CINECA

CodeX[ml] – Digital Library Management System

Premessa

Codex[ml] è un sistema sviluppato da CINECA per risolvere in maniera integrata le esigenze di gestione, conservazione e fruizione via web di risorse culturali digitali.
Codex[ml] trae il suo nome dall'accostamento tra il termine latino 'CODEX' (manoscritto, codice) e il termine informatico 'XML' (Extensible Markup Language); il prodotto consente, infatti, a qualunque ente che disponga di oggetti digitali di garantirne la gestione, la navigazione on line, la conservazione a lungo termine e la distribuzione a pagamento, attraverso l'utilizzo di standard di metadati amministrativo gestionali.

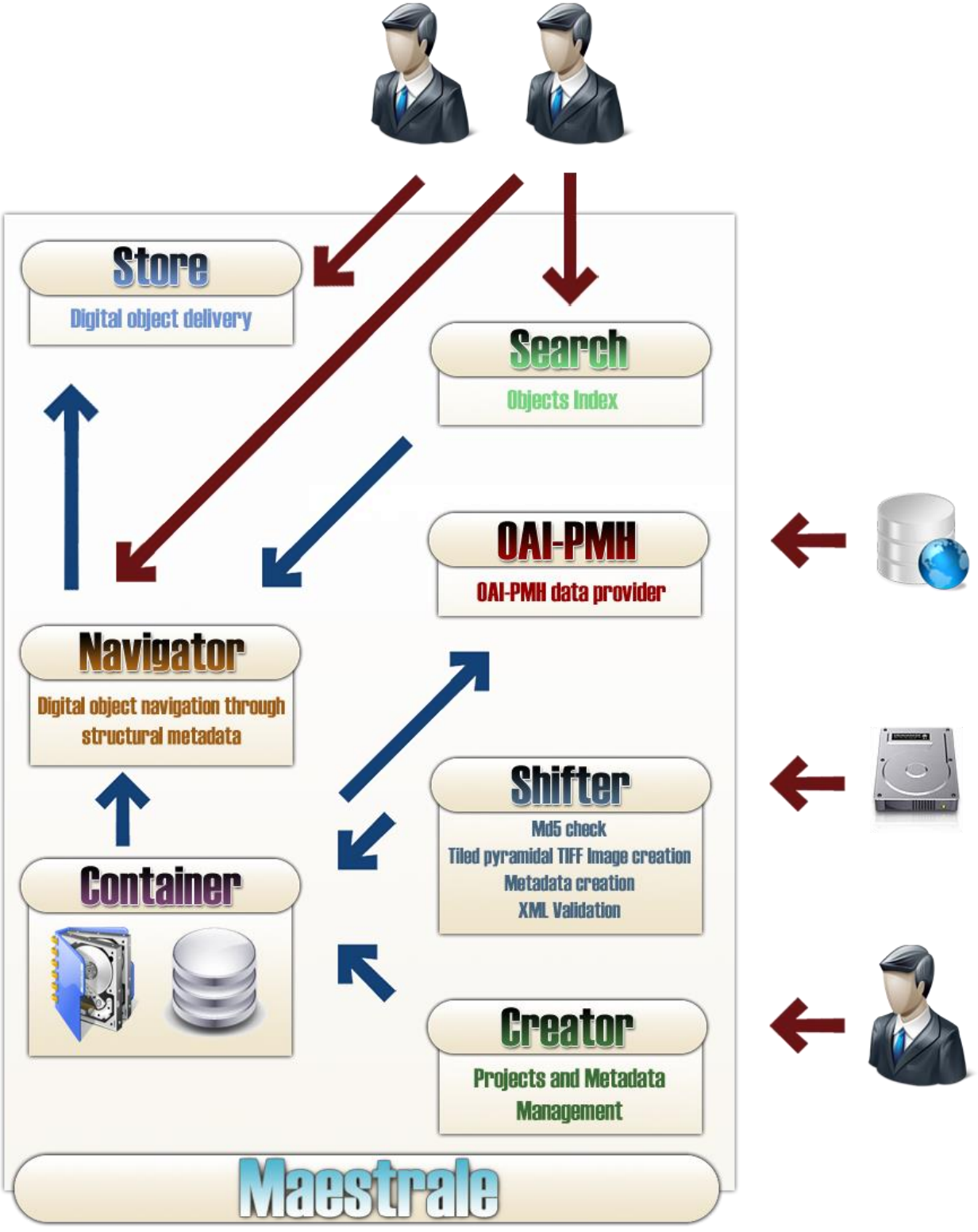
Codex[ml] è un sistema modulare e scalabile che può rispondere alle esigenze di:

- archiviazione orientata alla conservazione a lungo termine delle risorse digitali
- gestione, creazione, import ed export di metadati in formato XML
- fruizione via web degli oggetti digitali organizzati secondo le strutture e le associazioni previste dai metadati associati
- esposizione dei metadati XML secondo lo standard dell'Open Archives Initiative OAI-PMH,
- distribuzione a pagamento di risorse digitali.

Codex[ml] è una piattaforma aperta, strutturata per permettere l'utilizzo di diversi standard di metadati, il passaggio dall'uno all'altro sulla base delle mappature che via via si rendessero disponibili o anche l'adozione in futuro di nuovi schemi di metadati oggi non ancora esistenti.
La piattaforma, pur essendo dotata di un proprio motore di ricerca, è in grado all'occorrenza di interagire con tutti i sistemi di gestione e consultazione di metadati descrittivi per i Beni Culturali e con qualunque sistema di CMS (Content Management System).

Il flusso dei dati

Durante lo sviluppo di Codex[ml] si è mirato a creare un prodotto modulare, scalabile, dinamico ed elastico rispetto agli standard di metadati, in grado di gestire e preservare nel tempo notevoli quantità di dati, e dunque di fornire una risposta concreta, da un lato alle problematiche legate alla digital preservation, dall'altro alla richiesta di interoperabilità con altre piattaforme attraverso il protocollo OAI-PMH.
Da un punto di vista tecnico Codex[ml] è un insieme di applicativi collegati da un framework (Maestrale) che racchiude tutte le funzionalità di base condivise.
Le risorse fornite dai diversi enti vengono importate all'interno del data storage mediante il modulo Shifter che contestualmente, se necessario, provvede a creare le immagini in formato piramidale e i metadati amministrativo gestionali (anche a partire da semplici file guida forniti dagli enti proprietari delle risorse). Una volta caricati, i metadati relativi alle risorse digitali potranno poi essere modificati mediante il modulo Creator.
Le risorse digitali vengono poi navigate e fruite mediante il Navigator. Infine, attraverso il modulo OAI-PMH i metadati possono essere esposti per essere recuperati dai data harvester. Nel caso in cui l'ente proprietario delle risorse abbia deciso di utilizzare anche i moduli Search e Store, l'utente potrà poi effettuare ricerche all'interno dei metadati descrittivi relativi ai documenti digitalizzati, acquistare le risorse digitali o richiedere la scansione di risorse non ancora digitalizzate.



Il framework applicativo

Tutto il software che compone Codex si basa su Maestrale; un framework applicativo sviluppato in pHp a partire dai moduli dello Zend Framework, di PEAR e di altro codice open source come 'PHP-CAS'. Lo scopo di Maestrale è quello di fornire uno 'skeleton' di base da cui partire a sviluppare applicativi web senza preoccuparsi di riconfigurare o riadattare codice per gli 'accessori' indispensabili a quasi tutti i software di

questo genere (sistema di autenticazione, filtri, validatori, cifratura, Ajax, acl, configurazione, cache etc.). Maestrale è attualmente compatibile con i DBMS IBM DB2 e PostgreSQL e si basa sul paradigma MVC (Model View Controller).

Nonostante il compito principale sia quello di fornire una base per lo sviluppo di web application, Maestrale può essere utilizzato anche per realizzare applicazioni stand-alone ("daemon" compresi). Esiste sempre una sola copia di Maestrale per ogni macchina fisica/virtuale su cui si basano uno o più applicativi.
Le interazioni con il DB avvengono sempre tramite Maestrale.

L'architettura

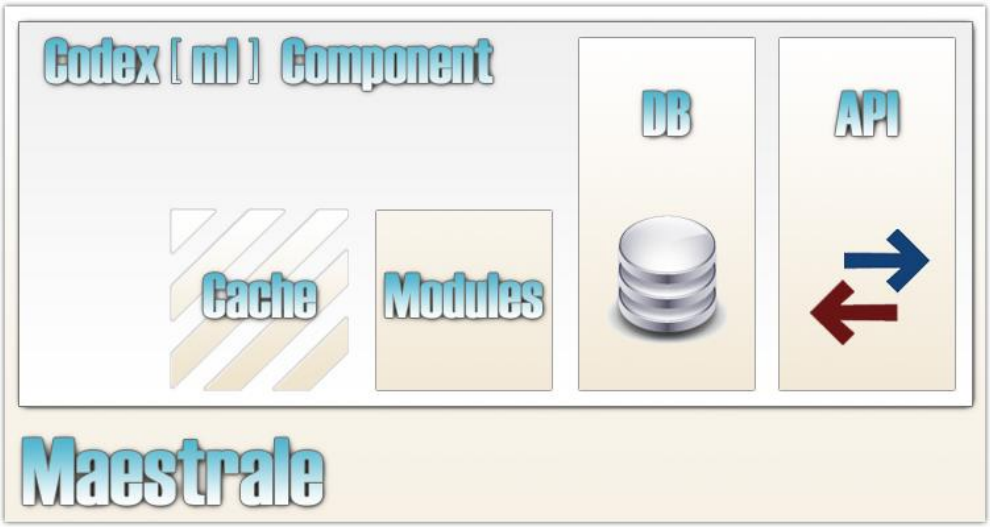
Codex[ml] è realizzato completamente in php per cui necessita principalmente dell'interprete pHp 5.4 o più recente. Codex[ml] è composto da due tipologie di script:

- script per il web;
- script stand-alone, essenzialmente i 'daemon' che si occupano della manutenzione e dell'operatività batch.

I primi operano attraverso Apache httpd mentre i secondi vengono lanciati da linea di comando direttamente da php. Tutti i componenti web di Codex[ml] utilizzano un DBMS. Gli script stand-alone usano, invece, connessioni non persistenti.

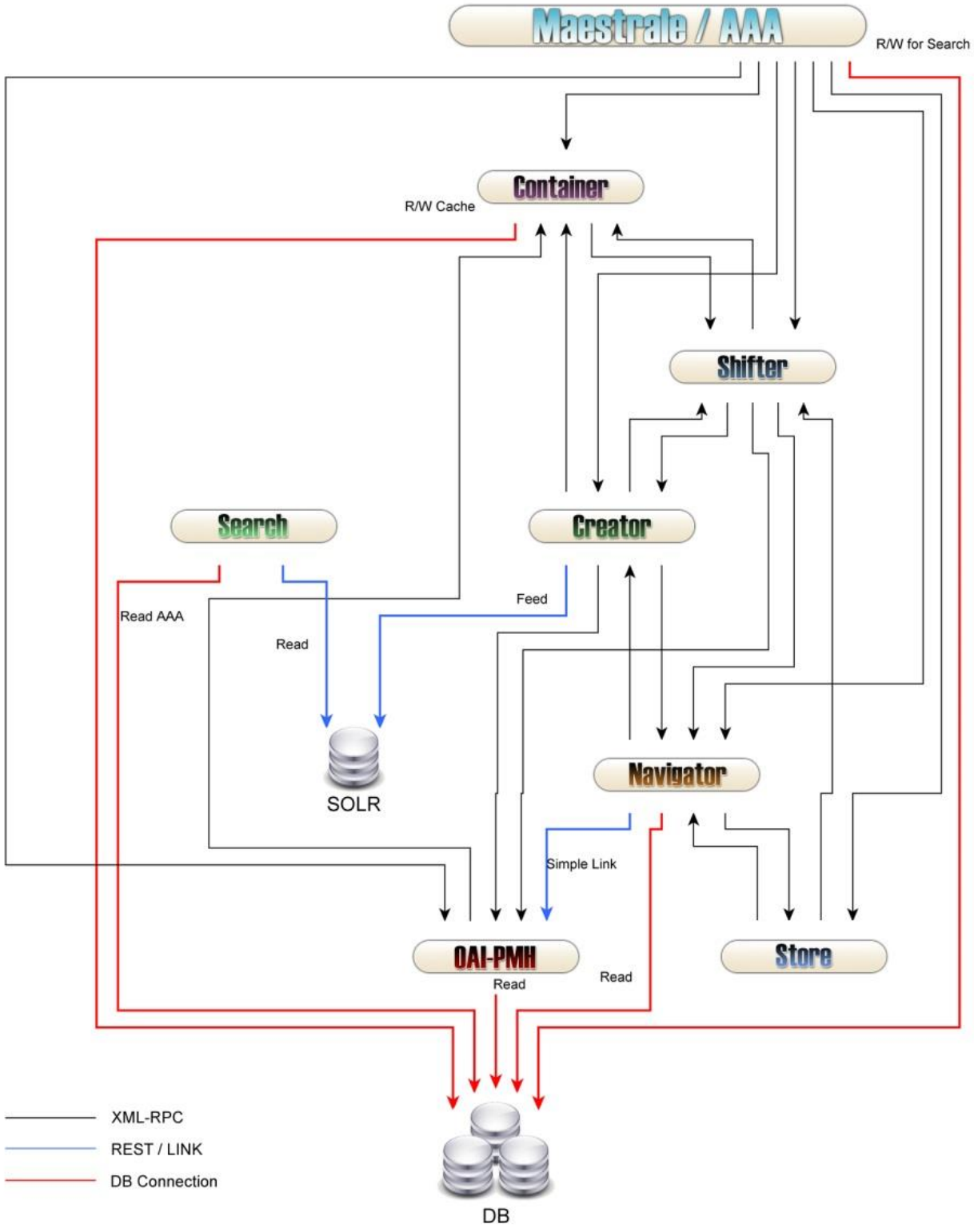
Ogni componente di Codex[ml] eredita tutto il codice di Maestrale e possiede tre layer interni:
- layer di moduli per la gestione dei differenti tipi di metadati (XML);
- layer di comunicazione con il DB;
- layer di API (Webservice XML-RPC).

Anche i sopracitati layer sono ereditati da Maestrale, ma configurati ed estesi per le esigenze del componente. Tramite il layer di interazione con i moduli per la gestione dei metadati è possibile aggiungere o rimuovere funzionalità legate ai metadati che si desidera gestire.
In alcuni casi è presente anche un layer di comunicazione con una cache comune contenente parte dei metadati inseriti con pacchetti di ingest. Questa cache è completamente trasparente per l'utente e viene aggiornata in tempo reale dal sistema.



Le interazioni tra moduli

Nello schema seguente è possibile osservare le interazioni tra i moduli di Codex[ml].



La maggioranza delle interazioni è mediata dai webservice XML-RPC.

Nello schema sopra riportato le frecce nere rappresentano le chiamate tra i diversi moduli. Ad esempio, una freccia che parte da Codex[ml] XML Navigator in direzione di Codex[ml] XML Creator indica una chiamata dal primo (client XML-RPC) al secondo (server XML-RPC).

La presenza di Maestrale al fianco dei vari moduli di Codex[ml] è legata alla presenza di un livello applicativo, di Maestrale stesso, utilizzabile dagli utenti per operazioni di amministrazione: ad esempio tramite l'interfaccia di gestione è possibile registrare nuovi utenti per qualunque modulo.
Dal momento che la creazione di un utente comporta differenti azioni in base al componente a cui è assegnato l'utente stesso, anche Maestrale utilizza webservices per mettere in grado il componente interessato di eseguire il codice necessario al completamento della registrazione.

Codex[ml] Search è il motore di ricerca di Codex[ml] e si basa interamente su VuFind 2. Per evitare di modificare codice all'interno di questo open source, l'integrazione è basata unicamente su SOLR e la creazione degli utenti front-end è ereditata da Maestrale direttamente all'interno del DB.
Da qui la freccia rossa con la notazione 'Read AAA' che parte da Codex[ml] Search e quella con notazione 'R/W for Search' che parte da Maestrale.

La gestione della cache condivisa conservata su DB, infine, è demandata unicamente a Codex[ml] Container che provvede al suo aggiornamento ogniqualvolta un metadato viene modificato. Gli altri moduli la utilizzano in sola lettura.

Le funzionalità principali

Amministrazione del Sistema

Le attività di Amministrazione possono essere gestite attraverso l'interfaccia del modulo "Maestrale". Tramite tale interfaccia è possibile:

- gestire l'elenco degli enti proprietari delle risorse,
- gestire gli "account" per gli utenti abilitati ad inserire le risorse in Codex[ml] e a lavorare su di esse,
- gestire i gruppi in cui sono suddivisi gli utenti,
- gestire i profili di autorizzazione degli utenti: ad ogni utente deve essere assegnato un profilo per ognuna delle applicazioni di CodeX[ml] che questi deve poter utilizzare. I profili sono:
    - Super User (può agire su tutte le risorse gestite dal componente in questione),
    - Power User (può agire solo sulle proprie risorse e su quelle degli users appartenenti al medesimo gruppo),
    - User (può compiere solo alcune azioni e solo sulle proprie risorse),
- attivare/disattivare i moduli per la gestione dei diversi standard di metadati,
- attivare/disattivare i diversi applicativi che costituiscono il sistema,

Caricamento delle risorse

I Submission Information Package (SIP) gestiti da CodeX[ml]

Nell'ambito di CodeX[ml] un pacchetto informativo in genere corrisponde a quella che il modello dei dati PREMIS definisce "entità intellettuale". Il pacchetto può essere composto da:

- metadati amministrativo-gestionali (che possono essere strutturati secondo lo standard MAG o METS, o essere costituiti da semplici file guida) e dati (immagini, file audio, file video, ecc.),

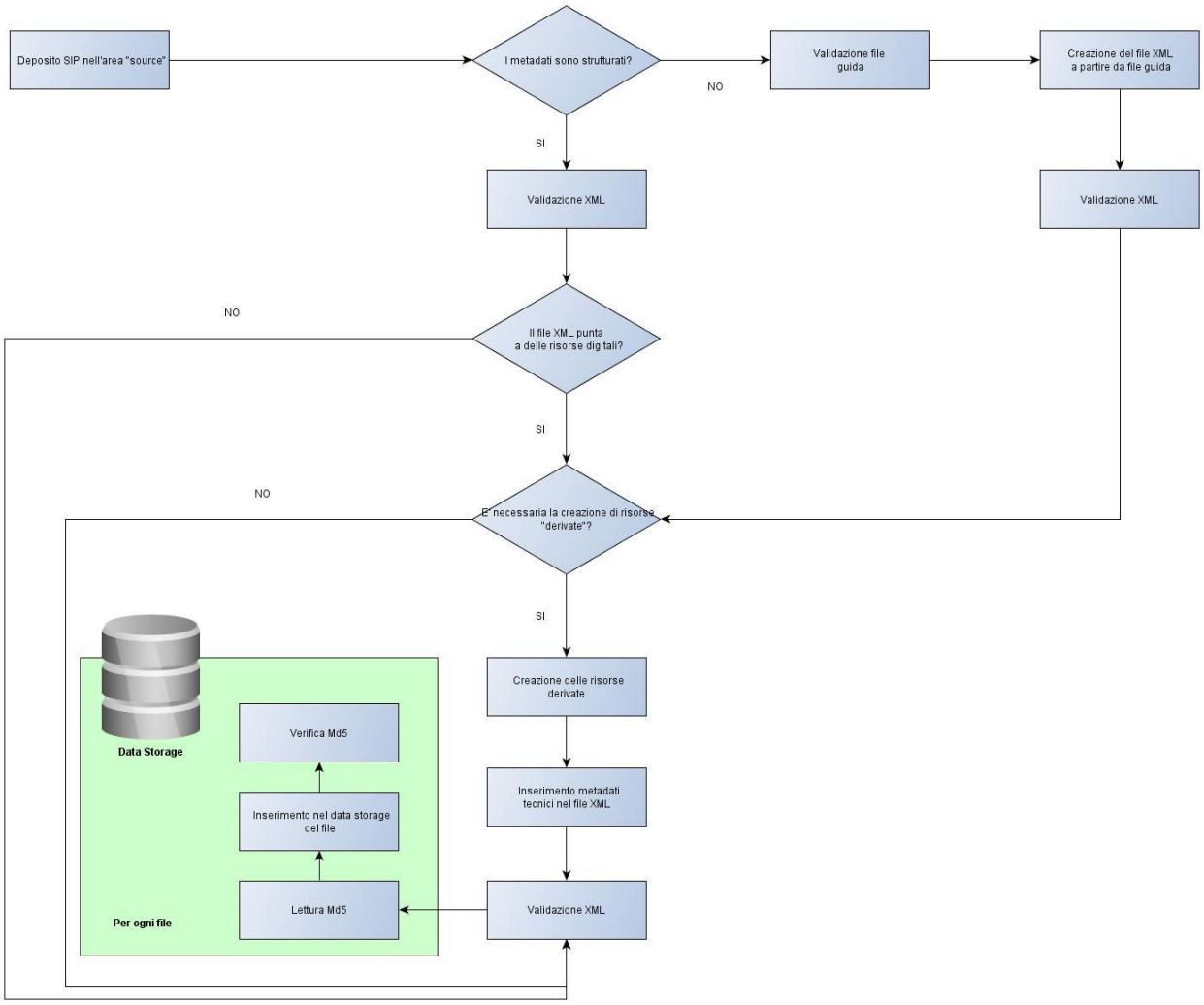
oppure da:

- metadati descrittivi (strutturabili, al momento, secondo gli standard MODS, SBNMarc o ICCD)

CodeX[ml], comunque, è predisposto anche ad accettare l'inserimento di dati privi di metadati, nel caso l'istituzione proprietaria del patrimonio culturale intenda procedere in questo modo, utilizzando la piattaforma semplicemente come *data storage*, oppure voglia inserire i metadati in un secondo momento.

Scendendo maggiormente nel dettaglio, per i contenuti provvisti di metadati, il SIP può essere di tre tipi:

1. composto da un file XML MAG o METS valido (e rispondente alle specifiche per l'inserimento in CodeX[ml]) e da uno o più file relativi a immagini, audio video, ecc.
2. composto da un file XML valido relativo ai metadati descrittivi, strutturabile attualmente secondo gli standard MODS, SBNMarc e ICCD
3. composto da tre file guida in formato XML[1] (che rappresentano rispettivamente i campi obbligatori e richiesti da CodeX[ml] dell'area GEN dello standard MAG, quelli dell'area BIB, e quei metadati dall'area IMG che non sono ricavabili automaticamente dal sistema) e da uno o più file immagini, audio, video, ecc.



[1] Per facilitare la creazione dei file guida, è disponibile un software stand-alone che genera un'interfaccia per l'inserimento dei metadati richiesti.

Il processo di caricamento

L'avvio del processo di ingestion avviene collocando il SIP o i dati (nel caso non siano presenti metadati) in un'apposita area del *data storage*, e quindi avviando il processo tramite l'interfaccia del modulo preposto alla gestione del processo di ingestion (Shifter)[2].

Shifter gestisce diverse tipologie di processi che possono prevedere: l'inserimento di risorse nel data storage, la loro cancellazione, la creazione di metadati in formato XML o di immagini in formato *Tiled Pyramidal TIFF,* la creazione di un collegamento tra i metadati descrittivi e quelli amministrativo-gestionali relativi alla medesima entità intellettuale.

Di seguito vengono descritte nel dettaglio le diverse tipologie di caricamento:

- inserimento nel *data storage* di immagini prive di metadati e loro eventuale piramidalizzazione
- inserimento nel data storage di metadati in formato XML, delle eventuali immagini ad essi associati, eventuale piramidalizzazione delle immagini stesse ed eventuale collegamento con altre tipologie di metadati relative alle medesime entità intellettuali
- creazione dei file in formato XML MAG a partire dai file guida prodotti dagli utenti, inserimento nel *data storage* del file XML MAG creato e delle relative immagini, piramidalizzazione delle stesse ed eventuale collegamento con i metadati descrittivi relativi alle medesime entità intellettuali.
- inserimento dei metadati in formato XML, successivamente all'inserimento delle immagini (caricate in precedenza senza metadati associati) e collegamento tra dati e metadati.

Il processo viene configurato sulla base di un file XML denominato "*driver*" dove vengono indicati tutti i parametri di cui tenere conto. All'interno del driver viene specificato, tra l'altro, se il SIP sarà caratterizzato dalla presenza di metadati strutturati o meno e se è richiesto di provvedere alla piramidalizzazione delle immagini per ottenere dei file in formato *Tiled Pyramidal TIFF*.

Nel caso in cui siano presenti metadati strutturati, il primo passo sarà la validazione del file XML rispetto allo schema o agli schemi (XSD) di riferimento. Successivamente, nel caso sia necessario, il modulo provvederà alla creazione delle immagini piramidali, all'inserimento dei metadati tecnici delle stesse (estratti mediante il software *ImageMagick*) nel file XML e alla validazione dello stesso così modificato. A questo punto, per ogni file contenuto nel SIP, verrà salvato il risultato del *checksum* md5 ed effettuato l'inserimento all'interno del data storage.

Nel caso in cui, invece, i metadati forniti con il SIP non siano strutturati, il primo passo del processo è la validazione dei file guida, cui segue la generazione del file in formato XML MAG e la validazione dello stesso. Una volta ottenuto un file XML valido, il processo prosegue allo stesso modo di quello precedentemente descritto.

In caso di errore in qualunque fase del processo (assenza dei file previsti, parsing dell'XML o delle immagini, creazione delle immagini piramidali), il *workflow* viene interrotto. Una volta risolto il problema, in genere, è possibile effettuare il *resume* del processo, in modo che esso riprenda dal punto in cui è stato interrotto. Nel

[2] La componente XML dei SIP può essere creata e inserita nel *data storage* anche attraverso l'interfaccia del modulo Creator di CodeX[ml].

caso in cui, invece, questo non sia possibile, e il processo vada rilanciato da zero, il sistema è in grado, sulla base del *checksum* md5, di non inserire nuovamente le immagini già caricate.

L'Archival Information Package (AIP)

Alla fine del processo di *ingestion*, quindi, se si eccettuano i casi, già ricordati in precedenza, di risorse prive di metadati, l'AIP (*Archival Information Package*), al momento, è necessariamente costituito da:

- un file XML MAG, che descrive l'entità intellettuale e tutte le immagini, identificato da un *persistent identifier*
- file immagini (TIFF e Tiled Pyramidal TIFF) identificati ognuno da un *persistent identifier*

Opzionale poi può essere invece la presenza nell'AIP di un file XML contenente i metadati descrittivi in formato MODS, SBNMarc o ICCD[3]

Le relazione tra l'XML e l'immagine è esplicitata all'interno dell'XML stesso, mentre la relazione tra i metadati descrittivi e quelli amministrativi è esplicitata sia all'interno dell'XML che nel database di CodeX[ml].

Creazione/modifica dei metadati

Le attività di creazione/modifica dei metadati possono essere effettuate attraverso il modulo Creator. Tale modulo è in grado di generare automaticamente una GUI (*Graphical User Interface*), sulla base di un qualunque XML Schema, attraverso la quale intervenire sul file XML. Il modulo permette di gestire il lavoro per Progetti, e di parametrizzare gli schemi sulla base delle esigenze del singolo ente. È possibile gestire la compilazione degli XML partendo da zero, oppure importare dei file preesistenti all'interno di Codex[ml] Creator, completarli o correggerli, e quindi pubblicarli sui moduli Navigator, OAI-PMH e Search.

Navigazione dei metadati

Una volta che le risorse e i relativi metadati sono stati importati, è possibile visualizzarle mediante l'interfaccia di navigazione (Navigator).
La struttura del documento viene presentata mediante un "albero" che può essere esplorato in tutte le sue parti sulla base delle informazioni recuperate in tempo reale direttamente dagli XML contenuti nel database.
Utilizzando l'interfaccia di fruizione, è possibile navigare le risorse digitali analizzandone i dettagli mediante potenti funzioni di zoom, in grado di permettere la visualizzazione di particolari spesso difficili da cogliere ad occhio nudo. È possibile altresì visualizzare i principali metadati descrittivi relativi al documento e i principali metadati tecnici relativi alle immagini che si stanno visualizzando.

Discovery Tool

Come già accennato, la piattaforma CodeX[ml] fornisce anche un Discovery Tool (CodeX[ml] Search) basato sul software open source VuFind, che permette di navigare ed effettuare ricerche *google-like* sui metadati descrittivi. La base dati di VuFind è costituita dagli indici di Lucene, interrogabili via web, per mezzo della piattaforma di ricerca Solr, con la possibilità di utilizzare anche funzionalità di faceted browsing.

Distribuzione a pagamento delle risorse digitali e attivazione di servizi di digitalizzazione "on demand"

Utilizzando il modulo CodeX[ml] Store è possibile distribuire a pagamento le risorse digitali ed attivare servizi di digitalizzazione "on demand". Tale modulo permette di gestire un flusso di lavoro, articolato, appunto in

[3] I dati possono comunque essere prospettati in formato METS via OAI-PMH

due macrofasi (digitalizzazione e vendita online delle risorse digitalizzate), che va dal momento in cui viene effettuata la richiesta di digitalizzazione di documento fisico posseduto da un'istituzione fino al momento dell'acquisto della risorsa digitale e della consegna di tale risorsa all'acquirente.

Il software è in grado di:

- gestire e monitorare i diversi passaggi relativi alla compilazione della richiesta di digitalizzazione e all'attivazione di tutte le procedure necessarie alla digitalizzazione o all'acquisto di risorse già digitalizzate,
- essere attivato a partire dagli OPAC delle diverse istituzioni,
- permettere agli enti di assegnare le proprie risorse a digitalizzatori diversi, con la possibilità di parametrizzare i singoli tariffari e la qualità delle immagini da distribuire
- preparare il “pacchetto” con le immagini e porlo nella directory da cui l'utente potrà scaricarlo.

Mediante il medesimo modulo è, inoltre, possibile per quegli enti che non permettono la visualizzazione pubblica delle risorse digitali, gestire la vendita di abbonamenti di diversa durata per accedere alla propria Digital Library.

Gestione del data-storage

Per la gestione del data storage è presente un'apposita interfaccia attraverso la quale è possibile visualizzare, scaricare, spostare, o aggiornare una risorsa.
I dati conservati nel data storage sono sottoposti ad una procedura che verifica periodicamente l'integrità dei dati conservati. È stato dimostrato, infatti, che su 1TB di dati non acceduti per un anno può registrarsi un “rate bit flop” 25 volte superiore alla norma. Nel caso in cui il processo identifichi la presenza di un file corrotto, viene inviata la segnalazione all'amministratore del sistema in modo che sia possibile sostituire la risorsa corrotta con quella integra conservata mediante backup.

Interazione con OPAC e altri motori di ricerca

Le immagini conservate in CodeX[ml] possono essere richiamate e visualizzate da qualsiasi OPAC o motore di ricerca mediante dei semplici link. In particolare è possibile:

- accedere direttamente dagli OPAC o da qualunque motore di ricerca o pagina web all'interfaccia di navigazione di CodeX[ml],
- integrare il client IIPImage sviluppato per CodeX[ml] all'interno di un'interfaccia di fruizione diversa da quella di CodeX[ml],
- integrare all'interno di un'interfaccia client IIPImage di terze parti (seppur adattati) che richiamino le immagini piramidali conservate all'interno di CodeX[ml].

Utenti

Ad oggi la piattaforma CodeX[ml] installata presso CINECA è utilizzata da 22 enti (Biblioteca Interdipartimentale “F. Petrarca” di Pavia, Conservatorio “G. Verdi” di Milano, Fondazione “G. Cini” di Venezia, Università IUAV di Venezia, Università “La Sapienza” di Roma, Veneranda Biblioteca Ambrosiana di Milano, Veneranda Fabbrica del Duomo di Milano, Banca Carige, Banca Intesa Sanpaolo, Banca Monte dei Paschi di Siena, Banca Popolare Commercio e Industria, Banca Popolare di Bergamo, Banca Popolare di Milano, Banca Popolare di Vicenza, Banca di Sassari, Banco di Brescia, Credito Emiliano, Cassa di Risparmio di Ravenna, Credito Valtellinese, UBI Banca, UBI Carime, Unicredit) ed ospita oltre 2.600.000 risorse digitali, per un’occupazione di memoria di oltre 36 TiB.
Installazioni autonome sono invece presenti presso la Biblioteca Nazionale di Napoli e il Comune di Padova.