

DIGITALIZZAZIONE DI PATRIMONI CULTURALI
Lotto 1: Sviluppo e arricchimento della Biblioteca Digitale Lombarda

Nota tecnica sulla realizzazione del servizio di digitalizzazione

Il documento intende fornire alcune informazioni e precisazioni di natura tecnica riguardo alle caratteristiche delle immagini e alle relative modalità di acquisizione, nonché al flusso di lavoro previsto, per il quale sarà messo a disposizione un applicativo software di gestione realizzato da Regione Lombardia tramite la società Lombardia Informatica SpA.

Parametri tecnici delle immagini

Acquisizione delle immagini

L'acquisizione dei documenti sarà effettuata per pagina e produrrà un file per ogni pagina (o per ciascun bene storico-artistico o naturalistico); potrà essere effettuata un'unica scansione che acquisisca le due pagine affiancate, dovranno poi essere prodotti due file, ciascuno dei quali corrispondente alla singola pagina. Nel caso di presenza di un'immagine che si estenda sulle due pagine affiancate, è opportuno creare un solo file corrispondente alle due pagine.

L'acquisizione dovrà comprendere tutte le pagine del documento, comprese quelle bianche. Si ritiene che debbano essere acquisite anche la coperta (1a e 4a di copertina), il dorso e i relativi piatti interni, e i tagli. In merito all'inquadratura, si evidenzia che dovrà comprendere tutti gli elementi informativi connessi all'aspetto materiale del documento (particolare attenzione sarà rivolta ad alcune tipologie documentarie: manoscritti, grafica...).

Ove la carta presenti caratteristiche di grammatura molto leggera o comunque sia molto sottile, per evitare che l'immagine digitale risulti scarsamente comprensibile a causa dell'effetto di trasparenza, sarà opportuno frapporre un foglio di carta bianco tra pagina e pagina, prima della scansione.

L'eventuale opportunità di acquisire il retro di fotografie, disegni, stampe e erbari (anche attraverso la ripresa fotografica digitale del bene nel caso di dipinti) sarà valutata nell'ambito dell'analisi delle collezioni e messa a punto del piano di digitalizzazione.

Risoluzione

Le risoluzioni previste, ampiamente utilizzate a livello internazionale, sono:

FILE MASTER:

file in formato TIFF con risoluzione finale effettiva di 400 dpi reali (la dimensione dell'immagine e quella del documento originale a 400 dpi);

THUMBNAIL (MINIATURA):

file in formato GIF / JPG con dimensioni di 150 x 150 pixel

TESTI:

file in formato JPG con risoluzione effettiva di 72 dpi (small);

file in formato JPG con risoluzione effettiva di 150 dpi (medium);

file in formato JPG con risoluzione effettiva di 300 dpi (large);

STAMPE-DISEGNI-FOTOGRAFIE-ERBARI

file in formato JPG con risoluzione effettiva di 400 dpi.

DIPINTI

La ripresa fotografica in formato digitale dovrà prevedere la produzione di immagine con dimensione maggiore o uguale a 4000 pixel (lato minore).

Colore

Per i volumi a stampa e i periodici si procederà alla scansione a colori o in bianco/nero rispettando le caratteristiche dell'oggetto originale. Per tutte le altre tipologie di documenti si procederà alla scansione a colori.

Il piano di lavoro complessivo delle attività del progetto dovrà contenere un'indicazione esplicita per ciascuna collezione o fondo in merito alla scansione a colori o in b/n.

I requisiti minimi per la profondità di colore delle immagini sono i seguenti:

- 8 bit scala di grigio per le pagine in b/n,
- 24 bit colore RGB mode per le pagine con elementi a colori e per i beni storico-artistici e naturalistici (per questi ultimi si raccomanda di non fornire file in formato colore CMYK)

Si richiede di acquisire un'immagine dell'oggetto che comprenda un color checker o un test chart al fine di dare la corretta percezione dei colori da parte dell'utente finale. Il color checker può venire posizionato:

- per il materiale librario, nella prima pagina bianca a disposizione;
- per gli altri documenti che non dispongono di alcuna pagina bianca, dovranno essere effettuate due scansioni, una con color checker e una senza.

Righello

Per l'esatta comprensione dell'originale da parte dell'utente è fondamentale dare anche una chiara idea delle dimensioni. Il software per la visualizzazione delle immagini che viene messo a disposizione legge nei metadati dell'immagine le dimensioni dell'originale, presentando di volta in volta agli utenti, accanto all'immagine, un righello adeguato per il confronto. E' necessario valorizzare i metadati riguardanti la grandezza dell'immagine.

Nel caso in cui si ritenga che ciò non sia possibile, è opportuno acquisire immagini che contengano, oltre all'originale analogico, anche un righello.

Rumore (noise) e luminosità

Si raccomanda di utilizzare periferiche di acquisizione e software di gestione delle immagini che garantiscano le migliori prestazioni riguardo a questi aspetti.

Post-processing

Una volta che l'immagine sia stata acquisita e salvata, potranno essere effettuati interventi correttivi di post-processing per l'ottenimento di un'immagine di qualità, interventi che potranno riguardare numerosi aspetti: contrasto, luminosità, brillantezza, riduzione dei colori accuratezza geometrica, distorsione, allineamento, rotazione, eliminazione dei bordi in eccesso, eliminazione dello sfondo.

La fase di post-processing consente la creazione, a partire dal file master, di immagini a diverse risoluzioni e con diversi formati di salvataggio.

Formati dei file immagini

Si prevedono i seguenti formati dei file immagine:

- formato TIFF per le immagini destinate alla conservazione nella cartella master;
- formato JPG per le immagini destinate alla diffusione su web (nelle cartelle small, medium e large);
- formato GIF / JPG per le miniature (nella cartella thumbnail).

Struttura degli oggetti e composizione

La struttura degli oggetti all'interno di BDL deve essere composta da cartelle che ricreino l'identificativo univoco all'interno del sistema che si declina in: istituto – progetto – collezione – oggetto.

All'interno della cartella oggetto devono essere presenti le cartelle:

- master – immagini per la conservazione in formato TIFF 400 dpi
- large – immagini per la gestione degli zoom all'interno della pubblicazione in formato JPG 300 dpi
- medium – immagini per la visualizzazione e gestione in formato JPG 150 dpi
- small – immagini per la gestione in formato JPG 72 dpi
- thumbnail – immagini da utilizzare come miniature 150x150 pixel in formato GIF / JPG
- pdf – contiene il file PDF dell'oggetto, denominato con il numero oggetto relativo
- pdfsingle – contiene i file PDF dell'oggetto, uno per ogni pagina dell'oggetto (utilizzati se i file complessivi sono di dimensioni molto grosse).

Eventuali ulteriori indicazioni riferite alla gestione delle immagini di fotografie, stampe, dipinti, disegni e beni naturalistici saranno fornite nell'ambito dell'analisi delle collezioni e messa a punto del piano di digitalizzazione.

I nomi file devono essere di 4 digit (es. 0001. 0002, etc) e devono essere sequenziali.

tmp_digital_files

|-4

|-5 # id_istituto

|-24 # id_progetto

|-67 # id_collezione

|-234 # id_oggetto_digitale

|-master # immagini master in formato TIFF

||-0001.tif

||-0002.tif

||-0003.tif

|

|-large # immagini JPG 300 DPI

||-0001.jpg

||-0002.jpg

||-0003.jpg

|

|-medium # immagini JPG 150 DPI

||-0001.jpg

||-0002.jpg

||-0003.jpg

|

|-small # immagini JPG 75 DPI

||-0001.jpg

||-0002.jpg

||-0003.jpg

|

|-thumbnail # thumbnail 150px X 150px

||-0001.gif

||-0002.gif

||-0003.gif

|

|-pdf

||-234.pdf

|

|-pdfsingle # pdf singola pagina (opzionale)

||-0001.pdf
||-0002.pdf
||-0003.pdf

Schema di workflow del processo di acquisizione con BDL

Nell'ambito dell'applicativo BDL sono previsti i seguenti ruoli:

catalogatore: identifica gli oggetti da digitalizzare e li cataloga nel sistema;

digitalizzatore: crea fisicamente la copia digitale dell'oggetto originale ed effettua i necessari collegamenti con i rispettivi oggetti;

supervisore: controlla e valida le varie fasi del processo.

Per ciascun istituto saranno inseriti nell'applicativo operatori definiti come *catalogatore* e come *digitalizzatore* per conto della Società aggiudicataria. Riguardo al ruolo di *Supervisore*, sarà registrato un responsabile per conto della Società aggiudicataria.

Gestione progetti-collezioni

I dati degli oggetti della BDL sono raggruppati gerarchicamente per istituto, progetto, collezione.

La gestione dei progetti – collezioni è in carico al ruolo di *catalogatore*. La definizione e gli attributi della collezione è particolarmente importante, essendo, la collezione, il raggruppamento degli oggetti presentato nella parte di applicativo esposta al pubblico per consultazione.

Identificazione

Il processo di identificazione oggetti consente al *catalogatore* di inserire in BDL gli oggetti originali gerarchicamente individuati dalla triade progetto/collezione/oggetto originale (ogni oggetto originale è afferente ad una e una sola collezione, a sua volta relazionata ad uno e un solo progetto). Al termine del lavoro il *catalogatore* manda in verifica al *supervisore* l'elenco degli oggetti inseriti. Il flusso di lavoro prosegue successivamente al completamento della verifica da parte del *supervisore*.

All'interno del sistema esistono dei legami tra oggetti originali (nel contesto BDL sono trattati solo legami verticali). I legami sono espressi attraverso il "titolo superiore", ossia un titolo che raggruppa titoli inferiori.

Per le monografie il legame può essere relativo a titoli in più volumi o a reticoli complessi, entrambi gli oggetti devono essere descritti in maniera esaustiva.

Per i periodici il legame è tipicamente relativo a testata-entità fisica (singolo fascicolo o più fascicoli raggruppati in annata, anche a prescindere dalle modalità fisiche di conservazione dei fascicoli in biblioteca).

Gli istituti partner del progetto forniranno alla società aggiudicataria l'elenco (file excel) dei documenti da digitalizzare, corredati dalla relativa collezione di riferimento. Per ciascun documento sarà presente l'identificativo univoco, se disponibile (BID SBN, IDK SIRBeC...); in sua assenza, saranno riportati gli elementi identificativi del documento (autore, titolo, dati relativi alla pubblicazione e all'esemplare fisico, identificativo / link al catalogo di riferimento, ad es. Manus).

Se l'identificazione degli oggetti sarà effettuata mediante identificativo univoco, la procedura consentirà di recuperare i dati di catalogazione in maniera automatica. In assenza di tale identificativo, gli elementi identificativi del documento forniti dagli istituti dovranno essere caricati nella procedura.

Verifica oggetti

La verifica degli oggetti originali è effettuata dal *Supervisore*: nell'ambito di questa fase di lavoro, viene "accettato" l'oggetto inserito da parte del *catalogatore* e viene staccato l'identificativo univoco all'interno del sistema BDL. Questo ID determina anche la struttura di cartelle che devono accogliere le immagini digitalizzate accodando, separati da "/", i diversi identificativi delle entità che compongono l'oggetto stesso:

Istituto - Progetto - Collezione - Oggetto digitale.

Digitalizzazione

Il ruolo di *digitalizzatore* effettua il caricamento delle immagini digitalizzate tramite client SFTP (o FTP) nell'area TMP nel repository di LISPA. L'area è opportunamente configurata ed è atta ad accogliere la mole di dati in entrata. Il trasferimento dati può avvenire anche offline tramite invio HD e successivo caricamento nell'area temporanea.

Terminato il trasferimento, il *digitalizzatore* accede al sistema ed esegue le procedure di conferma caricamento immagini.

Le procedure effettuano tutti i controlli necessari a validare il lavoro del *digitalizzatore* e trasferiscono le immagini nell'area di *storage* preposta.

Per ogni cartella/immagine l'applicativo effettua i controlli di coerenza (verifica struttura, *naming*, numero immagini, DPI). Nel caso in cui siano riscontrate anomalie, queste vengono scritte nel log dell'oggetto originale e la struttura dati non sarà copiata nell'area definitiva.

Nel caso in cui non ci siano anomalie, la struttura dati viene copiata nell'area di *storage* definitiva e viene generato l'oggetto digitale legato alle immagini appena inserite.

Infine, se il numero di immagini proposte risulta diverso dal numero di immagini digitalizzate, la verifica avrà comunque esito positivo: i dati saranno spostati nell'area definitiva e verrà inviata una email al *supervisore* e al *catalogatore* per segnalare la differenza riscontrata.

Aree dati

BDL mette a disposizione tre aree dati per la gestione delle immagini:

- l'area temporanea (TMP) dove le immagini vengono caricate temporaneamente, in attesa della verifica dell'applicativo, dal *digitalizzatore* o dalla gestione LISPA;
- l'area di *storage* (WRK) dove vengono trasferite le immagini dopo il processo di verifica e accettazione;
- l'area di archiviazione (DTD) dove vengono salvate su nastro per la preservazione le immagini master.

Esistono due modalità per il caricamento delle immagini:

- via SFTP, un'area privata, accessibile da internet previa autorizzazione, per poter caricare puntualmente gli oggetti nell'area temporanea BDL;
- via HD, l'hard disk esterno viene inviato direttamente a LISPA che si occupa di caricare i dati massivamente nell'area temporanea BDL.

Processo di validazione

Il processo di validazione e caricamento immagini prevede l'utilizzo di alcuni criteri per la determinazione della corretta formattazione e costituzione delle cartelle immagini presenti nell'area TMP:

- esistenza dell'istituto
- esistenza del progetto legato all'istituto
- esistenza della collezione legata al progetto
- esistenza dell'oggetto originale legato alla collezione
- progetto non concluso
- oggetto in stato "in catalogazione/catalogato"
- presenza delle cartelle master-large-medium-small-pdf-thumbnail-(pdfsingle opzionale)
- presenza del file <idOggetto>.pdf nella cartella pdf
- coerenza dell'estensione delle immagini presenti nella cartella
 - Master – file formato TIFF

- Large – medium – small – file formato JPG
- Thumbnail – file formato GIF /JPG
- controllo del nome immagine (ci si aspettano 4 cifre numeriche)
- controllo di sequenzialità delle immagini, ossia che non siano presenti salti di numero
- controllo del numero immagini rispetto al numero indicato dal *digitalizzatore*
- i DPI del formato sono controllati per ogni file di immagini
 - Master – 400 DPI
 - Large – 300 DPI
 - Medium – 150 DPI
 - Small – 72 DPI
 - Thumbnail – GIF/JPG 150x150px

Archiviazione

Il processo è eseguito in maniera automatica e, a fronte di copia su area TMP di *storage* di immagini MASTER, trasporta direttamente le stesse in un'area di *storage* differente configurata con policy di backup e ridondanza.

Catalogazione

Questo processo permette al *catalogatore* di completare la catalogazione dell'oggetto attraverso le interfacce di modifica degli attributi e di costruzione/gestione del TOC (Table of Contents). Al termine della lavorazione il *catalogatore* può inviare al *supervisore* l'oggetto da validare.

E' integrata nella procedura la funzionalità di Bookreader, ossia un'interfaccia web che consente la visione dell'oggetto digitale, inteso come insieme delle immagini che lo compongono, ordinate ed eventualmente corredate del TOC.

Per ogni oggetto è possibile:

- modificare i dati catalogafici nel caso di eventuale necessità;
- visualizzare la cronologia;
- inserire /modificare il TOC per l'oggetto (o importarlo da excel) e associare a ciascun nodo del TOC la rispettiva immagine.
- accedere all'anteprima immagini o alla visualizzazione finale;
- richiedere correzioni al *digitalizzatore* nel caso di immagini errate: invio segnalazione mediante email.

Concluse le modifiche / integrazioni, il *catalogatore* procede all'invio degli oggetti, arricchiti dei metadati aggiuntivi, al *Supervisore* per la relativa validazione per la pubblicazione.

L'inserimento del TOC è richiesto per tutti i documenti che presentano una significativa articolazione in parti riconoscibili autonomamente:

- pubblicazioni periodiche: in questo caso è necessario prevedere un indice con annate e – all'interno della singola annata – fascicoli;
- monografie: in questo caso il TOC permetterà di vedere direttamente l'indice della pubblicazione e l'inizio di ciascun capitolo.

Ulteriori precisazioni sulle scelte più opportune e coerenti riguardo a questo aspetto saranno concordate relativamente alle singole collezioni / fondi, nell'ambito della Fase 1 del progetto.

Processo di validazione per pubblicazione

Il processo di validazione per la pubblicazione consente al *Supervisore* di rendere disponibili per la pubblicazione gli oggetti che il *Catalogatore* ha verificato. Qualora il *Supervisore* non ritenga l'oggetto idoneo alla pubblicazione, lo rifiuta mandando un avviso al *Catalogatore*.

Successivamente, effettuata la corretta validazione, l'oggetto viene messo in stato di pubblicazione e risulta visibile all'esterno.

Il riconoscimento ottico dei caratteri (OCR)

E' prevista l'applicazione dell'OCR nei casi in cui le caratteristiche tipografiche e lo stato dell'esemplare lo consentano: a questo proposito si chiede che siano precisate le scelte in merito al software che sarà utilizzato e le modalità di effettuazione dei necessari successivi controlli.

L'applicazione o meno dell'OCR sarà valutata nell'ambito dell'analisi delle collezioni prevista nella Fase 1 del lavoro. Il piano di lavoro complessivo delle attività del progetto dovrà contenere l'indicazione esplicita per ciascuna collezione o fondo in merito all'applicazione o meno del riconoscimento ottico dei caratteri.

I metadati

Si prevede l'utilizzo di specifici schemi di metadati, scelti valutando anche le indicazioni delle principali istituzioni italiane (ICCU, Biblioteca Nazionale Centrale di Firenze, ecc.), ma anche tenendo conto del contesto internazionale di riferimento e dell'evoluzione al quale sono attualmente sottoposti alcuni progetti nazionali.

La scelta di quale schema di metadati utilizzare è caduta su METS, rispetto ad altre soluzioni più diffuse sul territorio nazionale, dal momento che l'ampio uso internazionale costituisce una garanzia di affidabilità nel tempo e di compatibilità. Inoltre:

- lo standard METS è un framework che definisce gli elementi che devono essere presenti, non definisce le "regole" secondo le quali scegliere e immettere i dati;
- METS è agnostico rispetto a quali schemi di metadati amministrativi o descrittivi verranno incorporati al suo interno

Per questo motivo, oltre a METS, sono stati individuati gli altri metadati tramite i quali verranno descritti da un punto di vista formale, amministrativo e tecnico gli oggetti digitali.

Per quello che riguarda i *metadati amministrativi gestionali* sono stati scelti:

- per la conservazione PREMIS
- per i metadati tecnici relativi alle immagini, MIX

Per quello che riguarda i *metadati descrittivi* sono stati scelti :

- per la descrizione a livello di collezione il profilo applicativo Dublin Core di Michael,
- per la descrizione dell'unità bibliografica (unità intellettuale, per utilizzare il lessico di Premis) gli Europeana Semantic Elements, ESE.

E' importante sottolineare che Europeana usa un profilo applicativo diverso rispetto a Cultura Italia, per la descrizione delle risorse. Europeana infatti ha adottato gli European Semantic Elements, ESE, mentre CulturaItalia ha creato un proprio profilo, PICO. La scelta di Regione Lombardia, caduta su ESE, consentirà comunque la compatibilità con CulturaItalia dal momento che è stata garantita la mappatura tra i due formati PICO ed ESE.

I dati prodotti dovranno essere disponibili all'harvesting tramite il protocollo OAI-PMH e OAI-ORE.

MIX

MIX (*Metadata for Images in XML Schema*), attualmente giunto alla versione 2.0, definisce un insieme di dati tecnici relativi alle immagini digitali necessari per la gestione di una collezione di oggetti digitali. Questi dati sono inseriti in uno schema XML.

MIX è uno standard internazionale sviluppato dalla Network Development and MARC Standards Office della Library of Congress, insieme al Technical Metadata for Digital Still Images Standards Committee del NISO (National Information Standards Organization) ed altri esperti del settore.

Tra gli elementi previsti vi sono:

- l'identificativo del file,
- la dimensione in bytes,
- il formato,
- l'eventuale compressione,
- la profondità del colore ...

I metadati MIX vengono creati in modo automatico o semi automatico: alcuni elementi vengono estratti dall'immagine stessa (la dimensione in bytes), mentre altri vengono impostati all'inizio dell'acquisizione (il formato di salvataggio e la profondità del colore).

OAI-ORE

Open Archives Initiative Object Reuse and Exchange⁵⁰ è un protocollo creato dalla Open Archives Initiative il cui obiettivo è permettere a repository differenti lo scambio di informazioni relative agli oggetti digitali. Ogni aggregazione di oggetti digitali, identificata tramite un URI (Uniform Resource Identifier), viene descritta dalla *Resource Map* che illustra le risorse che costituiscono tale aggregazione, le relazioni che vi intercorrono e le proprietà che le caratterizzano. Le *Resource Map* permettono quindi anche ad altre figure o istituzioni di comprendere gli oggetti digitali e di fornire servizi connessi (navigazione, stampa, visualizzazione, ecc.).

L'adesione al protocollo OAI-ORE non comporta alcun aggravio per il lavoro degli operatori impegnati nella realizzazione dei progetti di digitalizzazione, in quanto riguarda la struttura del repository dei dati.

METS e sua struttura

METS (Metadata Encoding and Transmission Standard) permette la codifica degli elementi necessari alla gestione degli oggetti digitali contenuti in un repository. E' stato sviluppato anche per permettere la creazione di strumenti e servizi finalizzati allo scambio di oggetti digitali e, di conseguenza, per favorire l'interoperabilità tra istituzioni, compresi i partner commerciali.

Grazie a METS è possibile registrare le relazioni che esistono tra le diverse componenti di un oggetto digitale, tra le sue sezioni e tra queste e i relativi metadati.

Un documento METS è costituito da sette sezioni principali:

1. Sezione Intestazione METS nella quale vi sono informazioni sul documento METS stesso (l'istituzione o l'autore responsabile, la data di creazione del file, ecc.)
2. Sezione Metadati Descrittivi (*dmdSec*) nella quale è possibile attivare un collegamento con una descrizione esterna (un record MARC, ad esempio), o inserire gli elementi descrittivi, oppure compiere entrambe le attività
3. Sezione Metadati Amministrativi (*amdSec*) suddivisi in:
 - metadati tecnici (*techMD*) relativi alla compressione dei file immagine, alla profondità del colore, ecc.,
 - metadati relativi ai diritti (*rightsMD*),
 - metadati relativi alla fonte (*sourceMD*), ossia all'oggetto digitalizzato,
 - metadati relativi alla provenienza digitale (*digiprovMD*)
4. Sezione File che presenta una lista di tutti i file che costituiscono l'oggetto digitale, anche riuniti in gruppi (*fileGrp*)
5. Sezione Strutturale nella quale viene delineata la struttura gerarchica che devono avere i file che costituiscono l'oggetto digitale per riproporre correttamente l'oggetto originario
6. Sezione Link Strutturali, utile soprattutto nel caso in cui si trattino siti web
7. Sezione Comportamento

La creazione dei metadati METS è gestita dall'applicativo BDL che, a partire dall'oggetto digitale pubblicato, genera in maniera automatica l'XML contenente i metadati catalografici, quelli relativi al TOC e quelli relativi alle immagini.