Lo standard Sgml e la biblioteca elettronica

Come la "marcatura descrittiva generica" rivoluziona l'uso e lo scambio di documenti

di Anna Maria Tammaro

Lo scenario della trasmissione dell'informazione sta cambiando. Finora chi aveva bisogno di informazione andava in biblioteca. Doveva usare cataloghi, bibliografie e abstract (o fonti secondarie) dove trovava la risposta a queste domande:
— Cosa c'è? (cosa c'è in biblioteca, oppure cosa è pubblicato su un dato argomento).
— Dove è? (dove è collocato in biblioteca).
Alla fine di questa ricerca, poteva accedere al documento (o fonte primaria) di cui aveva bisogno, sempre che fosse disponibile.
Il "cosa?" ed il "dove?", nel tempo, si sono estesi: dall'ambito della singola biblioteca ad un sistema di biblioteche che condividono le loro collezioni bibliografiche. I bibliotecari hanno quindi costruito i cataloghi unici per consentire l' accesso ad una collezione più ampia di documenti, disponibili con il prestito interbibliotecario. Successivamente i cataloghi automatizzati sono stati resi disponibili in rete. Per gli utenti attualmente non è più necessario andare in biblioteca per consultare i cataloghi o le bibliografie: da casa, in rete, si sa cosa c'è e dove è disponibile.
Tuttavia, ancora oggi, l'utente di una biblioteca ha molte difficoltà ad ottenere i documenti che gli servono: dovrà prenotare ciò che gli interessa e visitare la biblioteca per averlo, oppure dovrà chiedere un prestito interbibliotecario, che è una procedura lenta per chi lo chiede e gravosa per i bibliotecari. Tanta fatica per poi accorgersi, dopo aver sfogliato il documento, che questo non gli è utile. Ci si è quindi resi conto che, per rendere veramente efficace l'accesso ai documenti, bisogna organizzare il servizio in modo che sia l'utente stesso, dalla sua stazione di lavoro, a collegarsi direttamente ai documenti di cui ha bisogno, dovunque questi siano collocati. Questo è possibile con una collezione di documenti elettronici liberamente accessibile in una rete di calcolatori (o server).
L'ultima sfida per le biblioteche è oggi trasformarsi in biblioteche elettroniche, cioè sforzarsi di fornire l'accesso ad una collezione di documenti elettronici. Ogni biblioteca automatizzata dovrà quindi, gradualmente, diventare una biblioteca elettronica. Indispensabile prerequisito sarà l'adozione di sistemi aperti e di standard. Possiamo dire che gli standard sono il ponte tra le strategie di sviluppo dei servizi di biblioteca e la tecnologia. Quali standard? Certo non quelli tradizionalmente noti, come ad esempio quello conosciuto da tutti i bibliotecari: il Marc o le regole di catalogazione. Questi standard tradizionali trattano solo record bibliografici, cioè citazioni di documenti e non interi documenti o meglio documenti composti di testo, immagini e suono, come quelli che caratterizzeranno la biblioteca elettronica. Bisognerà quindi ricorrere a nuovi standard, nati fuori dall'ambito delle biblioteche, ma il cui uso si è esteso al di là dell'obiettivo iniziale per cui sono stati sviluppati. Come Sgml, Standard Generalized Markup Language.

1. COSA È SGML?

Sgml è definito da uno standard Iso con il numero 8879: 1986 ed il suo nome completo è Information Processing - Text and Office Systems - Standard Generalized Markup Language. Lo standard è nato come Gml (dal nome degli autori, Goldfarb, Mosher e Lorie) nel 1969 all'Ibm, ed è stato usato per l'editoria prima di diventare standard Ansi Z39.59 e successivamente standard Iso.
La biblioteca elettronica fornirà l'accesso ai documenti sia memorizzandoli in un proprio magazzino di dati, alternativa moderna dei depositi chiusi, sia facilitando l'accesso ad altri documenti archiviati in numerosi magazzini tutti collegati alla rete. I sistemi di editoria elettronica devono quindi essere interoperabili: ciò perché sistemi diversi distribuiti in rete dovranno lavorare come se fossero un solo sistema. In questo ambito flessibile e aperto, ogni sistema può mantenere le sue capacità locali ed interagire con altri calcolatori per condividere informazioni in rete. Sgml consente questo: lo scambio di documenti elettronici indipendentemente dal calcolatore utilizzato ed indipendentemente dall'applicazione

che si vuole usare (sistema di scrittura, programmi di information retrieval, ipertesti, sistemi esperti, concordanze, ecc.). I documenti in Sgml inoltre saranno sempre validi anche in futuro, quando i sistemi operativi, le macchine ed i programmi utilizzati avranno caratteristiche che non possiamo prevedere. L'informazione creata oggi potrà essere utilizzata in futuro in modo non previsto. In altre parole, i documenti preparati con Sgml possono essere utilizzati, senza nessuna conversione, per diverse funzionalità e con tutti i software. L'ideale della compatibilità totale, spesso proclamato dai fornitori ma sempre difficoltoso nelle realizzazioni, con Sgml è una realtà.

L'idea centrale dello standard è un tipo di marcatura generica chiamata "marcatura descrittiva" (descriptive markup) che definisce la struttura logica dei documenti. Il termine marcare (markup) è preso dall'editoria, dove viene usato per la codifica sul manoscritto delle istruzioni date dall'editore al tipografo. Questa marcatura usata dai tipografi riguarda essenzialmente lo stile di presentazione del testo a stampa e si chiama più propriamente "marcatura procedurale" (procedural markup). Oggi è spesso lo stesso autore che, usando i fogli di stile disponibili in tutti i programmi di scrittura, predispone il testo per l'editore.

Tutti hanno esperienza di un primo problema: le codifiche di marcatura specifiche di un sistema di scrittura normalmente non sono riconosciute da altri sistemi di scrittura o da edizioni diverse dello stesso sistema di scrittura. Così sono necessarie delle modifiche e procedure di conversione che a volte possono richiedere all'autore del documento o al tipografo dal 15 per cento al 50 per cento del tempo totale di produzione del documento. Se si cambia lo stile di presentazione, o se si cambia programma, c'è il rischio di perdere informazione: così si può condividere il documento solo con altri utenti dello stesso programma; inoltre il documento può essere utilizzato solo per la stampa. La marcatura descrittiva risolve questo problema. È un'idea di una grande semplicità quella di tenere separato lo stile di presentazione dalla struttura logica del documento, eppure è proprio questa semplice trovata che consente molteplici utilizzazioni della stessa informazione: con la marcatura descrittiva è possibile utilizzare lo stesso documento per la stampa oppure per la ricerca in linea, per la pubblicazione su cd-rom oppure per il riuso del documento integrandovi altro testo.

La novità della *marcatura descrittiva generica* dello standard Sgml è che identifica il contenuto del documento usando una notazione che esprime cosa è che compone il testo e non come deve apparire nel testo. L'organizzazione di un documento non è espressa quindi usando la codifica dei sistemi di scrittura, che è finalizzata alla presentazione grafica, ma sono evidenziate le parti in cui è strutturato il documento (ad esempio paragrafi, capitoli) insieme ad altre particolarità del testo (come note, tabelle, intestazioni). In un documento si possono trovare elementi ed entità. Ogni testo è infatti strutturato in alcuni elementi che sono parti identificabili, come la prefazione, i capitoli, i paragrafi, l'indice. Nei testi possono trovarsi anche entità diverse, cioè il testo può essere separato in più archivi, che vanno collegati al momento della presentazione (a video o a stampa) all'utente. Questi archivi separati possono essere esterni (ad esempio immagini, tabelle, liste) o interni al testo (come note, citazioni, notazioni speciali). Le entità devono essere visualizzate dove è presente la dichiarazione breve di entità (in pratica ciò richiede un programma particolare che richiama le entità identificate da un nome preciso).

La marcatura descrive la struttura logica di un documento usando un modello gerarchico che riflette la interrelazione nel testo di elementi ed entità. La struttura logica del documento è definita nel Document type definition (Dtd). Sgml può definire un numero infinito di tipi di documento, per ogni categoria di informazione prodotta: libri, articoli, bollettini, manuali, cataloghi, specifiche di disegno, lettere, ecc. Il Dtd è la caratteristica più importante di Sgml, perché per ogni tipologia di documenti codifica gli elementi che possono essere presenti in un documento o in un insieme di documenti e le relazioni tra questi. Il Dtd è come lo schema di una banca dati che descrive il tipo di informazione e le relazioni tra i vari campi che si possono trovare nei dati memorizzati. È una cornice concettuale per le parti che costituiscono il documento.

Lo standard Sgml può dirsi la grammatica di un documento in cui gli elementi che lo compongono sono il vocabolario utilizzato mentre la loro relazione (cioè il ruolo nel testo di elementi con altri elementi e con le entità associate che compongono il documento) rappresenta la sintassi astratta del documento. Il Dtd definisce questa sintassi. Ogni elemento nel Dtd è assimilato ad un concetto e quindi è un'astrazione delle caratteristiche proprie di tutte le occorrenze reali di una classe di oggetti simili e, come un concetto, è gerarchicamente connesso ad altri concetti. L'elemento nel Dtd è definito da un nome (*generic identifier*) e dalla descrizione del contenuto (*content model*). La descrizione del contenuto è una potente sintassi astratta dei vari contenuti che possono trovarsi in un dato elemento e consente di identificare parti distinte di un documento da un punto di vista concettuale e gerarchico. Questa particolarità dello standard è di grande rilevanza per i bibliotecari. Ad esempio, se nel documento si trovano alcuni dati che possono essere per gli utenti punti di accesso privilegiati al documento, si identificano con un "tag" che ne consentirà il riconoscimento. Ed ancora, occorrenze che si ripetono (o cosiddette parametriche) possono essere segnalate con un segno particolare (di solito il segno di percentuale %) per essere riconosciute. È così possibile attuare l'indicizzazione automatica dei documenti in Sgml.

Ogni documento in Sgml è diviso in tre parti:[1]

1) la prima parte si chiama "Sgml declaration" ed appare come segue: <!Sgml "Iso 8879:1986". Definisce il set di caratteri usato,[2] caratteri a cui è attribuito un particolare signi- ➤

[1] Per una descrizione dettagliata dello standard rimando al libro di A. Scolari, *Gli standard Osi per le biblioteche*, Milano, Editrice Bibliografica, 1995, p. 118-131.

[2] In testi multilingue o in documenti in lingue che richiedono più dell'Ascii, come ad esempio le lingue asiatiche Cjk-China Japan Korea, dove sia necessario un set esteso di caratteri, è stato proposto l'Extended Reference Concrete Syntaxes (Ercs) che si occupa sia della marcatura nel linguaggio nativo sia dello scambio della marcatura in diversi set di caratteri.

ficato ed altre particolarità che sono usate. Tutto ciò si omette se si usano i valori di default della sintassi del Dtd.
2) La seconda parte specifica il Dtd cioè la struttura del tipo di documento: <!DOCTYPE ... (seguito dal nome della classe di documenti).
3) La terza parte è il contenuto del documento (occorrenza o "instance" di quella classe di documenti specificata nel Dtd) ed include ad esempio il titolo, i paragrafi, gli indici, le tabelle, i grafici ed eventuale audio collegato. Ogni elemento è definito da un'etichetta (tag) che precede e conclude il contenuto. Ad esempio: <TITOLO>Contenuto di questo elemento</TITOLO>.

2. COME SI USA?

Se la grammatica e la sintassi definite dal Dtd di Sgml sono seguite correttamente, l'informazione sul contenuto del documento può essere trasportata insieme al documento stesso, senza problemi di conversione o incompatibilità. Sgml è un vero e proprio metalinguaggio, un sistema di codifica dei testi efficace in un sistema distribuito ed aperto di calcolatori che interoperano. Sgml è perfettamente neutro: non è un formato di dati ma può essere travasato in qualsiasi formato, non è un software ma può essere utilizzato con qualsiasi software, anzi dovrà essere necessariamente utilizzato con diversi software. Di per sé non fa nulla ma con Sgml si può fare di tutto. Occorreranno però, nell'ordine: software per creare o convertire documenti in Sgml, software per l'indicizzazione ed il recupero dei testi, software per l'uso di documenti Sgml (cioè la presentazione a video dei documenti o la stampa).

2.1 I software per l'input di documenti in Sgml

L'idea di Sgml è semplice ma l'applicazione dei vari Dtd presenta notevoli difficoltà, per cui si può dire che per realizzare una collezione di documenti Sgml è richiesta prima di tutto chiarezza di obiettivi e, una volta raggiunta questa chiarezza, un'approfondita analisi del Dtd del documento. Per questa analisi c'è un programma, in particolare, di cui non si può proprio fare a meno.[3] La segmentazione del documento in parti infatti non deve essere manuale ma esiste un programma che fa questa operazione in modo semiautomatico, con un semplice clic del mouse o aiutando il compilatore, mostrandogli solo le etichette usabili in una determinata parte del testo. Ciò consente anche ai non esperti di comporre la marcatura dei documenti in Sgml. Il programma che effettua queste funzionalità si chiama Parser.
Lo scopo di un programma Parser è di guidare la marcatura dei documenti Sgml e di controllare se i documenti sono o no conformi allo standard. Infatti, nel caso di errori di marcatura, il Parser li segnala: questo processo si chiama validazione. Dopo la validazione ogni documento può essere riformattato nell'applicazione scelta, esattamente come se fosse stato creato direttamente con quel programma. Insieme al gestore di entità (il componente che rende possibile segmentare i documenti Sgml in entità separate) forma la base di ogni sistema conforme a Sgml. Di solito il Parser è parte di un programma complesso che esegue l'editor, la conversione, registra i documenti ed eventualmente realizza anche i programmi di ricerca, navigazione e presentazione.
Attualmente la situazione più frequente è però non tanto quella di realizzare un documento Sgml, quanto quella di convertire un testo, costruito usando ad esempio uno dei vari sistemi di scrittura, in Sgml. Infatti sia la conversione retrospettiva delle pubblicazioni a stampa (lette con uno scanner), sia l'opera degli autori accettata nel formato elettronico originale, creano il problema di passare ad un formato neutro come Sgml prima di essere utilizzati con qualsiasi altro software. Lo scopo dei programmi di conversione esistenti è di rendere esplicita la marcatura che è implicita nel documento. Se la marcatura non esiste deve essere aggiunta automaticamente. Questa esigenza ha dato avvio a "Fred Sgml Grammar Builder", un progetto in corso presso l'Oclc per la formattazione automatica di documenti che non hanno un Dtd.[4] L'Oclc riceve infatti molti documenti con il testo formattato ma non in Sgml e con questi documenti deve passare prima ad un formato neutro come Sgml per costruire successivamente banche dati che si adattino alle interfacce già esistenti ed ai programmi di ricerca abitualmente utilizzati da Oclc. Per evitare una trasformazione arbitraria, Fred ha un linguaggio di traduzione che consente di confrontare la struttura stessa dei testi, muovendosi all'interno dell'albero gerarchico di Sgml, suggerendo al compilatore i tag relativi a elementi, attributi ed entità propri del Dtd di quel documento.
Può anche verificarsi il caso della traduzione in altri Dtd di Sgml, come ad esempio il più diffuso Html, o Icadd (per la produzione del Braille) o altri. Ogni programma di conversione dà accesso all'insieme degli elementi del Dtd (compreso inizio e fine degli elementi), alle eventuali entità interne ed esterne, al contenuto degli elementi e dà istruzioni procedurali per passare da un Dtd ad un altro.

2.2 Software per la gestione di documenti in Sgml

I documenti in Sgml, originali o convertiti, andranno memorizzati su supporto elettronico al fine di consentire sia la ricerca che la navigazione tra i dati. La memorizzazione comporta quindi anche l'attività di indicizzazione, a seconda dei casi manuale (fatta dal bibliotecario) o automatica.
Attualmente molti sono i pacchetti di recupero dell'informazione che accettano testi in formato Ascii o nei più comuni sistemi di scrittura. Questi programmi, che impiegano filtri per lo scambio di documenti elettronici, appiattiscono però la struttura dei documenti a testi Ascii e realizzano banche dati in cui si perde ogni struttura dell'informazione. Pochi ancora oggi sono invece i programmi che possono gestire documenti in Sgml. Tra questi si può fare una distinzione

[3] Un rapporto tecnico Iso (Iso/Iec Tr 10037) fornisce una guida alle funzionalità di un editor.
[4] Il software è reso disponibile gratuitamente presso il server Www di Oclc.

tra i sistemi quasi-Sgml e quelli prettamente Sgml. La differenza sta in questo: i sistemi quasi Sgml sono capaci di importare o esportare dati da documenti Sgml con la possibilità di estrarre informazione dalla struttura Sgml in modo batch; i sistemi Sgml invece arrivano alla completa o parziale mappatura di Sgml nella struttura di una banca dati ed operano direttamente sui documenti. Esistono quindi programmi che consentono la gestione di banche dati di documenti in formato Sgml, sia per la semplice archiviazione sia per funzionalità più complesse come il recupero di informazione (sia la ricerca esatta che la navigazione) da parte di molteplici utenti remoti. Tra questi programmi può essere ricordato ad esempio Basis, dell'Information Dimension, scelto anche da Sbn per la realizzazione dell'opac di Indice. In Basis viene conservata la struttura originale dei documenti, con grande vantaggio dell'indicizzazione automatica e del recupero dell'informazione a testo libero.
Il metodo più diffuso attualmente è però quello non di usare un sistema di ricerca Sgml ma di caricare su altri programmi di ricerca i documenti in Sgml e quindi tradurre i documenti Sgml nel formato usato da quei sistemi di ricerca. Va segnalato un particolare tipo di traduzione immediata (o più propriamente "on the fly"): Cgi Common gateway interface, un programma che consente a Web l'accesso a documenti più complessi di Html (ad esempio Sgml), traducendo da uno all'altro Dtd.[5]
Sia i sistemi Sgml che gli altri sistemi non Sgml che gestiscono testi pieni dovranno essere migliorati. Questo perché va approfondita l'adattabilità delle soluzioni che sono state realizzate per le problematiche tipiche della ricerca bibliografica (cioè della ricerca su citazioni bibliografiche che rimandano a documenti cartacei) alle nuove difficoltà che sono proprie della ricerca di documenti a testo libero e di documenti multimediali. I programmi che utilizzano completamente la struttura dei documenti Sgml potranno in futuro forse rendere superata l'attività di indicizzazione operata da parte dei bibliotecari cioè sia la catalogazione che la compilazione di bibliografie. I dati inseriti dai bibliotecari secondo certe regole manualmente potranno essere estratti dai documenti in Sgml, seguendo forse anche le stesse regole, automaticamente, a tutto vantaggio dell'immediatezza dell'informazione oltre che della correttezza ed univocità dei dati.[6] Lo stesso principio si potrà applicare anche alla indicizzazione semantica dei documenti. Non va infatti sottovalutata la possibilità dell'indicizzazione automatica per argomento. Occorrerà spirito critico ma si trarranno grandi vantaggi utilizzando al meglio le possibilità dello standard Sgml e le funzionalità dei nuovi programmi di ricerca a testo pieno. I bibliotecari non hanno un'esperienza adeguata per l'indicizzazione di documenti a testo pieno e per l'indicizzazione di documenti multimediali. Finora le biblioteche automatizzate indicizzano solo dati bibliografici che sono memorizzati negli opac; per il recupero dei dati memorizzati usano quasi esclusivamente la ricerca booleana basata sulla ricerca esatta di parole chiave: solo campi brevi e strutturati sono coinvolti nel recupero dei dati. Con il testo pieno ed i documenti composti di entità diverse, che comprendono anche suoni ed immagini, ben diverse sono le difficoltà e quindi non si può trasportare meccanicamente ed acriticamente l'esperienza bibliotecaria fin qui fatta.
Gli attuali sistemi di recupero dell'informazione elettronica che gestiscono anche testo pieno hanno esteso la ricerca booleana a sistemi basati sul *contenuto* in opposizione alle parole chiave (cioè la ricerca avviene su frasi, sentenze, paragrafi e argomenti e non su brevi campi). Per la ricerca possono essere usate varie tecniche come la ricerca su ogni parola del testo, la ricerca booleana, la ricerca per prossimità, la ricerca fuzzy. I sistemi più avanzati tra quelli esistenti si basano essenzialmente su un approccio statistico ed usano *modalità di ricerca "fuzzy"* che è l'opposto della ricerca esatta (o "match esatto") poiché i risultati sono allargati a documenti in qualche modo collegati alla richiesta e si ha una risposta elencata in ordine di rilevanza; altri sistemi basati sull'intelligenza artificiale sono ad un livello di sviluppo ancora insufficiente. Quello che andrebbe evitato è la proliferazione di programmi di ricerca e sistemi di memorizzazione di documenti elettronici, poiché sistemi non standard e chiusi potrebbero rendere difficile, in un prossimo futuro, l'obiettivo di una completa disponibilità in rete di magazzini di documenti elettronici.
Un ulteriore problema di ricerca dei documenti elettronici è quello connesso alla necessità di utilizzare un'architettura distribuita, sia a livello nazionale che internazionale, di server interconnessi come magazzini di documenti elettronici da cui recuperare informazione. La ricerca quindi dovrà avvenire su una molteplicità di archivi in rete e non essere limitata ad un solo sistema. Una stessa richiesta, trasmessa da un client a vari server, riceverà più risposte che il client presenterà al suo utente come se la ricerca fosse stata effettuata in un solo archivio di documenti. Wais, Gopher e Web sono i più diffusi e noti sistemi di ricerca utilizzati nell'architettura client-server che consentono la ricerca e l'uso di documenti a testo pieno.

2.3 Software per la presentazione di documenti in Sgml

Per gli utenti certamente il cambiamento più grande nell'uso della biblioteca elettronica sta nell'accesso in linea ai documenti stessi piuttosto che ad elenchi che rimandano ai documenti cartacei da cercare in varie biblioteche. La possibilità di recuperare l'informazione senza andare in biblioteca non è da poco: non ci sono limiti di orari o restrizioni di

[5] È molto importante per lo sviluppo futuro della biblioteca elettronica conoscere le funzionalità di Cgi. Per saperne di più vedere all'indirizzo: Url: http://hoohoo.ncsa.uiuc.edu/cgi/overview.
[6] Per assistere l'utente nella ricerca la Biblioteca dell'Università di Berkeley ha promosso il Finding Aid Project. I finding aid sono documenti usati per descrivere, controllare e consentire l'accesso a collezioni di documenti, cioè sono le bibliografie della biblioteca del futuro. Come le bibliografie, i finding aid sono documenti in Sgml che citano i documenti primari. Lo standard Sgml assicura che i finding aid rimarranno validi nel tempo, anche se cambierà hardware e software. Attualmente sono circa 100 le liste rese disponibili dalle biblioteche che aderiscono al progetto.

➤

servizio, l'intero sistema mondiale delle biblioteche sta su una scrivania. È possibile usare i testi elettronici in due modi: per la ricerca e la lettura dalla stessa stazione di lavoro o, in modo più tradizionale, per la stampa, appena completata la trasmissione.
Normalmente nella ricerca in rete di documenti a testo pieno, i risultati sono visualizzati in due fasi: la prima fase elenca in forma breve i documenti recuperati, nella seconda fase si accede ai documenti scelti dalla lista. C'è però un primo problema: la tendenza, dopo la presentazione a video del documento, è di tornare alla carta e stampare ciò che si vede a video.
Il problema è culturale e tecnologico insieme. È abbastanza normale che gli utenti preferiscano leggere gli articoli su carta invece che a video. Le abitudini di lettura e studio hanno un peso determinante nell'uso dell'informazione elettronica. Questa preferenza è però in parte motivata da limiti tecnologici come ad esempio dal fatto che è difficile visualizzare una pagina su un video. Se la pagina riesce ad entrare nello schermo, la risoluzione è bassa; aumentando la risoluzione la pagina è chiara ma si può leggere una parte alla volta. Anche scorrere a video un testo presenta problemi. Non è possibile avere sottomano una dozzina di articoli contemporaneamente, poiché attualmente la tecnologia dello scorrimento è lenta ed anche se è possibile aprire più documenti la ristrettezza dello schermo impedisce la possibilità di vedere simultaneamente più documenti. I sistemi elettronici andrebbero utilizzati per quello che riescono a fare meglio: prestazioni sofisticate nella ricerca di parole o frasi, possibilità di riordinare e di rielaborare i risultati dalla stessa stazione di lavoro con cui si effettua la ricerca, capacità di seguire i legami tra documenti cliccando su un'icona o una parola. Altre possibilità da non sottovalutare sono la capacità di fare zoom per i grafici o di visualizzare i documenti in fonti diverse, ad esempio in Braille per i non vedenti.
Un altro problema in cui la tecnologia è ancora abbastanza indietro è la trasmissione dei testi: bisogna oggi scegliere tra lentezza e qualità. Di solito la qualità della presentazione è migliore con la trasmissione delle immagini del testo e delle figure collegate, così come apparirebbero in un testo a stampa. Ma ciò comporta tempi lunghi per lo scaricamento delle immagini, anche se queste possono essere compresse e poi decompresse. Il limite più grande di questa scelta è che è impossibile la ricerca per parole dei testi come anche tutte le altre possibilità di navigazione ipertestuale in una collezione di documenti elettronici.
La presentazione e la trasmissione, nel caso di documenti in Sgml, può essere in testo ed immagini insieme. Originariamente la presentazione a video non era contemplata tra gli obiettivi dello standard Sgml ma attualmente si sta cercando di realizzare una migliore presentazione grafica dei documenti in Sgml: Dssl-Document Style Semantics and Specification Language (Iso Dis 10179) è uno standard collegato ad Sgml, in corso di realizzazione, per una presentazione grafica uniforme dei documenti in Sgml. Per la stampa di documenti Sgml occorre anche dotarsi di sistemi di stampa ad alta capacità. Se si vuole la stampa del documento occorrono anche sistemi di impaginazione, batch o interattivi, che gestiscono il salto pagina, che è legato alla carta e non a come viene visualizzato sullo schermo il documento.
Se si sceglie la fornitura elettronica come modo certamente più avanzato di usare i documenti elettronici, le differenze tra i vari programmi che realizzano le funzionalità di ricerca riguardano le possibilità attuate dall'interfaccia scelta per l'utente.[7] L'interfaccia utente deve infatti aiutare la transizione dalla carta ai documenti elettronici, per una lettura che è simile a quella dei libri a stampa (o meglio è simile a quella di un papiro) ma può fare anche molto di più. Molte le nuove possibilità di ricerca: il contenuto può essere sfogliato non solo sequenzialmente (dalla prima all'ultima pagina) come in una pubblicazione su carta ma anche in modo gerarchico, seguendo la struttura del Dtd del documento, oppure navigando tra i legami ipertestuali di più documenti collegati. La stazione di lavoro dell'utente dovrà essere appropriata a questo uso (cioè dotata di hardware e software che consenta la lettura di documenti elettronici).
Nel futuro c'è da prevedere che la combinazione perfetta di sistemi di ricerca che applicano l'architettura client-server in una molteplicità di magazzini di documenti in Sgml consentirà con sempre maggiore facilità la fornitura elettronica dei documenti, che sarà preferita alla stampa. Inoltre il servizio di fornitura elettronica dei documenti in Sgml sarà possibile con tutte le piattaforme e permetterà l'uso di un set esteso di caratteri. Con il crescere di collezioni di documenti in Sgml e la diffusione di sistemi di ricerca Sgml, sarà così per gli studiosi più facile sviluppare e seguire le idee.

3. SGML ED ALTRI STANDARD PER IL TRATTAMENTO DEI TESTI

L'unica valida alternativa ad Sgml attualmente è rappresentata da un altro standard Iso che è Oda. Oda (Open document architecture Iso 8613) come Sgml definisce la struttura di documenti elettronici e multimediali con la differenza che si preoccupa anche della presentazione all'utente, cosa in cui Sgml è ancora abbastanza carente. Mentre Sgml è nato nel mondo dell'editoria, Oda è nato per l'automazione d'ufficio. Entrambi dai documenti elettronici si sono poi estesi al trattamento dei documenti multimediali.
Oda definisce sia la struttura logica del documento (che è simile alla sintassi del Dtd) che la struttura di disposizione (la semantica del documento) legata all'apparenza fisica. Oda gestisce tre tipi di contenuto: testo scritto, grafici geo-

[7] Alcuni prototipi che studiano le possibilità di ricercare documenti Sgml sono nati per le specifiche esigenze delle biblioteche. Il Progetto Core ad esempio è un prototipo di biblioteca elettronica che consente l'acesso elettronico al testo pieno ed ai grafici delle riviste dell'American Chemical Society ed alle riviste di indicizzazione associate, a partire dal 1980 (circa 250 annate di riviste e più di 600.000 pagine). La banca dati è stata convertita in Sgml partendo dalla formattazione tipografica dei testi. Il prototipo è significativo per esplorare i problemi della struttura della banca dati e le soluzioni adottate per i problemi di ricerca di riviste scientifiche a testo pieno. Il Progetto Elsa (Electronic Library Sgml Applications) è un progetto europeo con tre partner: Jouve, Elsevier e De Monfort University. Il suo scopo è di utilizzare Sgml per pubblicare documenti in rete. Un client consentirà di accedere ai documenti e visualizzarli.

metrici e immagini. Dà più possibilità di descrizione dei documenti elettronici e multimediali di quelle possibili con Sgml ma questo a spese di maggiore complessità e quindi maggiori costi di applicazione. Per questo motivo è meno diffuso di Sgml nell'industria e più usato in ambienti accademici. La diffusione e l'accettazione sono la condizione più importante per la sopravvivenza di uno standard e gioca a favore di Sgml. La ricchezza delle applicazioni Sgml esistenti, anche a livello commerciale, non ha confronti con quelle di Oda, che è assai poco usato.
È molto noto anche PostScript che è però solo un linguaggio di descrizione della pagina. Descrive la presentazione fisica del documento e non la sua struttura logica. Gli archivi PostScript sono generati spesso dai sistemi di scrittura o da programmi per l'editoria. I documenti in PostScript possono quindi essere solo trasmessi in rete per la stampa.
Alcune estensioni di Sgml sono più diffuse ed importanti di Sgml stesso, perché consentono il trattamento di documenti multimediali con applicazioni dell'ipertesto. Questa caratteristica ha valore poiché in futuro l'ipertesto sarà una tecnologia fondamentale per i documenti elettronici ed i documenti multimediali. Un'estensione di Sgml è HyTime- Hypermedia-Time based structuring language (Iso 10774) che è lo standard Iso molto diffuso e corrispondente allo standard Ansi Smdl-Standard music description language. HyTime rende possibile lo scambio di documenti multimediali e ipermediali, cioè l'ipertesto è combinato alla tecnologia multimediale.
L'applicazione di Sgml più nota è però senza dubbio Html-Hypertext markup language, lo standard per i documenti ipertestuali nei siti Web in Internet. Ha un Dtd più semplice di Sgml, anche se le versioni più aggiornate di Html sono sempre più complesse e quindi più simili ad Sgml. Come per altri standard de facto di Internet, viene spesso visto in competizione con lo standard Iso che definisce Sgml, ma questa competizione non c'è perché Html deriva da Sgml. A riprova di ciò esistono esperienze in cui documenti in Sgml e documenti in Html vengono usati nella stessa applicazione con una traduzione automatica tra i due Dtd direttamente in linea.

CONCLUSIONI

L'accesso a documenti elettronici sarà il nuovo servizio organizzato dalla biblioteca. In futuro anche le procedure tradizionali di acquisizione e catalogazione saranno rivoluzionate dalla presenza del nuovo supporto elettronico d'informazione. Molti i problemi che i bibliotecari dovranno risolvere. Ad esempio: quale è il ruolo dell'indicizzazione per argomento in un testo pieno? L'indicizzazione automatica di parole o frasi potrebbe essere concentrata su parti significative del testo come ad esempio il frontespizio, gli indici, l'introduzione e la conclusione, o invece dovrebbe essere estesa a tutto il libro? Potrà l'indicizzazione automatica rendere superflua l'indicizzazione manuale del bibliotecario? Quale è per il bibliotecario il significato di acquisire, gestire e preservare una collezione elettronica? L'acquisto potrà essere limitato all'accesso presso un server remoto oppure si dovranno acquisire documenti elettronici presso il calcolatore locale? Nell'uno o nell'altro caso molti i cambiamenti necessari nella catalogazione: come è catalogato questo materiale? Vale la pena di distinguere tra ciò che è raccolto nella collezione della biblioteca, su supporto cartaceo o elettronico, e ciò a cui si può accedere in rete? La conservazione comporterà come nuovo problema quello di fissare l'edizione corretta di un documento elettronico: quali garanzie altrimenti che il documento non sia mutato nel tempo?
Il bibliotecario potrà liberarsi dalla supremazia (e dalle problematiche di gestione) del testo stampato? Quando si potrà riconoscere lo stesso valore all'informazione su qualsiasi supporto, anche non tradizionale? Ad esempio immagini, suoni, dati sono potenti strumenti d'informazione ancora sottoutilizzati in biblioteca dove si raccolgono solo documenti a stampa.
L'esemplificazione potrebbe continuare ma, in sintesi, potremmo dire che, per i bibliotecari, passare da una collezione di documenti a stampa ad una collezione di documenti elettronici e multimediali comporta quattro differenze fondamentali:
— i documenti multimediali ed i documenti elettronici sono fruibili con lo stesso canale di trasmissione e quindi vanno adeguate le stazioni di ricerca e l'organizzazione del servizio;
— sono diverse le esigenze di indicizzazione dei documenti in quanto ogni testo, o parte di esso, può essere oggetto di una catalogazione originale automatica per la ricerca immediata;
— la facilità di riproducibilità del documento elettronico richiede una nuova regolamentazione della proprietà intellettuale;
— infine cade la distinzione tradizionale della catena documentaria tra chi scrive e chi legge poiché ogni documento elettronico può essere ridigitato e modificato e per questo motivo si rende necessario rivedere il ruolo dei bibliotecari come conservatori.
Ai bibliotecari quindi compete ogni sforzo per il disegno e l'organizzazione dei sistemi di fornitura dei documenti e per garantire la qualità e la coerenza nel tempo dell'informazione elettronica. Sgml diventa in questa ottica uno standard fondamentale per la biblioteca elettronica così come è stato il Marc per la biblioteca automatizzata.

Riferimenti bibliografici

BRYAN MARTIN, *Sgml, an author guide to Standard Generalized Markup Language*, Addison Wesley, 1988.

ASSOCIATION OF AMERICAN PUBLISHERS, *Electronic manuscript series standards for electronic manuscript preparation and markup*, Aap, 1987.

GOLDFARB CHARLES, *Sgml Handbook*, New York, Oxford University Press, 1990.