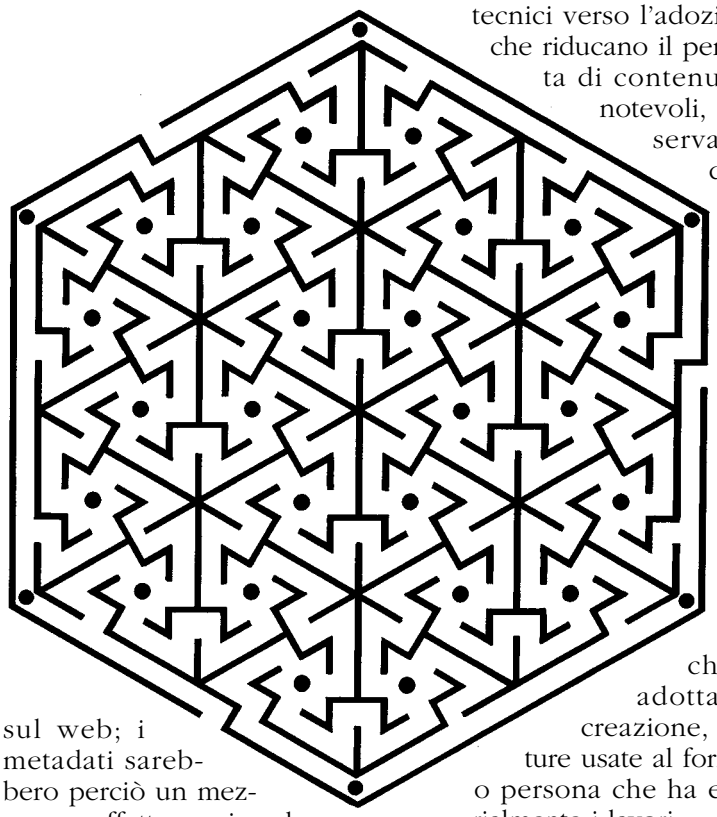
Nel labirinto dei metadati

A proposito di catalogazione e conservazione delle risorse digitali

di Guido Mura



Minacciato di estinzione, il bibliotecario ha reperito faticosamente un nuovo ambito d'intervento che potrebbe garantirne la sopravvivenza all'interno di un universo informativo in cui il libro, nella sua forma più usuale e tradizionale, non sarà più la fonte privilegiata di conoscenza. Tale nuovo ambito è la catalogazione delle risorse digitali mediante l'elaborazione di metadati.

Per chi ancora non avesse ben chiaro il concetto, i metadati consistono in una serie strutturata di informazioni relative a un oggetto costituito da un insieme di dati. In questa accezione, i metadati non sono altro che dati che riassumono, in maniera strutturata, informazioni relative ai dati originali. Non a caso le stesse schede bibliografiche, che forniscono i dati identificativi del materiale librario utilizzando una struttura organizzata per aree e campi, sono spesso indicate quale esempio di metadati. Una delle informazioni che sono o sembrano essere veicolate da una parte degli articoli e testi attualmente disponibili sui metadati motiva l'uso di questo strumento con la necessità di migliorare la ricerca sul web; i metadati sarebbero perciò un mezzo per effettuare ricerche valide in Internet, ponendo rimedio al caos dominante nell'ambito della rete.[1]

Si tratta in realtà di una definizione riduttiva delle finalità dei metadati, che come strumento possono essere applicati a vari contesti e soddisfare varie esigenze. Infatti, oltre che per facilitare e rendere più mirata la ricerca, i metadati possono e devono essere utilizzati per fornire informazioni sulla risorsa, allo scopo di garantirne la fruibilità nel tempo. Si parla in questo caso di metadati conservativi e il campo d'azione si sposta dall'universo web a quello del materiale acquisito in formato digitale con scanner, fotocamera o simili e temporaneamente registrato su un supporto che nella maggioranza dei casi è un cd recordable, cioè un disco registrato tramite un comune masterizzatore.

Sono soprattutto le preoccupazioni sulla durata di supporti, attrezzature e software di lettura a spingere i tecnici verso l'adozione di misure che riducano il pericolo di perdita di contenuti informativi notevoli, raccolti e conservati nell'ambito di progetti ad alto costo, spesso elaborati quando ancora gli strumenti informatici avevano costi proibitivi.

Si rende necessario, pertanto, catalogare la risorsa digitale indicando non solo i suoi contenuti, ma anche le tecniche adottate per la sua creazione, dalle attrezzature usate al formato, alla ditta o persona che ha eseguito materialmente i lavori.

Il problema fondamentale, che ha finora scoraggiato quanti fossero intenzionati a utilizzare in maniera concreta e non sperimentale i metadati, descrittivi e conservativi, nell'ambito del loro lavoro, è la proliferazione di proposte, standard, schemi, regole, tutti elaborati da prestigiosi enti e organismi per

lo più statunitensi o europei e tutti piuttosto criptici, come si conviene agli eminenti specialisti che elaborano e gestiscono i progetti. La compresenza di troppi e qualificati sistemi non può che consigliare di attendere la messa a punto di uno standard definitivo, soddisfacente e riconosciuto a livello internazionale, non solo dalle istituzioni che dominano l'universo della mediazione digitale, ma anche dal mercato, cioè dagli operatori e dagli utenti.

Per rendersi conto direttamente delle difficoltà collegate all'uso dei metadati, basta provare a compilare uno dei tanti schemi disponibili e soffermarsi ad esempio sul concetto di creatore della risorsa e su quello di *contributor*. Come indicare il creatore (persona o ente) di un'immagine digitale che riproduce un testo la cui paternità intellettuale sia chiaramente di altra persona o ente? Occorre quindi definire in modo chiaro le varie figure che intervengono nella creazione di una risorsa digitale, per evitare la confusione che attualmente pare regnare nelle pagine web, quando viene segnalato mediante un meta-tag l'autore della pagina.

Ulteriore cautela è da consigliare a chi volesse intraprendere in questo momento nuove iniziative di digitalizzazione. Siamo infatti in una fase di transizione, in cui tutto quello che finora è stato detto e consigliato sull'acquisizione di immagini digitali potrebbe tra breve tempo non essere più valido.

Gli stessi standard attualmente più diffusi per la circolazione e l'archiviazione di immagini sono in rapida e costante evoluzione. Già da qualche anno si ha notizia di un nuovo formato grafico, JPEG 2000, che dovrebbe sostituire il vecchio formato JPEG (Joint Photographic Experts Group), attualmente usato per le immagini collocate nei siti web.[2] L'elaborazione del prodotto è già a buon punto, tanto che la prima parte del nuovo standard è stata approvata nel febbraio del 2001 come standard internazionale.

Tra i numerosi e sostanziali miglioramenti apportati dal nuovo formato vi è la possibilità di realizzare immagini *lossless*, cioè immagini compresse senza perdita di qualità. Alcuni software che utilizzano JPEG 2000 sono attualmente in fase di sperimentazione. Il risparmio di spazio su disco di un'immagine JPEG lossless sembra essere più elevato rispetto al formato TIFF compresso col metodo LZW (Liv-Zempel-Welch compression),[3] ugualmente senza perdita di qualità. Nella sperimentazione che sto conducendo mediante un programma in versione di prova, un'immagine a colori in formato TIFF di circa 63 MB esportata in JPEG 2000 lossless è risultata di poco meno di 30 MB, il che costituisce un risparmio di spazio notevole.

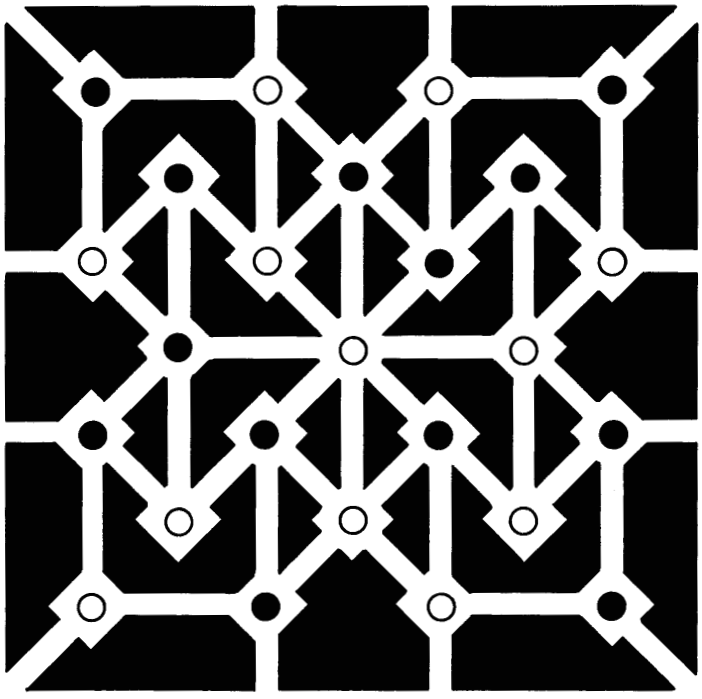
Altro vantaggio importante di JPEG 2000 è quello di basarsi su una tecnologia differente rispetto al precedente formato JPEG. Quest'ultimo si basava su una tecnica di compressione e di codifica dei dati nota come Discrete Cosine Transformation, o DCT, che otteneva alti livelli di compressione, peraltro regolabili a seconda delle esigenze, ma finiva col degradare l'immagine, che veniva suddivisa in blocchi quadrati di compressione, spesso chiaramente visibili, che venivano caricati secondo un ordine numerico, che rappresentava i dati dall'alto verso il basso. JPEG 2000 adotta invece la tecnologia *wavelet*, basata su espressioni matematiche che descrivono un'immagine in un flusso continuo; per questo motivo è da considerare un formato completamente diverso dal precedente JPEG. Con questo formato sarà possibile scaricare da un sito Internet un'immagine scegliendo la qualità desiderata, inserire nel file metadati in forma criptata, quali le informazioni relative al copyright, includere informazioni addizionali sul colore.

Accanto ai metadati, si sente parlare, sempre più spesso, di un linguaggio alternativo al comune e poco raffinato HTML, per la creazione di risorse testuali elettroniche. Si tratta di XML, che sta per eXtended Markup Language.[4]

Sia HTML che XML derivano da un linguaggio SGML, cioè Standard Generalized Markup Language, che mira a rappresentare un testo in forma elettronica mediante una serie di regole linguistiche che rendano la rappresentazione indipendente dalle varie piattaforme e dispositivi.[5] Però, mentre HTML è in definitiva un semplice sistema di formattazione di pagine di ipertesto, reso un po' più complesso dall'adozione dei fogli di stile (*style sheets*) e in particolare del CSSL (Cascading Style Sheets Language), XML è una sorta di sistema per definire altri linguaggi, che può essere gestito al meglio solo da chi abbia conoscenze di programmazione. Infatti alcuni concetti presenti in XML appartengono al bagaglio di conoscenze del programmatore. Inoltre, la sintassi XML è molto più rigorosa di quella del vecchio HTML, più limitato e approssimativo, e questo rende molto più difficile realizzare, anche con l'aiuto di editor specifici, pagine XML in grado di funzionare. Infatti, anche se i tutorial tendono a presentare XML come un linguaggio con cui è estremamente facile trattare, è indubbiamente frequente che un file XML non risulti ben formato (*well formed*), generando un messaggio di "fatal error". Spesso l'errore consiste nell'assenza di un tag di chiusura o nella mancanza di un elemento previsto nella DTD (Document Type Definition). Esistono comunque numerosi strumenti, ➤

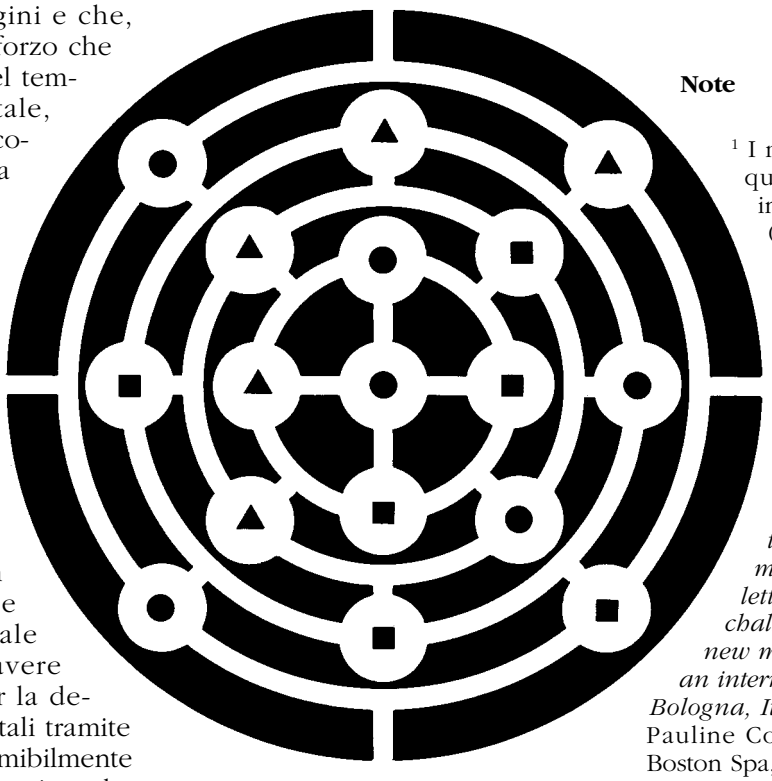
come ad esempio Lark, anche di uso gratuito, che consentono di analizzare e di apportare correzioni ai file XML.
Il vantaggio di questo linguaggio, che lo rende particolarmente appetibile al mondo accademico, è la capacità, insita nel sistema, di costruire sottoinsiemi di tag. In questo modo, dal numero definito di tag generici disponibili in HTML si può passare alla costruzione di gruppi di tag nuovi, di specifico utilizzo nell'ambito di particolari ambiti disciplinari ed elementi di riferimento di specifici motori di ricerca.
Lo svantaggio è che l'adozione di XML, rendendo più complessa la creazione e la gestione delle pagine, potrebbe ricondurre l'universo web nell'ambito operativo delle software house, anche per le operazioni di minore entità, come i lavori di manutenzione delle pagine e l'aggiornamento dei contenuti.
Per tornare ai metadati, è necessario notare che questi ultimi sono uno dei cavalli di battaglia del nuovo progetto relativo alla Biblioteca digitale italiana, per la cui realizzazione è stato costituito con decreto del Ministro per i beni e le attività culturali del 4 giugno 2001 un Comitato guida costituito da esperti, funzionari e docenti universitari. Lo studio di fattibilità della Biblioteca digitale italiana presentato recentemente a Padova si dibatte tra metadati e XML, propone esempi di progetti di digital library, dimenticando quello che è finora il più importante, utile, semplice, popolare ed efficace dei progetti italiani, il Progetto "Manuzio", forse il primo progetto dal volto umano in una selva di megaprogetti, nati con troppe pretese e con troppi finanziamenti, il che non incoraggia certamente a semplificare le procedure e ad agevolare il lavoro a studenti, ricercatori, amanti della cultura. Appena insediato, il Comitato guida ha preso probabilmente la decisione più sensata, lasciando a tempi migliori le attività per cui gli standard sono ancora in fase di definizione o di stabilizzazione e indicando

> come primo momento di sviluppo progettuale, anche in considerazione della sua più immediata realizzabilità, la scansione per immagini dei cataloghi manoscritti, sia in volume che a scheda, posseduti dalle biblioteche pubbliche statali.[6]



Per il momento, quindi, non ci si impegnerà nella più delicata riproduzione digitale di codici, stampe, libri d'arte e simili, ma si cercherà di dare un contributo importante, sia pure attraverso la creazione di immagini digitali, alla conoscenza del patrimonio librario nazionale, ricordando che la produzione di strumenti per la ricerca bibliografica rimane sempre l'obiettivo fondamentale delle biblioteche, insieme alla tutela dei beni librari nella loro fisicità e alla loro valorizzazione tramite l'attività espositiva.[7]
La Biblioteca digitale italiana è senza dubbio un progetto interessante e avrà sicuramente un futuro. Per poter risultare realizzabile, però, ha necessità di trarre alimento e fondamento da standard stabili e soddisfacenti e dall'uso di supporti e memorie di ampie dimensioni. In particolare, l'uso del cd come supporto privilegiato per le operazioni di archiviazione del lavoro realizzato non appare adeguato all'adozione di formati non compressi. Per far riferimento a un lavoro recentemente realizzato presso la Biblioteca nazionale Braidense, l'edizione ventisettana de *I promessi sposi*, in tre tomi, occupa nella versione destinata all'archivio digitale addirittura 21 cd in formato TIFF non compresso. Questa soluzione non sembra dunque accettabile, né in termini di economia operativa, né in termini di economia di spazio, sempre che non ci si accontenti di prodotti di qualità medio-bassa. Prima di affrontare gli oneri che comporta la costituzione di biblioteche digitali di ampie dimensioni, sarebbe meglio pertanto attendere l'affermazione di formati grafici durevoli più vantaggiosi di quelli attualmente disponibili e di sistemi di archiviazione dei dati più capaci e sicuri di quelli esistenti. A meno che non si rinunci definitivamente all'archiviazione su supporti teoricamente stabili ma di incerta durata, optando per soluzioni di tipo *caching*. Nel frattempo, sarebbe forse più opportuno

dedicare maggiore attenzione alla tutela del patrimonio bibliografico, con l'obiettivo di creare condizioni ambientali (microclima) idonee a garantire una migliore conservazione del libro nelle biblioteche destinate a tale compito, con un occhio di riguardo al materiale cartaceo del secondo Ottocento e del Novecento, che per le sue caratteristiche chimiche e strutturali corre maggiori rischi di degrado. Infatti, all'illusione che l'archiviazione in formato digitale possa costituire un'alternativa definitiva al supporto cartaceo va sostituendosi la consapevolezza che il libro, nella sua forma cartacea, costituisce il mezzo finora meno effimero per l'archiviazione di testi e immagini e che, pertanto, merita ogni sforzo che ne assicuri la durata nel tempo. La biblioteca digitale, piuttosto che proporsi come archivio eterno della cultura, deve prefigurarsi come servizio, che consenta una più ampia, veloce ed economica circolazione dei contenuti culturali, senza preoccuparsi troppo dell'eternità, che risulta essere finora un'aspirazione irrisolta della specie umana.
In questo contesto, non possiamo sapere quale fortuna, cioè quale reale diffusione, potranno avere gli attuali standard per la descrizione di risorse digitali tramite l'uso di metadati. Presumibilmente tali sistemi verranno adottati, anche in maniera obbligatoria, negli ambienti accademici e nelle strutture pubbliche o private che svolgono attività di conservazione e ricerca (biblioteche, archivi, musei), nei network proprietari, dovunque ci siano operatori da impegnare e notevoli finanziamenti. Nel mondo libero di Internet, finché esisterà e finché qualcuno, col solito pretesto di contrastare gli utilizzi illeciti o immorali della rete, non lo riempirà di regole ben più pesanti della netiquette, sembra difficile che i vari creatori di contenuti web trovino tempo e spazio, nella loro esplosione creativa, per catalogare tutti i loro effimeri prodotti riempiendo un congruo numero di campi in un editor per la costruzione di metadati. Il web libero, effimero e in larga misura gratuito tenderà probabilmente a differenziarsi sempre più dai servizi professionali che offriranno, a pagamento, un'informazione organizzata, correttamente veicolata tramite sistemi avanzati di ricerca.
Sarà l'utente a scegliere tra le diverse offerte e soluzioni e a decidere in base alle sue esigenze se affidarsi a una struttura che agevoli la sua ricerca o se procedere a caso, come il lettore di una volta, che anziché consultare il catalogo preferiva cercare direttamente negli scaffali e proprio in questo modo, incredibilmente, faceva le scoperte più interessanti, ritrovando notizie e percependo stimoli che nessun catalogo avrebbe mai potuto comunicargli. È chiaro che un lettore-navigatore di questo tipo, se qualcuno gli chiedesse il suo parere sui metadati risponderebbe: "Metadati? Grazie, ma non so che farmene!". Questo perché, malgrado il contrario avviso di qualche catalogatore, l'oggetto "bene culturale", libro, affresco, basilica o sito web, è sempre più ricco, complesso e importante della sua descrizione, per quanto raffinata questa possa essere. ■



Note

[1] I metadati sono presentati in questi termini nelle pagine introduttive curate dall'EUN (European Schoolnet), disponibili all'indirizzo <http://www.educat.hu-berlin.de/~kluck/datahandbook_V_300.htm>. Anche Riccardo Ridi pone l'accento sull'utilizzo dei metadati per la ricerca dell'informazione elettronica in rete: si veda la relazione *Metadata e metatag: l'indicizzatore a metà strada fra l'autore e il lettore*, in *The digital library: challenges and solutions for the new millennium. Proceedings of an international conference held in Bologna, Italy, June 1999*, edited by Pauline Connolly and Denis Reidy, Boston Spa, IFLA, 2000. L'articolo contiene una serie di spunti e condivisibili considerazioni; comprende inoltre le indicazioni bibliografiche necessarie per orientarsi nel mondo dei metadati.
[2] Su JPEG 2000 si possono reperire notizie in vari siti web. Si veda, ad esempio, la pagina in lingua italiana relativa al Progetto "Migrator", all'indirizzo <http://public.migrator2000. org/main.xalter>. Un articolo sintetico, ma abbastanza chiaro, sullo standard ➤

è consultabile all'indirizzo <http://www.zdnet.com/zdnn/stories/news/0,4586,2245956,00.html>. Un articoletto molto semplice, in italiano, si trova alla pagina <http://netart.stradanove.net/04/ graficaweb3.html>, che adotta un atteggiamento un po' troppo critico nei confronti del vecchio formato JPEG, utilizzato purtroppo, soprattutto in Italia, solamente ai livelli più alti di compressione, allo scopo di alleggerire le pagine web, con risultati a dir poco sconfortanti in termini di qualità dell'immagine. Molto chiaro l'articolo di ANTHONY CELESTE, *The future of web design*, presente alla pagina <http://www.designer.com/focus/articles/web_future/web_future_print.htm>, dove si parla anche di un formato ancora poco utilizzato, ma che dovrebbe presentare interessanti sviluppi, il formato PNG (Portable Network Graphics). Una serie di collegamenti a pagine di documentazione su JPEG 2000 si trova all'indirizzo <http://www.jpeg.org/JPEG2000.htm>.

[3] Il metodo di compressione LZW ha avuto origine da uno studio del 1977 di Jacob Liv e Abraham Lempel, aggiornato l'anno successivo dagli stessi ricercatori. Nel 1984 Terry Welch apportò delle modifiche al metodo, che divenne popolare con la sigla LZW. La compressione LZW non comporta perdita d'informazione, ma ha lo svantaggio di produrre nella maggior parte dei casi un file ancora troppo grande rispetto al file originario.

[4] Una pagina web che presenta vari link a siti che introducono il linguaggio HTML è la seguente: <http://www.echoecho.com/links/Tutorials/PageBuilding/XML/>.

[5] Cfr. la pagina web <http://www.isgmlug.org/whatsgml.htm>, dove si legge: "SGML (Standard Generalized Markup Language) is a language for defining markup languages. More specifically, SGML is a metalanguage formalism that facilitates the definition of descriptive markup languages for the purpose of electronic information encoding and interchange. SGML supports the definition of markup languages that are hardware – and software-independent, as well as applications-processing neutral. SGML is an International Standard, defined in the document ISO 8879:1986. Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML), as amended. A key philosophical commitment underlying SGML is separating the representation of information structure and content from information processing specifications. Information objects modeled through an SGML markup language are named and described (using attributes and subelements) in terms of what they are – from a defined perspective – not in terms of how they are to be displayed or otherwise processed".

[6] La citazione è ricavata da una recente nota del Ministero per i beni e le attività culturali.

[7] È auspicabile, comunque, che il processo di conversione dei cataloghi non venga attivato per quegli istituti dove già siano in fase di esecuzione progetti e procedure di recupero del pregresso in SBN.