# Archiviazione Documentale – il SERVIZIO

CILEA offre a partire da ottobre 2009 un servizio di Archiviazione Documentale, basato sullo standard OAIS.

Ricordiamo brevemente che con **Archiviazione Documentale si intende qui il processo di raccolta ordinata e sistematica di risorse digitali secondo determinati criteri e procedure, che garantisca il mantenimento di tutte quelle informazioni di carattere descrittivo e gestionale che possano assicurare l'interpretazione dell'informazione nel suo complesso** (contenuto, elementi, contesti)[1], secondo la definizione CNIPA. Più nel dettaglio intendiamo qui per "*Archiviazione Documentale*" l'archiviazione di **dati** e **metadati** atti a descriverli su di una architettura hardware e software adeguata ed in linea con lo standard OAIS. In buona sostanza, intendiamo quindi qui un sistema che archivi i dati garantendone la conservazione su un periodo di tempo medio-lungo (max. 10 anni), mantenendo consistenti i dati e i metadati ad essi associati, soddisfacendo requisiti di tipo OAIS ma non necessariamente requisiti legali (da soddisfare a seconda delle richieste del Cliente).

Senza entrare qui nello specifico dello standard ISO OAIS (Open Archival Information System), un sistema di AD deve permettere di:

- archiviare un set ampio di tipologie di formati
- caricare (***ingest***) gli OBJ sia in modalità singola che batch
- tracciare la "storia" di un OBJ archiviato
- ricercare OBJ nell'archivio, in modo da trovare sempre un OBJ desiderato
- garantire l'integrità dell'OBJ
- garantire la fruibilità dell'OBJ per il tempo dichiarato di archiviazione, provvedendo quindi sistemi di portabilità dei formati[2].

A questi requisiti vanno ad affiancarsi quelli normativi italiani in merito alla conservazione documentale a lungo termine. Senza entrare anche qui nel dettaglio, possiamo comunque affermare che a quanto detto in precedenza, la normativa aggiunge alcuni requisiti legati alla validità legale che si vuole dare all'oggetto archiviato. In particolare, se si devono preservare documenti con validità legale (fatture,

1 **Conservazione Sostitutiva** è invece il processo attraverso il quale è possibile conservare i documenti in modo che non si deteriorino e che risultino disponibili nel tempo mantenendo la loro autenticità e integrità (caratteristiche garantite da due azioni fondamentali: apposizione di firma digitale e marca temporale al documento o insieme di documenti che si desidera conservare.

2 In generale, si definiscono a priori comunque formati standard sia per il documento che per il metadato, quali PDF, TIFF, XML, che sono, ad oggi, i più portabili.

protocollo), gli oggetti andranno **digitalmente firmati con una firma a validità legale** e marcati con una **marca temporale a validità legale**.

A tal fine il sistema, a seconda delle necessità del Cliente, è in grado di soddisfare quanto indicato dal ***Tavolo tecnico Interministeriale - Aspetti tecnologici della conservazione permanente***, senza tuttavia fornire, al momento, funzionalità di dematerializzazione, protocollo informatico e fatturazione elettronica.

## Come funziona il servizio

Il sistema CILEA si basa sul middleware OpenSource ***Fedora-Commons (FC)***.

Fedora (Flexible Extensible Digital Object Repository Architecture), originariamente sviluppato dall'Università di Cornell, è una architettura per archiviare, gestire e accedere ad un qualche "contenuto digitale" nella forma di **oggetto digitale**. Fedora definisce un layer astratto di descrizione dell' OBJ, definendone relazioni con altri oggetti e servizi.

Tecnicamente, si tratta di una engine di webapps Java (implementato con Tomcat) e un RDBMS che contiene le informazioni necessarie al funzionamento di Fedora stesso, nonché alcune minime informazioni sull'OBJ stesso (quali identificativo univoco, path ai file su file system, alcuni campi *Dublin Core*).

Fedora-Commons è supportata da una ampia community internazionale, ed è ormai ampiamente diffuso in ambiti molto eterogenei [5] che vanno dalle biblioteche, ai consorzi universitari, alle università, alle realtà più strettamente IT (basti citare per queste lo svizzero SWITCH). Tutte ne hanno dimostrato la robustezza, la flessibilità e la scalabilità.

**Struttura dei metadati.**

Nella struttura di FC l'oggetto digitale è concepito come un insieme costituito da diversi componenti, identificato da un *persistent identifier* (PID) e dotato di una serie di metadati. I diversi componenti di un oggetto digitale sono definiti *datastream*. Ogni oggetto inserito in Fedora può avere uno o più TAG ***<datastream...>*** all'interno dei quali sono conservati il contenuto informativo e i metadati relativi all'oggetto digitale.

Al momento dell'inserimento di un nuovo oggetto digitale in Fedora vengono automaticamente generati 4 *datastream*:

1. DC (record Dublin Core che contiene metadati sull'oggetto, in caso di risorse risultato di un processo di digitalizzazione metadati sull'oggetto "fonte")
2. AUDIT (contiene traccia di tutte le modifiche relative ad un oggetto)
3. RELS-INT (relazioni interne relative cioè ai diversi *datastream* relativi al medesimo oggetto digitale)
4. RELS-EXT (descrive le relazioni con altri oggetti)
5. Datastream relativo alla risorsa digitale archiviata (ad es., file PDF)

All'interno di FC, i metadati relativi agli oggetti digitali, e a tutti i loro *datastream,* sono conservati in formato XML, secondo una specifica estensione dello standard *Metadata Encoding and Transmission Standard* (METS) 1.0, che è quella in grado di garantire una maggiore interoperabilità.
Preme qui sottolineare che la definizione della tipologia e del contenuto dei metadati è da stabilire e definire di caso in caso a seconda delle esigenze del Cliente. CILEA, dato il suo elevato livello di competenze in materia, può fornire consulenza a questo riguardo.

**Ingest**

L'ingest degli oggetti digitali avviene all'interno del server FC una volta che i dati vi sono stati resi disponibili. A seconda delle necessità e della realtà del Cliente, il sistema può gestire ingest di dati che risiedono presso cliente, purché resi (temporaneamente) accessibili via protocollo http/https, dati che Cliente fornisce direttamente a CILEA. In entrambe i casi CILEA può sviluppare le procedure per la creazione dei metadati e l'ingest vero e proprio.

**Indicizzazione**

Il sistema offre un sistema di indicizzazione base, con il quale è possibile cercare un oggetto digitale in base alle componenti di AUDIT e DUBLIN-CORE del metadato.
Su richiesta del Cliente è possibile affiancare a questa, un sistema di indicizzazione full-text basato su Lucene, che permette la ricerca su qualsiasi stringa di testo contenuta nell'oggetto digitale archiviato, se applicabile.

**Firma e marcatura temporale**

Su richiesta del Cliente o in base alle disposizioni di legge applicabili alla tipologia di dati trattati, è possibile affiancare un sistema nel quale gli oggetti digitali vengono firmati digitalmente con un

certificato digitale a piena validità legale rilasciato a CILEA; eventualmente si può anche apporre un time-stamp o una marca temporale all'elenco di file oggetto della singola operazione di ingest.

Il servizio offre nativamente un sistema di verifica di integrità all'atto dell'ingest di datastream relativi a file da archiviare. Appoggiandosi al RDBMS interno di FC e alle primitive di ricerca (REST) è stato inoltre implementato un sistema di verifica periodica dell'integrità e della consistenza dell'archivio. Nel dettaglio, giornalmente viene estratto dal DB un campione casuale di file, in ragione di circa il 10% del data sample, e ne viene calcolato il checksum MD5 su file system e confrontato con quanto salvato nel DB. Ogni differenza genera un allarme e una notifica e-mail allo staff di supporto Analogamente, vengono cercati nel DB file scelti a caso sul file system, al fine di verificare la consistenza del DB stesso.
Sebbene la normativa vigente non imponga siti di disaster-recovery per sistemi di AD, nemmeno per quelli a valore legale o fiscale, è ormai nel "sentire comune" che per alcune tipologie di documenti, essa venga quantomeno prevista o offerta.
La natura stessa del sistema di ingest permette facilmente di popolare in real-time almeno due repository che quindi possono essere uno il Disaster Recovery site dell'altro. Inoltre FC offre anche la possibilità di fare un export in modalità *Archive* che include nel file XML metadato anche un "dump binario" del datastream associato al file archiviato (ad es, file PDF). Ovviamente quest'ultima soluzione è percorribile solo nel caso di datastream di dimensione limitata e/o in casi di *disaster recovery off-line* di archivi medio-piccoli.
Qualora il Cliente decida di usufruire anche del servizio di disaster recovery, CILEA è in grado di offrirlo off-site presso la propria sede di Roma.
Per quanto riguarda invece le politiche di back-up, questi saranno back-up full e incrementali per la componente funzionale del server e back-up incrementali e full su apposito pool di cassette, con particolari regole di retention (10 anni).

## Contatti

Per tutte le informazioni relative al servizio e le relative condizioni economiche, contattare Dr.a Paola Tentoni (tentoni@cilea.it) o Dr. Matteo Boschini (boschini@cilea.it) tel. 02 269951.