalle fasi iniziali di valutazione dei database e di progettazione dei processi di conservazione digitale. Emerge come processi di conservazione di questo tipo coinvolgano una quantità così elevata di dati da rendere necessario l'impegno di numerosi professionisti ben qualificati per realizzarli.

[17] La SIARD Suite è disponibile su GitHub: http://github.com/sfa-siard.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

Anche in questo contesto, l'archivista può quindi svolgere la funzione di mediatore tra il ricercatore e la memoria conservata, in quanto protagonista dell'attività di valutazione e selezione e componente (umana) essenziale per la realizzazione dell'interoperabilità tra sistemi e contenuti e soluzioni di conservazione digitale all'avanguardia. È suggestivo e stimolante immaginare come le valutazioni compiute oggi, possano non solo guidare la ricerca di domani, ma determinare addirittura il complesso di informazioni che costituirà per una buona parte la memoria storico-istituzionale di questo millennio.

ERMS, GeoData e GIS, database, e-Health: gestione e amministrazione, territorio, ambiente e trasporti, sanità. Gli ambiti in cui trovano applicazione i risultati dell'*e-archiving building block* rendono evidente una precisa linea d'azione: fornire specifiche tecniche a supporto dell'interoperabilità di IP contenenti oggetti digitali prodotti nei contesti strategici per la politica dell'Unione.[18] Non si tratta soltanto di perseguire lo scopo di definire e mantenere *common information type specification* dedicate ad una rosa sempre più ampia di tipologie di oggetti digitali. Si tratta di sviluppare strumenti autorevoli e condivisi per tutelare l'identità di una federazione di Stati, nella misura in cui anche i prodotti tecnologici e il patrimonio informativo generato nell'attuazione delle politiche europee possono essere considerati evidenze materiali di questa Identità.

L'Italia non è rimasta indietro. In anticipo sui tempi sul piano normativo, ha saputo tenere il passo dei paesi europei aderenti al progetto E-ARK3, formulando le prime proposte teoriche per la conservazione digitale dei documenti dello Stato già alcuni anni fa (Guercio 2014, 303-318). E il progetto per la realizzazione di un servizio di conservazione digitale dell'Archivio centrale dello Stato (Trani 2019) è stato recentemente incluso tra le "*success stories*" di E-ARK (CEF Digital 2020). La partecipazione al progetto rappresenta una opportunità unica: la tradizione archivistica italiana e l'esperienza maturata negli ultimi anni nella definizione di metodi e strumenti per l'interoperabilità potrebbero arricchire il modello europeo, dando al tempo stesso al nostro Paese la possibilità di armonizzare con il modello E-ARK il complesso di raccomandazioni, standard, norme e prassi archivistiche che regolano la formazione, gestione e conservazione dei documenti digitali.

Le sfide da affrontare sono molto serie: come armonizzare il complesso di norme e linee guida che nel nostro Stato definiscono formati, processi e metadati di conservazione con un modello aperto e condiviso come E-ARK? È certamente auspicabile che i risultati dell'implementazione delle specifiche E-ARK da parte dell'Archivio centrale dello Stato spingano alla definizione di nuove *Linee guida* caratterizzate da una maggiore apertura.

Riferimenti bibliografici

AGID. 2020. *Linee guida sulla formazione, gestione e conservazione dei documenti informatici*. Roma: AGID.

[18] Cfr. i verbali di riunione del DILCIS Board su http://github.com/DILCISBoard/GroupDocumentation/blob/master/MeetingNotes.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12714

Bredenberg, Karin, Kuldar Aas, David Anderson e Jaime Kaminski. 2018. "The application of E-ARK tools for archival interoperability to support a long-term sustainable Digital Single Market". In *Proceedings of the 15th International Conference on Digital Preservation*. Boston: iPRES 2018.

DILCIS Board. 2019a. *Common Specification for Information Packages*, v. 2.0.2, 28 ott. 2019.

Id. 2019b. *E-ARK SIP Specification for Submission Information Packages*, v. 2.0.2, 28 ott. 2019.

Id. 2019c. *E-ARK AIP Specification for Archival Information Package*, v. 2.0.1, 9 sett. 2019.

Id. 2019d. *E-ARK DIP Specification for Dissemination Information Packages*, v. 2.0.2, 28 ott. 2019.

Id. 2019e. *E-ARK Electronic Record Management System (ERMS)*, v. 2.0.0, 31 mag. 2019.

Id. 2019f. *E-ARK Specification for digital geospatial data records archiving*, v. 2.0.0, 31 mag. 2019.

Id. 2020. *Specification for the E-ARK Content Information Type Specification for Relational Databases using SIARD (CITS SIARD)*, Review draft, 20 ott. 2020.

CEF Digital. 2020. "eArchiving Used as Italy's Reference Model in Permanent Digital Preservation". *CEF Digital News*, 27 nov. 2020. https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/News.

CEF eArchiving Building Block, E-ARK3. 2021a. *Preserving databases using SIARD. Case Study 1. Experiences with workflows and documentation practices RDB SIARD*, v. 1.0.

Id. 2021b. *Preserving databases using SIARD. Case Study 2. Experiences working with large databases and their preservation*, v. 1.0.

Guercio Maria. 2014. "Il futuro digitale degli archivi e il ruolo dell'Archivio centrale dello Stato: una riflessione sui rischi per la tutela dei patrimoni documentari dello Stato e del Paese". In *1943-1953. La ricostruzione della storia*, 303–318. Roma: Archivio centrale dello Stato.

Pigliapoco, Stefano. 2019. "La conservazione digitale in Italia. Riflessioni su modelli, criteri e soluzioni". *JLIS.it* 10, 1 (gennaio). DOI: 10.4403/jlis.it-12521.

Trani, Silvia. 2019. "Il progetto dell'ACS: dal *Repository* al *Polo di conservazione*". Workshop Documento elettronico, 6 nov. 2019. Torino: ANAI. http://www.documento-elettronico.it/workshop/workshop-2019/30-contenutiit/workshop-2019/193-il-progetto-dell-acs-dal-repository-al-polo-di-conservazione.

Vitali, Stefano. 2010. "La conservazione a lungo termine degli archivi digitali dello stato". In *Conservare il digitale*. A cura di Stefano Pigliapoco, 36–61 Macerata: eum.

UNI 11386:2020. *Supporto all'Interoperabilità nella Conservazione e nel Recupero degli Oggetti digitali (SInCRO)*. Roma: UNI.

Wessels, Bridgette e Rachel L. Finn, Kush Wadhwa, Thordis Sveinsdottir, Lorenzo Bigagli, Stefano Nativi, Merel Noorman. 2017. *Open Data and the Knowledge Society*, Amsterdam: Amsterdam University Press.