Paradigma Project: il deposito legale delle risorse remote nell'esperienza norvegese

Una ricerca di soluzioni per metadati e servizi agli utenti

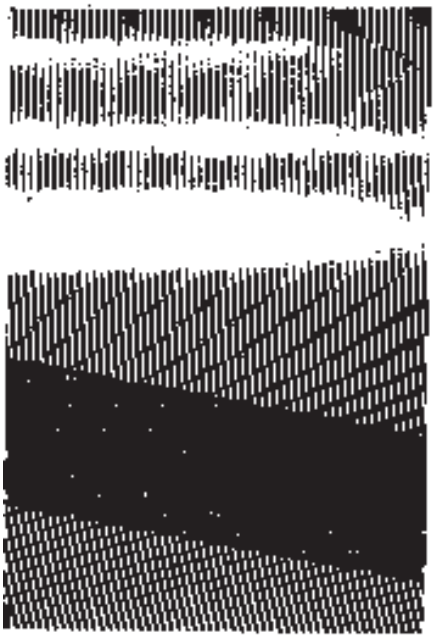
Carol van Nuys

Norwegian Archive, Library and Museum Authority
carol.vannuys@abm-utvikling.no

Ketil Albertsen
Linda Pedersen
Asborg Stenstad

The National Library of Norway
ketil.albertsen@nb.no
linda.pedersen@nb.no
asborg.stenstad@nb.no

Crediamo che sia utile ai lettori italiani proporre l'introduzione al Paradigma Project della National Library of Norway, presentata da Carol van Nuys, Ketil Albertsen, Linda Pedersen e Asborg Stenstad nell'Open Session della Cataloguing Section[1] al World Library and Information Congress: 70th IFLA General Conference and Council che si è tenuto dal 22 al 27 agosto 2004 a Buenos Aires.
L'archiviazione o la conservazione del web, ovvero della produzione digitale pubblicata in rete, è svolta sin dalla metà degli anni Novanta in Australia, Svezia e Stati Uniti[2] per la loro produzione nazionale. Il Paradigma Project, programma norvegese iniziato nel 2001, ha lo scopo di assicurare un soddisfacente deposito legale per tutti i tipi di documenti digitali, di cui vari milioni allocati sul dominio Internet norvegese. La conservazione del patrimonio culturale digitale viene svolta per consentire ai ricercatori un accesso all'archivio Internet tramite l'uso dei metadati e della ricerca a testo pieno.
Il contributo fornisce una breve descrizione del progetto e discute dei problemi che incontra nella ricerca di standard di metadati per la scoperta e la conservazione a lungo termine dei materiali; presenta l'uso dei livelli di entità FRBR opera, espressione, manifestazione e item per il design dell'archivio, nonché le idee sui servizi di verifica e di autenticazione da un lato, e di assegnazione di identificatori dall'altro, entrambi disponibili tramite Internet. Il Paradigma Project cerca di collegare l'harvesting e l'estrazione dei metadati alle disposizioni delle leggi nazionali relative al deposito legale, ponendo le basi per un accesso strutturato e amichevole al vasto patrimonio culturale digitale di rete della Norvegia, un paese in cui una percentuale elevata della popolazione utilizza abitualmente Internet.
In molti paesi, la normativa sul deposito, estesa a ogni tipologia di materiale documentario, costituisce solo una dichiarazione d'intenti nel caso del web, dato l'incontrollabile volume delle risorse remote, la complessità delle attività di manutenzione, la necessità di un'organizzazione dinamica degli strumenti di accesso alla base dati e la complessa articolazione della piattaforma tecnologica che l'impegno del controllo richiede di approntare.
In Italia, dopo la pubblicazione della legge 106 del 15 aprile 2004 Norme relative al deposito legale dei documenti di interesse culturale destinati all'uso pubblico, *la direzione della Biblioteca nazionale centrale di Firenze, in seguito a "numerose richieste d'informazioni su come depositare i siti web", ha diffuso il 20 maggio 2004 il comunicato "Archiviazione dei siti web"[3] con cui informa che "le biblioteche nazionali stanno cooperando a livello internazionale e [...] concordemente indicano nell'*harvesting *– ossia nella raccolta*



*delle pagine web effettuata tramite un software (*crawler*) – la modalità più efficiente e sostenibile di deposito. In pratica con questa tecnologia chi pubblica siti web liberamente accessibili in rete non deve 'depositare' assolutamente niente: è il* crawler *gestito dall'istituzione depositaria che provvede a 'raccogliere' il sito web (per maggiori informazioni su archiviazione del web e cooperazione internazionale: <http://netpreserve.org>). In ogni caso nell'attesa dell'emanazione del regolamento applicativo si raccomanda di non inviare (su cd o via posta elettronica) copie di siti web alla nostra biblioteca".*
Abbiamo stilato un Glossario per facilitare la comprensione di alcuni termini di conio recente e corrediamo la presentazione del Paradigma Project di alcune immagini tratte dal suo sito web presso la National Library of Norway.
(s.g., m.g.)

***

1. Introduzione

1.1. L'archiviazione del web in altri paesi

I documenti digitali stanno scomparendo ogni giorno e uno studio[4] mostra che solo il 20% dei documenti che troviamo sulla rete vi rimane – immutato nei contenuti – dopo un anno. Di conseguenza andranno scomparendo, per le nuove generazioni dei lettori, anche le possibilità di studiare i documenti digitali odierni. La conservazione del nostro patrimonio culturale digitale è un problema di crescente importanza e insieme una sfida, e la National Library of Norway[5] è una delle molte istituzioni che operano, in modo sistematico, per trovare risposte ai problemi legali, tecnici e bibliografici che l'accompagnano. Un aspetto del lavoro di conservazione digitale consiste nel raccogliere e archiviare documenti provenienti dai domini Internet nazionali. Diversi paesi hanno scelto differenti strategie di raccolta: la Danimarca [1] e l'Australia [2] hanno intrapreso l'approccio selettivo, mentre la Svezia [3], l'Islanda e la Finlandia hanno raccolto (*harvesting*) il loro intero spazio web nazionale. La Norvegia appartiene a un piccolo gruppo di biblioteche in Europa che stanno realizzando l'*harvesting* e l'archiviazione dei documenti digitali dai propri domini Internet nazionali basandosi sulla legislazione relativa al deposito legale.[6]

1.2. Il Legal Deposit Act

Lo scopo del Legal Deposit Act [5] è di "garantire che i documenti che contengono informazione pubblicamente disponibile siano depositati in collezioni nazionali, affinché questi record della vita sociale e culturale norvegese possano essere conservati e resi disponibili in quanto fonti per finalità di ricerca e documentazione" (§; 1). Ritenuto estremamente moderno quando apparve nel 1989, l'attuale Legal Deposit Act considera tutti i documenti norvegesi *pubblicamente disponibili* memorizzati su *qualsiasi* supporto: ad esempio carta, microforme, fotografie, documenti su più supporti, registrazioni sonore, film, video, documenti digitali e programmi radio o teletrasmessi. Sono rappresentati anche i documenti pubblicati all'estero da o per editori norvegesi e quelli adattati specialmente per il pubblico norvegese. Naturalmente nel 1989 il World Wide Web non aveva fatto ancora la sua comparsa. I documenti digitali – la gran parte dei quali aveva la forma delle basi dati – erano scarsi in rapporto ai milioni di odierne pubblicazioni sulla rete Internet, e persino difficili da trattare dal punto di vista tecnico. Oggi il National Library's Long-Term Preservation Repository ha la capacità di immagazzinare 100 TBytes di dati, l'equivalente di un numero molto grande di documenti digitali.

2. Il Paradigma Project

La National Library of Norway varò il Paradigma Project[7] nell'agosto del 2001. Suo obiettivo era ed è garantire un deposito legale soddisfacente dei documenti digitali norvegesi; ciò include lo sviluppo di tecnologia, metodologia e routine per la selezione, la raccolta, la descrizione e l'identificazione di *tutti* i tipi di documenti digitali – compresi i documenti pubblicamente disponibili sulla rete Internet. Il progetto sta anche occupandosi di fornire agli utenti accesso al proprio archivio Internet in ottemperanza all'attuale legislazione. Le attività del progetto, per le quali quattro operatori sono impiegati a tempo pieno, affondano le radici sul precedente lavoro svolto dalla biblioteca in varie sfere d'interesse e in settori importanti. Sono infatti coinvolti anche una trentina di altre persone in lavori complementari del progetto, il cui termine era fissato per il 31 dicembre 2004.
I paragrafi seguenti descriveranno sinteticamente gli aspetti principali del progetto nella selezione, raccolta e accesso a materiali digitali acquisiti tramite deposito legale dalla rete Internet, e insieme la natura e la dimensione del dominio Internet norvegese.

2.1. Strategie di raccolta e selezione

2.1.1. Raccolta

Sulla base del Legal Deposit Act e delle raccomandazioni del Paradigma Project, la National Library

Glossario

Batch. L'operazione condotta su più file contemporaneamente; oppure la tecnica di elaborazione da parte di un sistema operativo che organizza i lavori in sequenza e li esegue a lotti.

Crawler o spider, ragno, robot. Software, parte del motore di ricerca, che naviga il web, memorizza gli indirizzi e indicizza parole chiave e testo delle pagine che incontra. Con *spidering* o *crawling* si intende l'attività svolta dal software di scaricare documenti web seguendo i collegamenti ipertestuali. Poiché è impossibile effettuare lo *spidering* di tutte le pagine della rete, i motori di ricerca esercitano alcuni compromessi per indicizzare il maggior numero di risorse web (indicizzazione limitata alle parole del titolo, o ai livelli più alti del sito). Alcuni spider evitano di indicizzare termini comuni (articoli, congiunzioni), che sono inclusi in elenchi di *stop word*. Vi sono differenti metodi di *crawl*: *focused* (ristretto da parametri quali il tipo di file, la dimensione o localizzazione), *smart crawl* (ristretto in base a criteri dinamici, ad es. la valutazione di un sito in base alla struttura e al contenuto, o alla presenza di metadati; più il software setaccia la rete, più impara a essere selettivo), *incremental* (aggiorna un *crawl* precedente, scaricando solo le pagine web che hanno subito modifiche, aggiunte o cancellazione di dati) e *customized* (ottimizzato per visitare siti web particolari in base alla conoscenza umana della loro struttura e del loro contenuto).

Crosswalk. Tabella di comparazione delle relazioni ed equivalenze tra due o più formati di metadati. Le tabelle sono dette anche di conversione, poiché supportano la capacità dei motori di ricerca di interrogare con efficacia diverse basi dati eterogenee; i *crosswalks* aiutano dunque a promuovere l'interoperabilità.

Deep web. Web profondo o invisibile; risorse web pubblicamente accessibili, ma non incluse negli archivi dei motori di ricerca. Sono compresi materiali difficili o impossibili da indicizzare, ad es. basi dati, alcuni formati di file, pagine successive a quella in cui si richiede un'autenticazione, pagine dinamiche.

Harvester. Software che realizza la funzione di *harvesting*, ossia la "mietitura" o raccolta automatica, in un unico deposito digitale, dei metadati presenti nelle basi dati remote. I metadati descrivono le risorse informative disponibili ai relativi indirizzi. L'*harvesting* può riguardare anche i dati: per *full-text harvest* si intende una raccolta di testo volta a costruire un indice a pieno testo con collegamenti alle risorse; per *full-content harvest* si intende una raccolta che registra parti estratte dalle pagine per presentare, all'interno di un indice a pieno testo, i termini di ricerca nel loro contesto.

Snapshot. Salvataggo o istantanea di un sito web. Il termine è utilizzato anche per i programmi di grafica, a indicare una fase di lavoro che si vuole "congelare" nel caso in cui le variazioni operate successivamente non siano soddisfacenti. Oltre alle versioni alternative di un'immagine, *snapshot* indica le istantanee di uno schermo o di una sequenza in movimento.

Streaming media. Tecnologie che consentono, tramite Internet, una trasmissione continua e crescente del flusso di informazione in formato audio o video. L'utente ascolta o visiona il file contemporaneamente al suo "passaggio": le immagini o i suoni, compressi, sono inviati dal server in successione e immediatamente decompressi e visualizzati sul client. La tecnologia permette la visione in diretta di eventi. *(s.g., m.g.)*

of Norway ha deciso di iniziare l'*harvesting* generale di *tutti i documenti digitali pubblicamente disponibili* nello spazio web norvegese (".no"). In futuro anche i documenti reperiti su domini come ".com", ".org" e ".net" saranno oggetto di *harvesting*.
Ci sono diverse ragioni per adottare questo approccio a un *harvesting* di carattere generale. In primo luogo, non possiamo prevedere quali documenti saranno di valore nella ricerca e documentazione futura; in secondo luogo l'archiviazione digitale sta divenendo ogni giorno più economica; in terzo luogo, un *harvesting* senza filtri permette di risparmiare una selezione manuale costosa in termini di risorse al momento della raccolta, infine l'utente di un archivio Internet può ricuperare i documenti tramite utilità di ricerca a testo libero, potendo così passare in rassegna tutti i documenti, compresi quelli che non meritano una catalogazione manuale. Criteri di selezione per un qualsiasi uso, come per una successiva descrizione bibliografica, possono anche essere messi alla prova e modificati in qualsiasi momento. Naturalmente questo sarebbe impossibile se il materiale fosse escluso al momento dell'*harvesting*.
A partire dal 2001 la Legal Deposit Section ha indicizzato (*harvesting*) in modo semi-automatico una selezione di documenti web, usando il software HTTrack;[8] i documenti sono descritti nel catalogo della National Library of Norway (BIBSYS).[9] Il lavoro continuerà sino al momento in cui l'attività di *harvesting* generalizzato del Paradigma Project e procedure collegate non saranno pienamente opera-

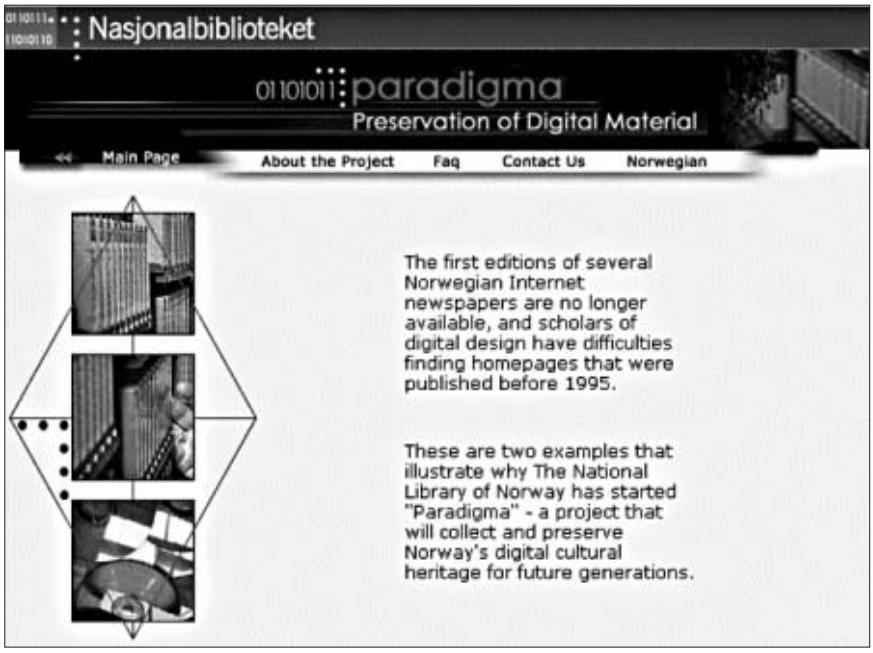


Fig. 1 – Home page del Paradigma Project nel sito della National Library of Norway

tive. La stessa sezione si occupa anche delle collezioni di risorse basate sugli eventi e ha raccolto siti web di partiti politici, prima, durante e dopo le elezioni. Anche altre sezioni sono impegnate in attività di deposito legale digitale, e la Library's Sound and Image Archive sta lavorando per trovare soluzioni al deposito legale di programmi radio e televisivi nati in formato digitale in cooperazione con la Norwegian Broadcasting Corporation.
Un problema estremamente stimolante e impegnativo è il deposito del *deep web*, ad esempio quotidiani Internet (*streaming media*), documenti prodotti da camere web, media interattivi e materiali in formato elettronico di qualsiasi tipo che sono archiviati nelle basi dati. Il Paradigma Project ha iniziato la raccolta quotidiana di circa 65 giornali Internet, e compirà quanto prima il download di intere basi dati di giornali, integrando così le istantanee (*snapshots*) quotidiane. Stiamo discutendo i problemi del *deep web* nel contesto dell'International Internet Preservation Consortium,[10] ma un vasto numero di problemi amministrativi, giuridici e tecnici sono ancora irrisolti.
In sintesi, la National Library of Norway può attendersi di ricevere oggetti digitali tramite diversi canali: l'*harvesting* automatico dei documenti dalla rete Internet, gli aggiornamenti delle basi dati acquisite con procedure *batch*, le sottoscrizioni a periodici e a mailing list ricevute tramite e-mail, gruppi di discussione NetNews e documenti distribuiti su supporti fisici quali cd-rom.

2.1.2. Selezione

Ci sono molti documenti di valore reperibili sulla rete Internet, e stiamo lavorando alla definizione di *criteri di selezione* per quei documenti che riteniamo "meritino" una descrizione bibliografica manuale. I criteri di selezione sono basati sulla legislazione relativa al deposito legale e sulle politiche generali di raccolta della Biblioteca come vengono formulate nel nostro piano strategico. I criteri di selezione per i documenti digitali sono stati integrati con quelli riguardanti tipi più tradizionali di documenti nel *Library's selection manual*.
Il Paradigma Project ha intenzione di implementare un'architettura di sistema che consenta un procedimento di *selezione* a tre fasi, affinché i bibliotecari ricevano aiuto tecnico per trovare quei pochi documenti che dovrebbero essere catalogati. Nella prima fase vengono ricuperati e raccolti da Internet i documenti norvegesi e Sami. Il secondo stadio permette ai bibliotecari di produrre *automaticamente* elenchi ordinati basati su ricerche specifiche. Gli elenchi sono basati sull'uso di vettori contenenti metadati che sono stati automaticamente estratti dai documenti raccolti. Nella terza fase, i bibliotecari scelgono documenti specifici dagli elenchi ordinati per una registrazione manuale, con l'uso dei criteri di selezione di cui abbiamo detto. Un giorno potremo essere in grado di monitorare risorse integranti (*integrating resources*) che sono state catalogate manualmente, assistendo così i bibliotecari a ricuperare e modificare queste registrazioni bibliografiche, ad esempio a determinati intervalli di tempo, quando mutamenti testuali eccedono una data percentuale.

2.2. Il dominio Internet norvegese

L'esatta dimensione del dominio Internet norvegese è ancora sconosciuta. Il primo ciclo di *harvesting* del Paradigma Project, nel periodo dicembre 2002 – gennaio 2003, ha prodotto circa 3,1 milioni di URL (ossia file), di cui circa il 53% (per conteggio) è costituito da immagini (.jpg, .gif, .png). L'*harvester* NEDLIB[11] iniziò con circa mille URL e l'*harvesting* fu limitato al protocollo HTTP, al dominio nazionale norvegese (".no"), e agli URL senza parametri. La seconda tornata di *harvesting* fu

condotta nell'agosto 2003, e ha prodotto circa 4,1 milioni di URL. Della terza, in corso, non disponiamo al momento di alcuna statistica. Se assumiamo una distribuzione simile a quella che si può osservare nelle serie di *harvesting* condotte in Svezia e Finlandia, ci aspettiamo di trovare il 45-55% dei siti Internet norvegesi in domini esterni a ".no". Va da sé che è impossibile il trattamento manuale e la valutazione di ogni oggetto; la stragrande maggioranza di essi deve essere elaborata automaticamente.

2.3. Strategia di accesso

2.3.1. *Chi* cercherà *che cosa* nel nostro archivio?

Quando cerchiamo di identificare delle soluzioni per i metadati che descrivono il ricco e vario materiale digitale conservato nel nostro archivio è importante chiederci: chi userà il materiale e per quali finalità? È difficile immaginare le interrogazioni specifiche del ricercatore dei prossimi dieci, venti o cinquanta anni, ma possiamo tentare di immaginare qualche *gruppo di utenti* e *tipi* di domande.
Un gruppo può consistere in persone interessate allo studio di Internet e dei materiali digitali in quanto *medium*, ossia al motivo per cui il materiale è stato raccolto da Internet, e in quanto mostra le caratteristiche legate a questo mezzo. Possiamo osservare che alcuni utenti potrebbero aver bisogno di studiare l'uso del linguaggio sulla rete e il sistema delle relazioni tra differenti forme di linguaggio; i ricercatori dei mezzi di comunicazione potrebbero voler studiare i rapporti tra media a stampa e media digitali o tra linee di sviluppo tecnologico e contenuti; gli utenti che studiano il design delle pagine web potrebbero essere interessati all'uso della pubblicità, del layout ecc.; ricercatori nel campo della *computer science* potrebbero studiare i diversi protocolli di comunicazione, l'uso dei formati attraverso il tempo o persino dei virus che hanno attaccato i dati; gli scienziati sociali potrebbero essere interessati al modo in cui l'informazione disponibile su Internet ha influenzato la società e viceversa. Possiamo inoltre attenderci di trovare ricercatori con aree di interesse che si sovrappongono.
Un altro gruppo di utenti potrebbe essere formato da quelli che hanno bisogno di usare i documenti digitali come *source material* allo stesso modo in cui oggi si impiegano le fonti tradizionali. Questo gruppo consisterà principalmente di ricercatori che appartengono a tutti i campi di studio e sarebbe perciò interessante scoprire le loro aspettative, in particolare riguardo ai materiali digitali. Il materiale rilevante è disponibile solo in formato digitale? Sono giudicati importanti contenuto dinamico, animazioni, display interattivi, suono e video integrati ecc.? I ricercatori necessitano di accedere al materiale tramite una ricerca a testo pieno o di correlare grandi quantità d'informazione proveniente da fonti differenti?

2.3.2. L'attuale legislazione

Fornire agli utenti accesso all'archivio del deposito legale di Internet è una questione complessa e la National Library of Norway deve trovare soluzioni soddisfacenti a dispetto delle numerose regole, in alcuni casi contraddittorie, che troviamo nel Legal Deposit Act, nel Copyright Act e nel Personal Data Act. Stiamo tentando di trovare risposta a domande come le seguenti: quali utenti possono ricevere accesso ai differenti tipi di materiali digitali? Quali possono accedere alle raccolte dai computer che sono all'esterno della National Library of Norway?

2.3.3. Strumenti di accesso

Le caratteristiche degli utenti descritte in precedenza ci interessano, dal momento che stiamo tentando proprio di sviluppare strumenti d'accesso per la ricerca nel nostro archivio Internet. Dobbiamo, naturalmente, prendere in considerazione il fatto che i bibliotecari cataloghereanno una piccola quantità dei documenti disponibili. Su un piano tecnico, ci si augura che il Paradigma Project fornisca agli utenti accesso all'archivio Internet tramite il Nordic Web Archive's[12] (NWA) Access Tool (vedi figura 2, p. 26). Oggi sono opzioni standard la ricerca a testo libero con operatori booleani, la ricerca di un particolare URL e la presentazione della storia del documento tramite una linea del tempo (*timeline*). Ci auguriamo che lo strumento fornirà sempre maggiori possibilità: uso di combinazioni di ricerca booleana per integrare elenchi differenti di record, ricerca parallela sui documenti descritti in cataloghi esterni, ricerca su metadati estratti automaticamente, opzioni che ci permettano di salvare i risultati in una "biblioteca di progetto", accesso a gruppi di documenti ordinati in base a vari criteri (editore ecc.), restrizione dei risultati a un solo record in caso di presenza di duplicati, raggruppamento di un documento logico che consiste di numerose pagine web separate in un solo record nell'elenco dei risultati. È previsto l'adattamento dell'interfaccia del NWA Access Tool per soddisfare varie funzioni speciali dell'utente e il nostro uso del modello FRBR giocherà un ruolo importante sulle modalità con cui forniremo accesso al materiale archiviato.

3. Una ricerca per soluzioni di metadati

Il Paradigma Project è nel pieno della sua ricerca per identificare formati di metadati adatti e soluzioni ad essi relative. La definizione di metadati per la *ricerca* è stata tra le nostre principali attività nell'anno passato, insieme ai nostri tentativi per trovare soluzioni soddisfacenti nell'estrazione automatica dei metadati. Nel paragrafo che segue cercheremo di dare un'idea del *perché* e del *come* progettiamo di descrivere quella massa di documenti digitali che risiedono nel nostro archivio Internet.

3.1. Perché dovremmo catalogare le risorse Internet?

Nell'introduzione al suo libro *Cataloging Internet resources* [3] Nancy Olson[13] propone tre ragioni basilari per cui le risorse Internet dovrebbero essere catalogate:
1) esiste una grande quantità d'informazione di valore disponibile tramite Internet;
2) le risorse richiedono di essere organizzate per una loro accessibilità;
3) l'uso delle tecniche e delle procedure biblioteconomiche e la creazione di record per il loro ricupero tramite i cataloghi online costituiscono il metodo più efficiente per accedere a tali risorse.
Siamo d'accordo con Olson su tutti e tre i punti, ma allo stesso tempo stimiamo che *molto meno dell'1%* del materiale raccolto dal dominio Internet norvegese possa diventare oggetto di una registrazione bibliografica. Ciò dipende dalla grandezza dell'archivio. Sebbene una percentuale molto elevata dei materiali tradizionali della Biblioteca sia soggetta a registrazione bibliografica, materiali differenti sono trattati in modi diversi: agli *ephemerais* si dà una registrazione semplice, mentre ai libri e ai periodici viene attributo un livello più elevato di catalogazione. Per contrasto, il 100% dei documenti Internet sarà completamente indicizzato con FAST,[14] un software di indicizzazione che opera dopo l'*harvesting*. Ciò permetterà allo staff della Biblioteca e agli utenti di ricercare nell'archivio Internet con una ricerca a testo libero e con altri indici. La minuscola frazione dei documenti catalogati manualmente sarà disponibile tramite la ricerca a pieno testo nell'archivio e tramite le registrazioni bibliografiche nel catalogo – possibilmente collegate tra loro in qualche modalità amichevole. Oltre alla catalogazione di alcuni documenti e all'indicizzazione di tutti, raccoglieremo con l'*harvesting* i metadati incorporati nella risorsa insieme ai documenti Internet che essi descrivono, e la National Library of Norway sta progettando un servizio che permetterà agli editori di generare e distribuire i metadati insieme ai relativi documenti al momento del deposito.

3.2. Cosa sono i metadati?

Una ricerca di soluzioni relative ai metadati ci ha, naturalmente, condotti a una ricerca di adeguate definizioni. Il termine "metadati" è stato definito varie volte in letteratura. "Dati relativi ai dati" è forse la definizione più ricorrente, e i metadati comprendono un'intera gamma di tipi di informazione.[15] Abbiamo scoperto che gli schemi di metadati sono numerosi quanto diversi, ma che hanno una cosa in comune: possono aiutarci a *descrivere* e *trovare* i numerosi documenti di valore nella nostra raccolta, compresi quelli che non sono candidati per una catalogazione di alto livello.

3.3. Cos'è un documento Internet?

3.3.1. Definizione di un documento Internet da un punto di vista tecnico

Quando un documento Internet viene selezionato per l'*harvesting* e quindi per la conservazione, la semantica di un "documento" può essere altamente ambigua: quali componenti dovrebbero essere sottoposti a *harvesting* e archiviati come parti integrali del documento? Quali componenti dovrebbero essere soggetti a una valutazione individuale? Riteniamo che qualsiasi componente che influenzi l'aspetto di una pagina web (compresi *suono* e altri elementi *non grafici*) dovrebbe essere incluso senza condizioni se una pagina web viene selezionata, ad esempio immagini di sfondo, contenuti dei frame, immagini per bottoni ecc. I documenti cui si fa riferimento tramite i collegamenti ipertestuali si distinguono, pur essendo loro connessi, dai documenti che li citano. A un livello semantico più elevato, spesso desideriamo considerare un gruppo intero di documenti collegati l'un l'altro come un unico, ampio documento. Se li valutiamo come documenti completamente indipendenti, corriamo il rischio di realizzare un *harvesting* di pochi capitoli in una relazione, escludendo gli altri (ciò potrebbe avvenire poiché contengono citazioni ampie, riepiloghi ecc. in lingue diverse dal norvegese). Dunque, per rispondere alla nostra domanda "In che consiste un documento Internet?", possiamo dire che un documento Internet è composto di numerose parti correlate o file, ad esempio testo, immagini, suono, animazioni ecc., e che esse sono connesse molto spesso tramite link e a volte contenute in insiemi di frame.

3.3.2. Definizione di un documento Internet da un punto di vista bibliografico

Non possiamo naturalmente fare mai affidamento sul computer per sapere dove inizia e dove finisce un documento Internet – anche se lo avessimo programmato per seguire determinate istruzioni con questo obiettivo. Fortunatamente i bibliotecari sono molto esperti nel decidere quali parti (delle molteplici di un documento Internet) costituiscono un insieme dal punto di vista logico. Così, da un punto di vista bibliografico, possiamo definire un documento Internet come un'unità d'informazione che può essere descritta bibliograficamente. Questa definizione *non* specifica intenzionalmente un insieme di componenti definite o uniche del documento, ma lascia al bibliotecario il compito d'identificare l'oggetto da descrivere: un intero sito web può essere descritto con un record e a una particolare risorsa di quel sito può essere assegnata una descrizione. Il bibliotecario può includere o omettere suoni di sottofondo, fogli di stile ecc. e può raccogliere in un documento varie pagine web strettamente collegate, ad esempio i paragrafi di una relazione. Le nostre procedure automatiche suggeriranno al bibliotecario definizioni del documento basate sull'analisi del contenuto, insiemi di link ecc.; per default, immagini incorporate, clip audio-video direttamente collegati e fogli di stile vengono inclusi nel documento. Sono inclusi anche collegamenti particolari che identificano una pagina web referenziata, per esempio un sommario o una sezione. Di conseguenza un'unità informativa che può essere bibliograficamente descritta è il punto di partenza per la formulazione di una descrizione in termini di metadati, nel caso in cui il materiale digitale sia depositato su supporti fisici quali cd-rom o dvd, e nel caso in cui sia raccolto dalla rete Internet in termini di file separati. Ciò significa che tutti i documenti digitali – dai tipi *tradizionali* di documento quali monografie, tesi ecc., a quelli *effimeri* quali i quotidiani Internet, hyper poetry, hyper drama ecc. e ai *nuovi* tipi di documento quali homepage, web blog ecc. – sono i candidati per una descrizione tramite metadati nell'ambito del nostro archivio Internet.

3.4. Un'indagine sui metadati

3.4.1. Di quali tipi di metadati abbiamo bisogno?

Abbiamo considerato interessante chiedere quali formati di metadati per la descrizione di diversi tipi di materiali digitali siano oggi in uso presso la National Library of Norway. Questa informazione può essere utile, poiché si spera di essere in grado di esportare e importare i dati nel nostro archivio. I risultati della nostra indagine mostrano che sono in uso vari formati: BIBSYS-MARC (il formato MARC del sistema BIBSYS) per i testi digitali, Dublin Core Metadata Element Set[16] per programmi radio, MAVIS[17] (sistema e formato australiano) per il materiale cinematografico e televisivo (*broadcasting*), suoni e immagini, e inoltre altri formati usati in sistemi locali. I formati di metadati sono adatti, ma non costituiscono soluzioni soddisfacenti per tutte le nostre esigenze relative ai metadati. L'archivio Internet richiede molti tipi di metadati:
– *amministrativi*, riguardano ad esempio la creazione e modifica dei record di metadati;
– di *gestione dei diritti e dell'accesso,* per registrare informazione sul copyright e definire quali insiemi di utenti possano ottenere accesso all'archivio e quali documenti possano leggere;
– *strutturali*, per evidenziare le relazioni logiche tra oggetti, tra metadati o tra oggetti e metadati;
– *per la conservazione a lungo termine*, per la specifica, ad esempio, dei tipi di file, del software necessario e della storia della conversione (o migrazione) del documento;
– *tecnici*, per esprimere la dimensione dei documenti, script, dettagli sulla comunicazione ecc.;
– ultimo, ma non da meno, abbiamo la necessità di metadati *descrittivi* e *analitici,* per la ricerca e il ricupero.

3.4.2. Quale modello descrittivo dovremmo scegliere?

Vi sono opinioni diverse circa il livello di descrizione che un documento digitale dovrebbe ricevere. Nel nostro lavoro, volto a definire le caratteristiche dei metadati descrittivi e analitici, abbiamo preso in considerazione due modelli alternativi. Il primo prevede l'uso di tre livelli descrittivi:
1) catalogazione per inclusione nella Bibliografia nazionale e nel catalogo BIBSYS della National Library of Norway e in altre basi dati;
2) catalogazione a un livello più semplice in un formato comune;
3) estrazione automatica di metadati dal documento, come da protocolli di comunicazione ecc.
L'alternativa consiste nell'uso di un metodo a due livelli, ad esempio "catalogare o non catalogare":
1) catalogazione per inclusione nella Bibliografia nazionale e nel catalogo BIBSYS della National Library of Norway e in altre basi dati speciali;
2) estrazione automatica di metadati dal documento, come da protocolli di comunicazione ecc.
Esistono diverse ragioni per la seconda soluzione:
a) il ricupero di materiale digitale (a testo pieno e libero ecc.) non è dipendente dalla registrazione come nel caso di materiale analogico non registrato;
b) non è necessario per la biblio-

teca registrare materiale per tener traccia dei suoi aspetti logistici e amministrativi, ad esempio quali biblioteche universitarie abbiano ricevuto le copie;
c) possiamo sempre rimpiangere la nostra decisione di non catalogare un certo tipo di materiale digitale.
Descriviamo brevemente ciascuno dei tre livelli.

a) *Catalogare ai fini dell'inclusione nella Bibliografia nazionale ecc.*
Le nostre proposte attuali su quali tipi di documenti debbano essere catalogati a livello più elevato sono incomplete, ma si può dire con certezza che un piccolo numero di documenti digitali di valore continuerà a essere catalogato in qualche formato MARC per una loro inclusione nella Bibliografia nazionale.[18] Il codice di catalogazione norvegese è basato sulle *Regole di catalogazione angloamericane* (AACR2), di cui sono disponibili in norvegese i capitoli 9 e 12. La catalogazione dei materiali audiovisivi per la conservazione a lungo termine richiede un elevato livello di dettaglio, specialmente quando giunge a dar conto dell'informazione tecnica connessa al restauro degli originali, delle copie ecc. La Biblioteca continuerà senza dubbio a usare MAVIS per questo lavoro.

b) Catalogazione a un livello più semplice in un formato comune
Come si è detto, la National Library of Norway sta progettando un servizio che permetterà agli editori di produrre e distribuire metadati insieme ai loro documenti al momento del deposito legale. Oggi il Paradigma Project sta operando per definire i formati di metadati che costituiranno le fondamenta di uno strumento amichevole. I bibliotecari potranno trattare i record di metadati forniti dagli editori, utilizzandoli quale base per registrazioni bibliografiche di più alto livello. Abbiamo preso in esame e comparato alcuni formati di metadati nella nostra ricerca di soluzioni adeguate: MAchine Readable Cataloguing (MARC) e Dublin Core Metadata Element Set (DCMES), poiché sono usati nelle biblioteche e in istituzioni simili; Metadata Object Description Schema (MODS)[19] e Metadata Encoding & Transmission Standard (METS),[20] sviluppati da biblioteche per la comunità bibliotecaria e Online InformationeXchange (ONIX),[21] formato implementato nell'ambito dell'editoria e delle industrie del libro. Si noti inoltre che la comunità ISBN ha suggerito che gli editori possono fornire alle agenzie ISBN dei metadati ONIX compatibili, in coincidenza con l'assegnazione di ogni ISBN. Abbiamo confrontato questi formati chiedendoci: "Chi è responsabile della gestione del formato? È uno standard internazionale? In quale area è usato? Quali tipi di media descrive? Include definizioni semantiche e/o sintattiche? Come descrive le relazioni tra i documenti? Il formato è dipendente da regole o codici specifici? È compatibile o collegato ad altri formati? Quanto è diffuso e da quali comunità è usato?".
Ci auguriamo che questa indagine possa condurre a una discussione più ampia sui metadati in biblioteca con riferimento al loro esame in corso. È nostra intenzione prendere anche in considerazione attentamente come sia possibile soddisfare i requisiti funzionali del nostro utente adottando il *Common core records* proposto dall'IFLA Working Group on the Use of Metadata Schema [6] e Functional Requirements for Bibliographic Records (FRBR) dell'IFLA [5]. Sul tema delle soluzioni relative ai metadati sono previste inoltre collaborazioni con un progetto bibliografico in corso all'interno della National Library of Norway e con un altro progetto a livello nazionale: la Norwegian Digital Library. Speriamo che questo lavoro conduca a suggerire i formati di metadati da impiegare nella descrizione ai differenti livelli.
Allo stesso tempo, abbiamo lavorato per specificare i requisiti dei metadati tecnici per il nostro software di sistema d'archivio, identificando vari fattori che possono condizionare la scelta di formati di metadata per una catalogazione di livello più basso; ne riportiamo alcuni tecnicamente desiderabili.
– *Interoperabilità semantica con MARC*: è importante che le proprietà dei formati di metadati siano semanticamente armonizzate con il formato MARC, dominante all'interno della comunità bibliotecaria. Laddove possibile, il formato dovrebbe essere un sottoinsieme funzionale del MARC. Ciò faciliterebbe lo scambio dei dati.
– *Semplice ma ricco*: è importante identificare un formato di metadati che sia semplice da usare e insieme abbastanza ricco per rappresentare un'adeguata quantità di dettaglio.
– *Facilità di conversione in altri formati*: una tabella di crosswalk di conversione tra il formato scelto e il MARC dovrebbe essere disponibile o relativamente facile da definire. Osserviamo che già esistono crosswalk tra MARC21 e MODS (e tra MARC21 e ONIX), come pure crosswalk tra il Dublin Core non qualificato e MODS.
– *Compatibilità con XML*: XML è uno standard *de facto* e un formato XML compatibile; ci permetterà di gestire il formato con il software disponibile. Sarà anche definita in XML una più ampia cornice strutturale, permettendo all'archivio di accettare metadati da fonti diverse, gestire modifiche dei metadati, definire metadati originali, tener traccia della storia delle versioni ecc. (ad esempio METS).

– *Estensibilità*: un formato di metadati dovrebbe permetterci, se necessario, di definire nuovi elementi.
– *Elementi di base* (core elements): è importante definire elementi di metadati di base, ad esempio un denominatore comune che può facilitare la ricerca e il ricupero dei documenti tra differenti tipi di materiali.
Se raffrontiamo questi fattori con i formati di metadati descritti nella nostra indagine, osserviamo che sono preferibili i formati MARC e XML compatibili. Ad ogni modo, non esiste una ricetta semplice. Nuovi elementi per i metadati tecnici, strutturali e di gestione dei diritti e dell'accesso devono essere definiti, e forse consolidati all'interno dell'ambiente METS. Lo stesso è vero, naturalmente, in relazione ai metadati per la conservazione a lungo termine. A questo fine il Library's Long-Term Digital Repository ci richiede di usare metadati compatibili OAIS.[22]

c) Un'estrazione automatica di metadati
Purtroppo i bibliotecari non cataloghereranno mai un'incredibile percentuale dei documenti Internet: il 99% del nostro archivio. Questa è la ragione delle indagini sull'estrazione automatica di metadati dai documenti Internet come parte del nostro lavoro con i metadati e il design del sistema. I metadati estratti saranno archiviati insieme agli oggetti digitali e ad altre descrizioni di metadati e resi disponibili per la ricerca strutturata nell'archivio Internet. La tecnologia non è ancora abbastanza evoluta per decidere in modo automatico qual è il tipo di documento, ma può aiutare a ridurre la quantità di documenti cui si presta attenzione nella seconda fase del nostro lavoro di selezione. Esempi di proprietà del tipo di documento sono:
1) lingua, vocabolario e grammatica;
2) dimensione e struttura del documento;
3) fonte, editore, web-server;
4) uso dei *cookies*;
5) età e aspettative di vita di un documento;
6) suono, immagini, animazioni, video e tipi evoluti d'informazione;
7) interazione dell'utente tramite "moduli", bottoni ecc.;
8) numero, tipo e origine dei link;
9) valori dell'URL, ad esempio uso di caratteri o parole speciali nell'URL;
10) uso di script lato client;
11) dettagli tecnici sulla comunicazione.
La tecnologia per l'analisi del vocabolario e della grammatica sta migliorando, e abbiamo la sensazione che questo tipo di analisi possa essere un elemento importante nelle future procedure automatiche. Infine, proprietà tipologiche definite automaticamente saranno rese disponibili per una ricerca strutturata nell'archivio Internet. Il valore di queste proprietà sarà limitato, ma in combinazione con altri criteri di ricerca potrà rivelarsi positivo.

4. Il ruolo di FRBR nell'archivio Internet

Il Paradigma Project desidera presentare i documenti digitali archiviati e i metadati in una modalità strutturata e organizzata, facilitando così la navigazione dell'utente. Abbiamo considerato il modello FRBR dell'IFLA uno strumento essenziale in questo lavoro e useremo il modello per le fondamenta del design dell'archivio Internet. Riteniamo che aggiungere meccanismi complementari al modello FRBR darà in continuazione un beneficio al nostro lavoro con media dinamici quali i documenti Internet, i multimediali e altre risorse. Meccanismi complementari possono essere aggiunti come semplici estensioni del modello, non richiedendo mutamenti significativi agli attuali concetti di FRBR.[23] Per adattare il modello FRBR ai documenti Internet dinamici, è richiesta una moderata reinterpretazione dei concetti di *manifestazione* e *item* descritti nel paragrafo seguente.

4.1. Adattamenti di FRBR per il loro uso con documenti Internet dinamici

4.1.1. Documenti dinamici

I documenti Internet sono spesso *dinamici*, ad esempio un quotidiano Internet, aggiornato molte volte nell'arco di una giornata. Un utente può riferirsi a questo tipo di documento dinamico come a un forum o canale informativo: "'The Daily News' riferisce che...". Possiamo forse dire che un documento dinamico corrisponde all'incirca a un URL. I concetti di "stampe" e di successive "edizioni" devono essere ripensati anche nel contesto di Internet: da un punto di vista formale, l'aggiornamento di una pagina web può essere simile all'edizione di un nuovo libro. Tuttavia, i lettori vedono la pagina iniziale, in continua mutazione, di un quotidiano Internet come un'entità singola, variabile, non come una serie di edizioni distinte, separate. Usando il modello FRBR con le estensioni per il materiale complementare, abbiamo definito il concetto di documento dinamico come "l'intero ciclo di vita di una pagina web, o di un analogo documento Internet che muta in continuazione". Se dovessimo catalogare un documento web aggiornato di questo tipo con le AACR2 useremmo le regole per le risorse integranti, ossia quelle risorse bibliografiche che sono aggiunte a, o che mutano per mezzo di, aggiornamenti che non riman-

gono discreti e sono integrati nel tutto, nell'insieme. I documenti quali i quotidiani Internet sono tuttavia simili a un canale radio, un flusso continuamente mutevole d'informazione transitoria. Non si "integrano nel tutto". Catturare i contenuti di un documento in continuo mutamento a un dato momento è come registrare un campione di un'effimera trasmissione radiofonica. Definiamo ognuno di tali campioni, o *snapshots,* un "documento specifico*"*.

Quando accediamo a un documento dinamico sul web, l'*item* (ad esempio, esemplificazione) ricuperato da un utente può essere differente da tutti gli altri *item* dello stesso documento: può dipendere da una combinazione di un numero di fattori: l'identità dell'utente, gli strumenti di accesso impiegati (i browser web), l'informazione relativa a precedenti accessi allo stesso documento (conservati nei *cookies*), i parametri specificati dall'utente (ad esempio in un modulo) e ultimo, ma non da meno, lo stato attuale della base dati. Spesso l'*item* è generato sul momento (*on the fly*) quando un utente richiede un'esemplificazione. In altri termini, una chiamata HTTP agisce come un servizio di "print on demand": la copia consegnata riflette il contenuto del documento della base dati, qualunque esso sia al momento della stampa. La base dati può essere considerata come una rappresentazione fisica (semi) permanente del documento dinamico, da cui *item* specifici possono essere derivati. Gli stessi *item* non hanno una rappresentazione permanente, sono transitori a meno che non siano conservati in un archivio Internet.

4.1.2. Documenti specifici

Abbiamo definito un *item* che esemplifica un documento dinamico come un "documento specifico", che differisce da un *item* tradizionale per un aspetto principale: è un membro di un insieme di *item* che esemplificano lo stesso documento dinamico. Un documento conservato nell'archivio, o visualizzato all'utente, è ovviamente un documento specifico, ma ciò viene attenuato: una ricerca full-text ci darà per un documento dinamico tutt'al più *una* voce nell'elenco dei risultati. Se l'utente richiede una visualizzazione di un hit, il documento dinamico viene presentato come un'unità e l'utente può selezionare un *item* specifico "su una linea del tempo" (*timeline*), ad esempio una *menu line* che rappresenta la durata di vita del documento. Ogni versione conservata, ossia ogni documento specifico, è indicata su questa linea del tempo con un indicatore. L'utente può avere accesso a qualsiasi documento specifico cliccando l'indicatore per una certa data o tempo, ricuperando così l'*item* (vedi figura 2).

4.2. Definizioni del documento e dei metadati proposte dall'editore o dall'utente

La presentazione di documenti archiviati per ricerca e documentazione è solo uno dei servizi che verranno erogati dalla National Library of Norway. In aggiunta, abbiamo suggerito revisioni al servizio di assegnazione di identificatori della Biblioteca, basate sulle idee che abbiamo esposto. Oggi, questo servizio web assegna URN:NBN [7] alle università e altre istituzioni per il ramo norvegese del *name space* (dominio) URN:NBN. Possiamo, tuttavia, intravedere la possibilità di distribuire anche numeri ISBN stand-alone da questo servizio.

4.2.1. Lo scenario futuro

Uno scenario che mostri le future funzionalità è il seguente: la serie dell'identificatore primario attribuita da questo servizio richiede che l'utente/richiedente fornisca sia un insieme minimo di *metadati* che

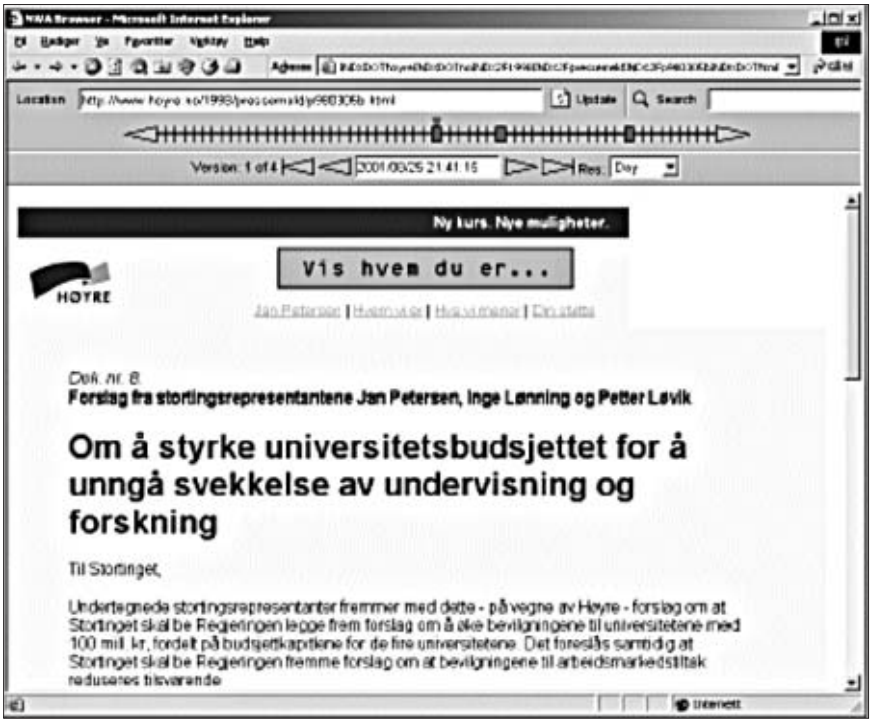


Fig. 2 – Presentazione di un documento dinamico nella NWA Access Tool User Interface

un'esatta definizione del documento identificato. Possono essere assegnati identificatori per *opera*, *espressione*, *manifestazione* (che comprenda la definizione di documenti dinamici) e *item* (che comprenda la definizione di documento specifico). *Item* (documenti specifici) dovrebbero essere specificati da una lista completa dei documenti (ad esempio un file HTML, file di immagini, di suoni ecc.); *manifestazioni* (documenti dinamici) possono essere specificate con definizioni quali: "La prima pagina di un quotidiano Internet a questo URL e tutte le pagine, collegate direttamente a partire dalla prima pagina, che risiedono sullo stesso sito web". Per gli identificatori dell'*espressione* e dell'*opera* l'utente può facoltativamente identificare *espressioni*/documenti dinamici e *item*/documenti specifici, che sono rappresentazioni, concretizzazioni, di questa *opera/espressione*.
Piuttosto che essere proposte automaticamente, vengono considerate conclusive e validate le definizioni avanzate dall'editore o dall'utente. Viene registrata l'identità dell'editore o dell'utente assegnando l'identificatore; una definizione di documento specificata da una casa editrice o università riconosciuta può essere valutata come più significativa di quella richiesta da un utente arbitrario, ovvero da un utente qualsiasi

4.2.2. Campi di metadati

Campi di metadati obbligatori e non obbligatori potrebbero essere disponibili per la descrizione del documento a ogni livello FRBR, e ogni livello verrebbe identificato con un URN:NBN. I valori dei metadati sarebbero archiviati con l'identificatore e gli utenti del nostro servizio di risoluzione basato su Internet saranno in grado di ricuperare il documento contraddistinto da questo numero. Dopo aver inserito l'informazione nei campi di metadati di un futuro strumento di assegnazione di metadati e identificatori, sarebbe possibile a un editore cliccare su un bottone, ad esempio <HTML Dublin Core>, per visualizzare questi metadati in HTML all'interno di un'altra finestra. L'utente potrebbe copiare i metadati e incollarli nell'elemento <HEAD> del documento web che si sta descrivendo prima di continuare il procedimento di assegnazione dell'identificatore. Dopo aver salvato il documento digitale, che ora comprende i metadati incorporati, l'utente potrebbe facilmente conservare la copia del documento arricchita di metadati nell'archivio della biblioteca cliccando il bottone di aggiornamento del browser.

4.3. Possibili servizi di verifica e di autenticazione dei documenti?

Si racconta di alcune personalità che, dopo aver sottoposto a revisione rapporti ufficiali su Internet, hanno in seguito rifiutato di riconoscere l'esistenza delle precedenti versioni. Abbiamo anche udito di aziende commerciali che pubblicizzano i loro prodotti a un dato prezzo, e poi richiedono al cliente una somma maggiore.
Facendo tesoro di queste e altre storie, il Paradigma Project propone un servizio di verifica e autenticazione che potrebbe consentire agli utenti di richiedere un download di un documento Internet, ad esempio un'istantanea (*snapshot*) di una pagina web con una particolare offerta commerciale, la formulazione della responsabilità legale, un'asserzione diffamatoria ecc. Se dubbi si sollevassero in merito al contenuto di questi documenti a una certa data, la biblioteca potrebbe allora confermare (o rigettare) qualsiasi rivendicazione/richiesta a tale proposito. Anche quando non è coinvolto alcun aspetto legale, un *item* di un documento specifico che si è preservato può servire come una ben definita immagine di un documento dinamico a un tempo determinato, ad esempio per una citazione. Ciò è importante, specialmente quando si consideri che gran parte dei documenti Internet non ha numerazione delle pagine, numero di versione ecc.
Nel nostro archivio Internet, un documento specifico è definito nella forma in cui è stato ricevuto dal server web. Esiste un ben definito flusso di bit per ogni componente del documento (testo, immagini ecc.). La resa grafica del documento *non è* parte della sua definizione – questo processo è lasciato allo strumento di accesso. Il documento specifico è identificato come il contenuto di un documento dinamico dato da certi componenti e metadati:
– l'origine di ciascun componente (ad esempio un URL);
– l'insieme dei parametri specificati dal client quando ricupera i componenti;
– *the wall clock time* in cui ciascun componente è stato ricuperato;
– il set di componenti compresi nel documento.

5. Conclusione

Il Paradigma Project della National Library of Norway sta operando intensamente per determinare, entro il periodo che rimane alla definizione completa del progetto, tecnologie soddisfacenti, metodologie e routine per il deposito legale di tutti i tipi di documenti digitali, compresi i milioni di documenti allocati sul dominio Internet norvegese. Si spera di essere in grado di fornire ai nostri utenti accesso ai materiali archiviati tramite le registrazioni bibliografiche, diversi tipi di metadati e strumenti di ricerca full-text già nel 2005.
Il nostro archivio Internet strutturato su FRBR sarà di certo uno dei primi di questo tipo e speriamo anche

di realizzare la nostra idea circa l'uso dei livelli delle entità *opera*, *espressione*, *manifestazione* e *item* di FRBR in un futuro servizio di autenticazione su Internet. Il tempo lo dirà, ma nel frattempo la National Library of Norway continuerà a esplorare nuove modalità per conservare il patrimonio culturale digitale della Norvegia e per fornire ai suoi utenti gli strumenti che possano aprire le porte di questa entusiasmante biblioteca digitale.

Tutti i siti citati nelle note e in bibliografia sono stati visitati dagli autori il 15 aprile 2004 e controllati il 22 settembre 2004.

(Traduzione di Stefano Gambari e Mauro Guerrini)

Note

1 La relazione in inglese è disponibile all'indirizzo: <http://www.ifla.org/IV/ifla70/papers/009eNuys.pdf>. Code number: 009-E Meeting-89. Cataloguing.

2 Cfr. Stefano Gambari – Mauro Guerrini, *Definire e catalogare le risorse elettroniche*, Milano, Editrice Bibliografica, 2002, p. 197-201.

3 <http://www.bncf.firenze.sbn.it/notizie/testi/comunicatositiWeb.htm>.

4 Johan Mannerheim, *The WWW and our digital heritage,* [online], <http://ifla.org/IV/ifla66/papers/158-157e. htm>.

5 Per maggiori informazioni relative alla National Library of Norway, vedi <http://www.kb.nl/gabriel/libraries/pages_generated/no_en.html>.

6 Halgrímsson Torsteinn (2003, February 28), *Web archiving in Europe* [discussion]. – NWA [online]. E-mail address: nwa@nb.no.

7 Per maggiori informazioni sul Paradigma Project, vedi: <http://www.nb.no/paradigma/eng_ index.html>.

8 Per maggiori informazioni sul software HTTrack, vedi: <http://www.httrack.com/>.

9 Su BIBSYS, vedi: <http://www.bibsys.no/english. html>.

10 Per maggiori informazioni su questa attività relativa al *deep web*, vedi: <http://www.nla.gov.au/ntwkpubs/gw/66/html/p15a01.html>.

11 Per maggiori informazioni sull'*harvester* NEDLIB, vedi: <http://www.csc.fi/sovellus/nedlib/ver11/documentation11.doc>.

12 Informazioni sul Nordic Web Archive Project all'indirizzo: <http://nwa.nb.no>.

13 Nel testo inglese compare: Olsen (*ndt*).

14 Per maggiori informazioni su FAST Search & Transfer (FAST) ASA, vedi: <http://www.fast.no>.

15 Una delle numerose indagini sui metadati che abbiamo studiato è: *DESIRE: a review of metadata: a survey of current resource description formats*, 1997, all'indirizzo: <http://www.ukoln.ac.uk/metadata/desire/overview/rev_toc. htm>.

16 Per maggiori informazioni su Dublin Core Metadata Initiative, vedi: <http://www.dublincore.org>.

17 Per un approfondimento sul Wizard's MAVIS system, vedi: <http://www.wizardis.com.au/ie4/products/mavis/introducingmavis.html>.

18 La versione norvegese di MARC è chiamata NORMARC; alcuni sistemi hanno adottato versioni locali, ad esempio BIBSYS MARC, e stiamo discutendo a livello nazionale dell'uso di MARC21. Per maggiori informazioni su MARC21, vedi: <http://www.loc.gov/marc/bibliographic/ecbdhome.html>.

19 Su MODS vedi: <http://www.loc.gov/standards/mods/>.

20 Per approfondire sul formato METS, vedi: <http://www.loc.gov/standards/mets/>.

21 Per maggiori informazioni su ONIX, vedi: <http://www.editeur.org/onix.html>.

22 Per maggior informazioni sull'OAIS Reference Model, vedi: <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>.

23 Un contributo con la nostra proposta di meccanismi complementari sarà disponibile sul fascicolo monografico di "Cataloging & classification quarterly" dedicato nel 2005 a FRBR.

Bibliografia selezionata

[1] Final Report for the Pilot project "Netarkivet.dk" [online]. – URL: http://www.netarkivet.dk/rap/webark-final-rapport-2003.pdf.

[2] Guidelines for the selection of online Australian publications intended for preservation by the National Library of Australia [online]. – URL: http://pandora.nla.gov.au/selectionguidelines.html.

[3] The Kulturarw3 Project – The Royal Swedish Web Archiw3e – An example of "complete" collection of web pages [online]. – URL: http://www. ifla.org/IV/ifla66/papers/154-157e. htm.

[4] Nancy Olson (2002). Cataloging Internet Resources : A Manual and Practical Guide [online]. – OCLC. – URL: http://www.oclc.org/support/documentation/worldcat/cataloging/internetguide/1/1.htm.

[5] Norway. [The Legal Deposit Act (1989)] Act relating to the legal deposit of generally available documents : no. 32 of 9 June 1989 : with regulations / [published by the Ministry of Church and Cultural Affairs ; unofficial English translation published by the National Library of Norway. – [Oslo] : National Library of Norway, 1997. – p. 21.

[6] IFLA Cataloguing Section Working Group on the Use of Metadata Schemas (2003). Guidance on the structure, content, and application of metadata records for digital resources and collections : draft for worldwide review 27 October, 2003 [online]. – URL: http://www.ifla.org/VII/s13/guide/ metaguide03.pdf.

[7] RFC 3188 Using National Bibliography Numbers as Uniform Resource Names [online] / J. Hakala, 2001. – URL: http://www.ietf.org/rfc/ rfc3188.txt.

[8] Carol Van Nuys (2003). Identification of network accessible documents : problem areas and suggested solutions [online] / Carol van Nuys, Ketil Albertsen. – p. 13-25. – 1: Proceedings : in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries, ECDL 2003 / Julien Masanès, Andreas Rauber, Gregory Cobena (eds). – URL: http://bibnum.bnf.fr/ecdl/2003/index.html.

[9] Ketil Albertsen (2003). The Paradigma web harvesting environment. – p. 49-62. – I: Proceedings : in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries, ECDL 2003 / Julien Masanès, Andreas Rauber, Gregory Cobena (eds). – URL: http://bibnum.bnf.fr/ecdl/2003/index.html.

[10] Carol Van Nuys (2003). The Paradigma Project [online]. – I: RLG DigiNews. – Vol. 7, no. 2. – URL: http://www.rlg.org/preserv/diginews/v7_n2_feature2.htmlBack to the Programme: http://www.ifla.org/IV/ifla70/prog04.htm.