DigLib: un’infrastruttura tecnologica per i progetti di digitalizzazione dell’Università di Bologna

Autori: Simone Sacchi – Fabrizio Morroia

Alma-DL e DigLib: il contesto generale.

La necessità di agevolare l'accesso a materiale bibliografico antico, ha portato alla nascita di molti progetti di digitalizzazione, a livello nazionale o legati a singole Università e centri di ricerca.
Le biblioteche italiane sono notoriamente ricche di materiale antico e di pregevole valore, la cui consultazione è resa problematica dalle difficoltà oggettive che il trattamento di questo materiale comporta: spesso infatti, si tratta di documenti di facile deperibilità, la cui salvaguardia impone limiti alle biblioteche rigide norme intesea limitarne l’accesso.
Nondimeno l’importanza di questi documenti risulta tale che una loro valorizzazione non potrebbe che arricchire il patrimonio storico-culturale comune e facilitare notevolmente il lavoro di ricercatori e storici i cui studi spesso necessitano di un accesso diretto alle fonti.
La digitalizzazione sistematica di questo tipo di documenti, risponde alle esigenze di una comunità scientifica sempre più attenta alle fonti primarie e che considera internet uno strumento quotidiano per la consultazione di materiale bibliografico a testo pieno.
All'interno di questo contesto l'Alma Mater Studiorum Università di Bologna, tramite il CIB Centro Inter-Bibliotecario aderendo alle direttive del Progetto Biblioteca Digitale Italiana, ha promosso il progetto Alma-DL Biblioteca Digitale dell'Alma Mater Studiorum[1].
Il progetto nasce con l’intento di fornire nuovi strumenti e servizi di supporto alla ricerca e alla didattica di Ateneo e si articola in alcuni sottoprogetti tra i quali DigLib il cui scopo è lo sviluppo di un’infrastruttura tecnologica di supporto all’archiviazione e alla fruizione di oggetti digitali.
Il sistema è complesso ed investe organi accademici, biblioteche e centri di servizio, tra cui: il Sistema Bibliotecario di Ateneo, referente per le biblioteche dell’Università, il Centro Inter-Bibliotecario[2], responsabile delle competenze tecnologiche e il Comitato Scientifico a sovrintendere la scelta del materiale bibliografico.
A livello organizzativo risulta evidente la necessità di rendere efficiente la gestione dei flussi di lavoro quali la scelta del materiale, la gestione dei rapporti con i fornitori di servizi in out-sourcing, dei processi di digitalizzazione, archiviazione e organizzazione dei materiali digitali nonché l’implementazione e aggiornamento del sistema di accesso tramite web.
In questo contesto verrà approfondita la componente più spiccatamente tecnologica e di approccio metodologico alla creazione, degli archivi digitali e del sistema di accesso.

La riproduzione digitale e la conservazione

La digitalizzazione delle opere, viene affidata a service esterni specializzati, ai quali sono fornite le specifiche tecniche per la realizzazione di immagini ad alta risoluzione dalla scansione di materiale cartaceo, utili sia per la conservazione che per un eventuale utilizzo editoriale.
Un apposito capitolato è stato preparato per garantire uniformità agli standard qualitativi ed indicare tutte le esigenze di progetto e i vincoli che il service dovrà rispettare per il corretto svolgimento del proprio lavoro e per la salvaguardia degli originali.
Gli standard di qualità e sicurezza considerati, impongono vincoli sul tipo di apparati da utilizzare, sulle tecniche di ripresa, sul trattamento delle immagini e non ultimo sulle modalità di trasporto, conservazione e trattamento degli originali cartacei.

Tra le specifiche più rilevanti riguardanti la sicurezza e la salvaguardia delle opere citiamo l'utilizzo di scanner a torre con lampada a luce fredda per ridurre al minimo le sollecitazioni fisiche dannose alle rilegature e il surriscaldamento dell'ambiente circostante l'opera. Per quanto concerne il controllo di qualità delle immagini vogliamo ricordare alcune soluzioni utilizzate: ogni immagine è ottenuta da una ripresa digitale unica anche per originali di grandi dimensioni, l'uso di processi software di ricostruzione dell'immagine da riprese parziali è escluso; ogni immagine deve contenere il documento originale per intero e non soltanto la porzione scritta, uno sfondo monocromatico (bianco o nero) di larghezza non superiore ai 5 mm verrà usato per delimitare la parte dell'immagine occupata dal documento; inoltre un originale delle scansioni deve essere conservato presso il service per il tempo necessario alla verifica e controllo delle immagini (sequenza corretta, lacune, rispondenza cromatica).

La seguente tabella[3] riassume gli standard utilizzati per ottenere immagini ad alta risoluzione ai fini della conservazione e copie a bassa definizione utilizzate per l'accesso:

Tipologia: materiale a stampa (rilegato o a fogli sciolti)				
Dimensione	Risoluzione	Scala di grigio	Colore	Formato
Fino ad A5	600 dpi	8 bit (256 toni di grigio)	24 bit (16.7 milioni di colori)	TIFF non compresso
Da A5 ad A4	400 dpi			
Da A4 ad A3	300 dpi			
Maggiore di A3	200 dpi			

Come si evince dalla tabella abbiamo stabilito risoluzioni diverse per originali di dimensione diversa: questo è solo uno dei possibili criteri di scelta (un'altra possibilità è quella di definire una dimensione standard di output, ad esempio A4 e una risoluzione standard per questa dimensione ad esempio 300 dpi e riportare a questi valori tutte le scansioni a prescindere dalla dimensione degli originali) ed è il più efficace per visualizzare o stampare le immagini a grandezza naturale senza l'applicazione di elaborazioni software.
Le risoluzioni scelte consentono una riproduzione fedele anche a livello tipografico considerando come dimensione media di visualizzazione e stampa quella del formato A4 (21x29.7 cm).
La profondità dello spazio colore, 8 bit per immagini in bianco e nero e 24 bit (True color) per immagini colorate, consentono di rappresentare al meglio qualsiasi sfumatura percepibile dall'occhio umano. Infine il formato TIFF non compresso rappresenta lo standard internazionale per lo stoccaggio di immagini ad altissima definizione.
I master così ottenuti vengono conservati in tre o quattro copie: due archiviate su supporto DVD-R (una conservata presso il Centro Inter-Bibliotecario e una presso la biblioteca che fornisce gli originali), una su supporto DAT come copia di backup, e infine una nel filesystem del server per la consultazione.

Dalla conservazione all'accesso: i formati di compressione

Il processo di riproduzione digitale ad alta definizione fornisce immagini il cui peso in termini di byte non ne consente la diffusione in rete (va tenuto presente che un'immagine in formato A4 a 300 dpi e 24 bit di colore in modalità CMYK non compressa ha una dimensione di circa 33 MByte), è

necessario quindi creare copie di dimensioni ridotte che possano essere efficacemente trasferite in rete, limitando i tempi di accesso.
Nel nostro progetto abbiamo optato per due diverse modalità di presentazione delle immagini via web applicando due diversi algoritmi di compressione: JPEG[4] e DjVu[5]; la motivazione che ci ha spinto verso questi due formati sono diverse e rispondono a diverse esigenze.
Il formato JPEG è considerato lo standard per la diffusione in rete di immagini a colori, grazie al buon rapporto tra qualità dell'immagine e fattore di compressione, è supportato da tutti i più comuni browser web e la sua visualizzazione non necessità di strumenti software aggiuntivi.
Il formato DjVu è un formato proprietario le cui specifiche sono state pubblicate sotto tutela della licenza GPL, ed ha dimostrato di essere estremamente efficace nella compressione di documenti testuali e immagini tratte da materiale antico. Le sue peculiarità sono insite nella struttura a livelli in cui separa le immagini: un livello per lo sfondo e uno per il testo scritto, ad ognuno di essi vengono applicate due diverse tecniche di compressione ciascuna maggiormente efficace sulla parte che va ad elaborare.
Lo sfondo o livello inferiore(che per quanto riguarda materiale a stampa coincide con la pagina di scrittura senza il testo) viene compresso tramite un algoritmo wavelet, adatto alla compressione di immagini con contrasti di colore ridotti che permette di ottenere fattori di compressione elevati e quindi ridurre enormemente le dimensioni dei file immagine. Il testo o livello superiore, viene prima ridotto ad uno spazio colore bianco-nero, eliminando tutte le informazioni sul colore, poi reso trasparente nelle parti bianche. La ricomposizione dei due livelli restituisce l'immagine originale (pagina più testo) con occupazioni di memoria paragonabili a quelle dei file jpg, ma di dimensione visualizzabile 4 volte superiore ed una fedeltà di riproduzione delle forme e dei colori di qualità simile se non superiore. Per poter visualizzare immagini DjVu nel proprio browser è necessario installare un plugin gratuito molto simile a quello utilizzato per i file PDF che offre molti tools molto utili all'utente tra i quali un potente zoom.

L'utilizzo di questi formati ci permette di offrire un doppio servizio: coloro che non vogliono o non possono installare il plugin DjVu sul proprio PC riescono comunque ad accedere alle immagini JPEG, mentre chi ritene opportuno installarlo potrà usufruire delle funzionalità aggiuntive.
Di seguito le caratteristiche delle immagini nel formato di pubblicazione in rete:

Formato	Dimensioni a video	Risoluzione	Tipo di compressione
JPEG	Larghezza: 600 px Altezza: in proporzione	100 dpi	?????????????
DjVu	Larghezza: dimensione della finestra del browser Altezza: in proporzione*	300 dpi	Multilayer wavelet

*Il plugin DjVu si ridimensiona a seconda della finestra di visualizzazione, all'interno della quale può poi essere applicato lo strumento zoom.

I metadati: una scelta difficile

La scelta del set di metadati utili alla gestione degli oggetti digitali è stata la più complessa e ci ha impegnati nello studio e nel confronto tra diversi progetti che hanno esposto lo schema adottato, cercando di sfruttare al meglio le esperienze che ci hanno preceduto applicandole alle nostre esigenze di progetto.
Il nostro intento era quello di astrarre il più possibile la scelta dei metadati dal contesto che ci ha legati ad utilizzare un periodico come opera per lo sviluppo del prototipo: si vuole infatti sviluppare una piattaforma in grado di gestire qualsiasi tipologia documentaria, indipendentemente dal formato bibliografico (libri, periodici, documenti sciolti, fotoriproduzioni, materiali musicali, etc.), dalla provenienza (biblioteche, archivi, centri di documentazione) e dal contenuto, cercando al contempo

di creare una struttura in grado di permettere l'accesso alle opere con modalità di consultazione proprie per ciascuna specificità.
Abbiamo quindi separato la gestione degli oggetti dalle specifiche di accesso, creando un set di metadati strutturali e amministrativo-gestionali in grado di rappresentare e gestire efficacemente qualsiasi tipo di unità documentaria strutturata gerarchicamente a più livelli.
Gli oggetti digitali vongono quindi associati ad un set di metadati descrittivi specifici per ogni tipologia documentaria (attualmente viene utilizzato un sottoset del Dublin Core[6] per economie di sviluppo).
Dal punto di vista sistemistico i più significativi sono certamente i metadati strutturali, poiché il carattere generale della formalizzazione che introducono, ci permette di considerarla non soggetta a modifiche sostanziali e quindi ideale per essere identificata come scheletro della progettazione del sistema.
L'oggetto digitale completo, che esprime l'unità bibliografica nel suo complesso, è rappresentata da un'entità detta oggetto *primario*, che può essere suddiviso in oggetti detti *intermedi*.Questa struttura permette di gestire unita bibliografiche annidate su più livelli, come ad esempio volumi, capitoli e paragrafi. La suddivisione in oggetti intermedi è opzionale poichè non tutte le risorse lo richiedono.
Gli oggetti detti *terminali*, rappresentano infine le entità base di cui è formata una risorsa digitale (pagine, tavole, spartiti, etc.) e possono essere raggruppati in sottoinsiemi e associati agli oggetti intermedi o essere raccolti in un unico gruppo e legati direttamente all'oggetto primario.
L'ordine sequenziale è limitato agli oggetti che su uno stesso livello gerarchico hanno il padre in comune. In termini prettamente informatici possiamo dire che la struttura appena descritta è un albero n-ario, in cui i nodi fratelli sono in ordine sequenziale tra loro.

I metadati amministrativo-gestionali descrivono e caratterizzano le entità digitali fisiche, creano cioè la "carta d'identità" degli oggetti digitali e ne descrive il formato, la data di creazione, le dimensioni, la localizzazione, gli strumenti di visualizzazione e le condizioni di accesso: questo per conservare una traccia indelebile del processo di creazione, gestione e accesso agli oggetti.
Ogni tipologia di formato ha i propri parametri caratteristici, ad esempio: i file immagine sono tipicamente definiti da un'altezza e una larghezza espresse in pixel, dalla risoluzione di stampa, dalla profondità di colore o di toni di grigio e dal tipo di scanner usato, mentre nel caso di file di testo avremo altri parametri, come il tipo di font dei caratteri, il numero di righe e colonne, il software OCR utilizzato.
Gli oggetti fisici sono legati alla struttura tramite gli oggetti terminali in modo del tutto simile a come questi sono legati agli oggetti di rango superiore, ma la natura del legame è concettualmente diversa; infatti mentre i figli di un dato nodo rappresentano le sue parti costituenti, gli oggetti fisici definiscono le viste delle entità terminali e cioè l'insieme dei file in uno stesso formato con cui è possibile visualizzare le entità. La possibilità di un nodo di avere più figli si traduce nella possibilità di associare ad un oggetto terminale più rappresentazioni visive della stessa entità.
Anche in questo caso le modifiche da apportare al sistema per l'introduzione di una nuova vista non si propaga alla struttura, ma si limita all'aggiunta di un legame tra ogni oggetto terminale interessato e le entità fisiche nel nuovo formato.

Software gestionale

Il software è stato interamente realizzato in Perl e la sua progettazione ed implementazione è orientata agli oggetti; questo tipo di organizzazione aumenta la modularità del codice, migliora l'organizzazione logica dei componenti e dei loro rapporti, inoltre ne facilita l'estensione e il riutilizzo. Il software è stato dotato di un'interfaccia grafica per l'interazione con l'utente-catalogatore che tramite essa inserisce le informazioni indispensabili ai fini dell'archiviazione

(metadati descrittivi). Oltre a questo l'interfaccia permette la scelta della tipologia di archiviazione tra quelle messe a disposizione .
La funzionalità principale del software è la costruzione della *rappresentazione interna* delle risorse, e la sua trasformazione in una delle *rappresentazioni esterne*. Per rappresentazione interna si intende, la formalizzazione definita internamente al sistema per rappresentare la risorsa e cioè l'insieme degli oggetti e delle loro relazioni (oggetti primari, intermedi, terminali e fisici); mentre per rappresentazione esterna si intende la presentazione delle informazioni implicitamente ed esplicitamente contenute nella rappresentazione interna, in una forma che le renda fruibili ad un utente esterno. Nel nostro caso le rappresentazioni esterne sono l'insieme delle tabelle del database che archiviano la risorsa e un set di file XML[7] che presentano i metadati in forma testuale. E'importante definire questa astrazione dei componenti esterni al sistema software per indurlo ad essere estremamente flessibile verso la possibile introduzione di nuovi sistemi di archiviazione o presentazione dei dati raccolti. In altre parole l'aggiunta di un sistema di impaginazione per la stampa dei metadati o il passaggio ad un'altra piattaforma di database, deve tradursi nella scrittura di nuovo codice, integrabile nel sistema senza la necessita di alterare quello esistente.
L'operazione di archiviazione di una nuova risorsa ha come dato iniziale, l'insieme dei file che rappresentano la forma digitale dell'opera, raccolti all'interno di una porzione di file-system, raggruppati in cartelle e sottocartelle secondo la disposizione scelta e numerati all'interno di esse in ordine sequenziale. Il componente software dedicato alla costruzione della struttura, raccoglie informazioni di carattere bibliografico da un form presente sull'interfaccia grafica, quindi esegue una scansione della porzione di file-systeminteressata, ricavando automaticamente le informazioni richieste dai metadati strutturali e gestionali.
Il secondo metodo di costruzione si ottiene dal parsing di file XML, associati a risorse gia archiviate in precedenza e non necessita nessuna operazione di inserimento dati aggiuntiva poichè il file contiene già tutte le informazioni necessarie.
Una volta terminata la costruzione sarà possibile scegliere, quale delle rappresentazioni esterne generare. Nel caso della creazione ex-novo di una risorsa, il popolamento del database si traduce nella compilazione di nuove tabelle o nel caso di file XML, nella creazione di nuovi file. Il secondo metodo di costruzione invece, ha lo scopo facilitare la modifica manuale di dati già presenti nel database; l'idea parte dall'assunzione che sia in generale più semplice e sicuro editare un file di testo piuttosto che modificare direttamente le tabelle del database, per cui la procedura di modifica prevede prima la modifica del file, quindi la costruzione della rappresentazione interna a partire dal file modificato e la sovrascrittura delle tabelle interessate attraverso il software.

Database

Il database scelto è di tipo relazionale, è cioè capace di instaurare relazioni tra i dati presenti su tabelle diverse, e permette interrogazioni i cui risultati dipendono dai valori messi in relazione. Nel nostro caso questa possibilità risulta indispensabile per mantenere tra le tabelle le stesse relazioni esistenti nella struttura della risorsa.
In questa fase prototipale il database utilizzato è MySql, un prodotto freeware e open source, le cui peculiarità sono l'efficienza e la velocità, ottenute a scapito di controlli sulla consistenza delle tabelle e delle relazioni tra esse, normalmente presenti in prodotti più complessi e articolati. L'utilizzo del software per il popolamento del database ci permette di poter confidare sulla sua consistenza anche senza questi controlli, mentre le modifiche manuali dirette sono sconsigliate, e dovrebbero essere attuate solo in casi di assoluta necessità da personale competente e disciplinato.

Il linguaggio d'interrogazione usato è SQL standard, linguaggio con cui gran parte dei database relazionali mantengono per lo meno la compatibilità, per cui il trasferimento dell'archivio su un database diverso, ma basato sullo stesso linguaggio, non comporta modifiche sostanziali al software.

Interfaccia web e consultazione delle risorse.

L'interfaccia di consultazione web è stata realizzata in PHP, un linguaggio ideato per la costruzione di pagine web dinamiche il cui contenuto ed aspetto varia in base all'interazione tra l'utente e la pagina; nel nostro caso a seconda della risorsa che l'utente sceglie di visualizzare, il motore php interroga il database, secondo le richieste inoltrate attraverso la pagina ed estrae le informazioni che gli sono utili per la costruzione della URL del file scelto.
La pagina contiene tutti gli elementi necessari alla presentazione e navigazione di un oggetto primario e dei suoi componenti, fino alla visualizzazione dell'oggetto terminale nel formato prescelto. La varie parti dell'opera sono presentate e selezionabili tramite una stringa testuale che denota il nome associato alla parte(es. cap.1, pag.1, etc). La selezione dà come risposta, l'apertura di sottomenu per la selezione di altre sottoparti se presenti o con la visualizzazione del file se l'oggetto scelto rappresenta un oggetto terminale inoltre se sono presenti più viste della risorsa sarà possibile scegliere quale di queste forme di visualizzazione utilizzare.
Sul menu di navigazione sono inoltre presenti, pulsanti per lo scorrimento delle pagine in sequenza e per selezionare direttamente la pagina iniziale o finale.

Conclusioni e sviluppi futuri

Lo studio di fattibilità e lo sviluppo del prototipo hanno permesso di analizzare dettagliatamente le problematiche che un progetto di digitalizzazione sistematico comporta, quali la conservazione del digitale, l'accesso permanente e l'usabilità del servizio.
L'evoluzione della nostra piattaforma tecnologica, strettamente interconnessa con le raccomandazioni dell'Ateneo e le nuove sperimentazioni italiane e straniere in questo campo, comporterà: l'arricchimento dei metadati descrittivi con l'implementazione dell'intero set di metadati Dublin Core per la descrizione delle unità bibliografiche e la creazione di uno schema descrittivo per gli spogli dei periodici; lo sviluppo di un sistema di gestione del Copyright Digitale (DRM) per garantire i corretti diritti di accesso; il perfezionamento dell'interfaccia con la possibilità di consultare in parallelo le immagini digitali e il testo strutturato corrispondente (codificato secondo lo standard TEI Lite[8]).

---

1 Alma-DL Biblioteca digitale dell'Università di Bologna
http://almadl.cib.unibo.it

2 CIB Centro Inter-Bibliotecario Università di Bologna
http://www.cib.unibo.it

3 Per una rassegna delle Best Practices sui formati e gli standard per le immagini digitali si rimanda ad alcuni esempi:
Guides to Quality in Visual Resource Imaging
http://www.rlg.org/visguides/

Digital capture, format & preservation
http://www.slq.qld.gov.au/pub/digital/sd2_digcapture.htm

Digital Imaging

http://palimpsest.stanford.edu/bytopic/imaging/

Digitizing Images and Text
http://sunsite.berkeley.edu/Imaging/

4 Joint Photographic Experts Group
http://www.jpeg.org/

5 DjVu Zone
http://www.djvuzone.org/

[6] Dublin Core Metadata Element Set
http://dublincore.org/documents/dces/

[7] Extensible Markup Language (XML)
http://www.w3.org/XML/

[8] Text Encoding Iniziative
http://www.tei-c.org/