

Conservazione e accesso alla memoria digitale della società: esperienze e proposte¹.

Nel mondo del digitale, la scrittura e più in generale l'informazione non è più legata a un determinato supporto. Si parla di "rappresentazione informatica" ovvero di "una sequenza di bit, che, elaborata da un sistema informatico, può essere resa visibile su uno schermo, stampata sulla carta o inviata a distanza". Si tratta di "un cambiamento radicale nella concezione e nell'uso del documento, così come lo conosciamo da migliaia di anni, nella sua natura di res signata, cioè di una cosa che riporta dei segni, delle informazioni".

Più in generale la risorsa digitale non dipende quindi necessariamente da un supporto, ma assume "una funzione autonoma rispetto alla sua (eventuale) fissazione su un supporto materiale²". Per risorsa digitale si intende qui fare riferimento a un insieme di bit dove "informazioni di tipo diverso possono essere tutte ridotte allo stesso codice di base, alle lunghe catene di 0 e di 1 dell'informazione digitalizzata³": la posta elettronica, le pagine web, la musica su CD, i film su DVD sono esempi di risorse digitali che negli ultimi hanno profondamente modificato la nostra vita e le nostre abitudini.

La permanenza nel tempo delle risorse digitali non può essere quindi vista come permanenza del supporto del digitale. In altre parole gli articoli che di solito in prossimità dell'estate compaiono sui quotidiani nazionali e che mettono a confronto la durata dei CD o dei DVD con la durata della carta o con la Stele di Rosetta sono del tutto fuorvianti. Se il digitale è indipendente dal supporto allora conservare la risorsa digitale non significa conservare il supporto, ma mantenere leggibile (tramite un computer) nel tempo la "catena di 0 e di 1".

In sintesi la risorsa digitale è copiabile e trasferibile con facilità e - se la catena di 0 e di 1 rimane inalterata - non ha senso parlare di copia, clone o duplicato⁴. Avremo senza alcuna difficoltà più originali, ma tutti fragili perché sempre dipendenti da dispositivi hardware e da programmi software (ma non dal supporto sul quale in quel momento di trovano). La risorsa tradizionale ha invece forti dipendenze dal supporto che la veicola, ma non vi sono in generale forti dipendenze da dispositivi particolari (si pensi ad esempio alla immediata leggibilità del testo a stampa). Si può copiare, ma conservare la copia non significa conservare l'originale: un libro a stampa digitalizzato non può sostituire per se l'originale (anche se può risultare di indubbia utilità in molti contesti).

Quando si parla di conservazione del digitale occorre dire che molte cose sono state risolte⁵ e molte altre restano ancora da fare (come del resto in tutte le attività umane)⁶. Resta il fatto che una enorme massa di risorse digitali è presente oggi sul

web e che una percentuale davvero alta (in alcuni studi si parla del 17 per cento⁶) ogni anno scompare e non è più disponibile.

La mancata conservazione nel tempo delle risorse digitali comporta almeno tre rischi: la manipolazione delle fonti, la non verificabilità delle citazioni e la privatizzazione della memoria⁷.

Per quanto riguarda il primo rischio si può citare una famosa pagina web appartenente al sito ufficiale della presidenza degli Stati Uniti. Si tratta di una pagina che informa sullo stato della guerra in Iraq. Nel maggio del 2003 la pagina risultava pubblicata con questo titolo⁸: “President Bush Announces Combat Operations in Iraq Have Ended”. A dicembre dello stesso anno il titolo risultava “leggermente” modificato, ma di questa modifica non si trovava nessuna nota redazionale⁹: “President Bush Announces Major Combat Operations in Iraq Have Ended”. Non sfuggono le ragioni di questa modifica: in maggio la guerra veniva data come totalmente terminata, in realtà - come è apparso subito evidente - così non era. La modalità dell’aggiunta di quel *Major* nel titolo sembrano invece indirizzate a far sembrare che quel titolo sia sempre stato formulato in quel modo: in altre parole - visto dalla prospettiva del futuro lavoro dello storico - un caso di “manipolazione delle fonti” ottenuto grazie alla indipendenza dal supporto del digitale ovvero alla facilità con la quale le catene di 0 e di 1 possono essere modificate senza lasciare traccia (una manipolazione di questo tipo non sarebbe stata possibile nel mondo testo a stampa). Solo grazie al lavoro di conservazione di Internet Archive che dal 1996 archivia “istantanee” dello spazio web a livello mondiale, il caso è stato documentato.

Il secondo rischio si riferisce alla messa in discussione dei fondamenti della ricerca scientifica. Senza ad esempio la possibilità di verificare le citazioni, un articolo scientifico non può essere valutato e i suoi risultati non possono essere assunti come base di partenza per ulteriori sviluppi (occorrerebbe ogni volta rifare tutto il percorso).

L’ultimo rischio riguarda la progressiva “privatizzazione del sapere”. Google è una grande azienda quotata in borsa, Internet Archive è una fondazione, non ha scopo di lucro, ma non è una istituzione pubblica. Il rischio della “posizione dominante” nel campo del sapere è da prendere in conto con estrema attenzione. La recente iniziativa di Google di digitalizzare 15 milioni di libri appartenenti alle collezioni di 5 grandi biblioteche e di rendere il testo di questi libri ricercabile su Google (“cerchiamo di mettere il contenuto dei libri dove è più facile trovarlo, ovvero nei risultati della ricerca di Google¹⁰”) ha fatto molto discutere.

La ben nota reazione del Presidente della Bibliothèque nationale de France Jean-Noël Jeanneney¹¹ che parla di “sfida di Google all’Europa” si basa sostanzialmente su due punti. Il primo riguarda il fatto che sono state scelte solo biblioteche angloamericane con la conseguente predominanza della lingua e della cultura inglese. In realtà - come hanno ben documentato studi successivi sulla composizione delle

collezioni selezionate - la percentuale di libri di area angloamericana è poco meno della metà¹².

E' il secondo punto invece che ci riporta al rischio di "privatizzazione del sapere". Secondo il Presidente della Bibliothèque nationale de France non bisogna lasciare che in questo campo sia "il mercato" a decidere, occorre un grande investimento pubblico e invita l'Europa a promuovere un grande progetto di Biblioteca Digitale Europea.

Non si vuole qui approfondire le differenti visioni sul ruolo del mercato. Quello che è qui importante rilevare che - almeno dalla rivoluzione francese in poi - le biblioteche si presentano come "servizio pubblico": con un mandato e con un controllo della comunità¹³.

Naturalmente non si tratta di "chiudere Google" o di impedire a Google l'accesso ai contenuti europei¹⁴. Google secondo molti è il punto di sviluppo più significativo di 15 anni di Internet e in questo momento uno strumento fondamentale per l'accesso all'informazione. Per impedire le "posizioni dominanti" occorre che l'accesso al sapere sia considerato come servizio pubblico: occorre prevedere grandi investimenti pubblici nel campo dell'organizzazione del sapere (dalla digitalizzazione dei libri su larga scala alla creazione di strumenti innovativi per l'accesso e la conservazione delle risorse digitali). Nuovi equilibri potranno sicuramente nascere dalla dialettica e dalla cooperazione di ruoli e visioni differenti: iniziative di tipo "protezionistico" sarebbero sicuramente destinate al fallimento.

E' interessante a questo proposito la scelta di Internet Archive di aderire a un Consorzio di biblioteche nazionali per la conservazione di Internet: IIPC, International Internet Preservation Consortium¹⁵. Si tratta di un interessante esempio di cooperazione tra pubblico e privato. Fanno parte del Consorzio la Biblioteca Nazionale Centrale di Firenze, la Biblioteca Nazionale Francese (che ha attualmente il ruolo di coordinatore), la Library of Congress, la British Library, la Biblioteca nazionale australiana e quella canadese; inoltre le biblioteche nazionali di Svezia, Finlandia, Norvegia, Danimarca e Islanda portano nel Consorzio l'esperienza di molti anni di raccolta periodica dello spazio web nazionale del Nordic Web Archive. Il Consorzio si propone di:

- rendere possibile la conservazione (intesa come salvaguardia e possibilità di accesso nel lungo periodo) della ricchezza dei contenuti di Internet provenienti da tutto il mondo;
- favorire lo sviluppo e l'uso di strumenti, tecniche e standard per la creazione di archivi internazionali;
- sostenere le biblioteche nazionali che intendono occuparsi dell'archiviazione di Internet.

Oltre ai rischi occorre tener conto anche di una obiezione: è davvero utile e necessario conservare tutte le risorse digitali che si presentano in rete? Come sappiamo il web è dominato dalla immediatezza e dalla quantità. La mancanza di

mediazione che invece caratterizza l'editoria tradizionale (scelte editoriali, peer review, ecc.) mette a serio rischio la qualità di quanto viene messo in rete e il rischio è quello di investire risorse in un lavoro che richiede enormi investimenti e che non trova una giustificazione sul piano del "valore culturale".

La selezione tuttavia presenta innegabili aspetti problematici. Nel contesto del "deposito legale" le biblioteche nazionali hanno sempre cercato di assicurare la più ampia copertura possibile riducendo al minimo i rischi e i costi collegati alla selezione. In ogni caso la scelta dovrebbe essere fatta con criteri pubblici e oggettivi (non dipendenti dalla convinzioni e dai pregiudizi di chi sceglie). La formalizzazione dei criteri di selezione fa subito emergere contraddizioni di non facile soluzione. Se ad esempio un criterio di selezione è il genere potremmo escludere dal deposito legale tutti i weblog: ma con questa decisione perderemo la possibilità di archiviare diari in rete di autori citati anche nella letteratura scientifica "certificata"¹⁶.

La convinzione del Consorzio IIPC è che lo strumento dell'harvesting (della raccolta dei siti web) sia una tecnologia oggi in grado di far diventare il deposito legale dei siti web una attività sostenibile con risultati misurabili. Come è noto l'harvesting viene usato dai motori di ricerca (es. Google) per indicizzare il web; ma viene usato anche da oltre 10 anni per "l'archiviazione del web" da parte di Internet Archive¹⁷. Esistono ormai da tempo attività di harvesting portate avanti da biblioteche nazionali.

L'harvesting ha ovviamente anche "controindicazioni": con l'harvesting si ottengono "fotografie a intervalli di tempo" di un determinato spazio web (con la conseguente perdita degli intervalli); tutto il cosiddetto "web profondo" (deep web) rimane inaccessibile all'agente software che si occupa dell'harvesting (crawler). In altre parole l'harvesting non è una tecnologia in grado di risolvere tutti i problemi di deposito legale del digitale in rete, ma una tecnologia in grado di offrire un'ampia base di risultati. Per convenzione si parla di web profondo con riferimento a siti non raggiungibili dai tradizionali motori di ricerca (e quindi non raggiungibili nemmeno da un crawler). Tra questi si indicano di solito:

- siti non accessibili liberamente (per esempio a pagamento e protetti da password);
- siti che offrono le risorse digitali come risultato di una interrogazione da parte di un utente (per esempio la ricerca in un catalogo dove l'utente inserisce il titolo, l'autore ecc di un libro, oppure il servizio offerto da *Google maps* che genera su richiesta la carta geografica desiderata).

Nel primo caso l'harvesting potrà funzionare solo se il sito "aprirà le porte" al crawler (per esempio fornendo la password all'istituzione depositaria). Nel secondo caso occorrerà una forte collaborazione tra il produttore dell'informazione e la biblioteca depositaria. Non è ovviamente pensabile che l'istituzione depositaria installi e mantenga tutti i database e tutte le applicazioni che generano le pagine web. Ci sono sperimentazioni a questo proposito (Francia e Australia) di invio alla biblioteca depositaria di record esportati in formato XML da database che "alimentano" il deep web¹⁸.

Occorre ricordare che i crawler di nuova generazione permettono di impostare delle regole di priorità nella raccolta per evitare ad esempio che un sito di una agenzia di viaggi venga raccolto con la stessa frequenza di un sito di una università. In altre parole anche per quanto riguarda l'harvesting è possibile impostare regole di selezione che riguardano essenzialmente:

- la definizione - sulla base di determinati criteri - della lista di indirizzi di partenza (URL) chiamati anche "semi" del crawler. Nel caso del deposito legale italiano i semi di partenza potrebbero essere dati da tutti i domini "punto it";
- la definizione di regole con le quali i semi di partenza si accresceranno con nuovi semi durante una sessione di harvesting (ad esempio gli indirizzi di altri siti trovati nella pagina raccolta, ma non quelli che non sono "punto.it") ;
- la definizione della frequenza di raccolta (quante volte in un determinato periodo di tempo si desidera che quella pagina venga raccolta) .

In generale si può dire che la definizione degli insiemi di partenza (i semi) per l'harvesting sono il risultato di interrogazioni su archivi esistenti: ad esempio è possibile estrarre da Internet Archive i siti "punto it". Inoltre la possibilità di impostare la frequenza della raccolta (ogni quanto tempo si ritorna sullo stesso sito) è una caratteristica che differenzia i crawler di nuova generazione da quelli storici. Oggi non si parla più di "istantanee periodiche" dello spazio web ma di "raccolta continua" dove il crawler è in grado dinamicamente di gestire i ritorni a partire dalle "priorità" che vengono date in input. I criteri di priorità sono comunque liste di siti (modificabili dinamicamente nel corso della raccolta) basati su criteri "oggettivi" quali:

- le pubblicazioni di fonte pubblica (Stato, regioni enti locali ecc.);
- la produzione scientifica delle università e delle istituzioni di ricerca ;
- criteri usati dai motori di ricerca come la frequenza di aggiornamento¹⁹ del sito, i link in entrata (numero di siti che citano un determinato sito) ecc¹⁹.

La Biblioteca Nazionale Centrale di Firenze nel corso del progetto Crawler - finanziato dalla Biblioteca Digitale Italiana - ha dato vita ad una prima sperimentazione su larga scala di raccolta dello "spazio web nazionale". In collaborazione con Internet Archive nel corso di quattro settimane (tra maggio e giugno 2006) sono stati raccolti 7.22 Terabyte di dati in formato WARC²⁰ (oltre i 2 milioni i server contattati per circa 240 milioni di documenti raccolti). Il software usato per il crawler è Heritrix²¹ (open source prodotto da Consorzio IIPC). Punto di partenza sono stati 648.255 siti "punto it": la lista dei "semi" è stata prodotta da Internet Archive sulla base delle sue raccolte precedenti. Naturalmente i siti "punto it" non esauriscono lo "spazio web italiano" (come è noto molti siti italiani sono registrati come "punto com", "punto net", ecc), ma questa scelta è stata un compromesso inevitabile. A parte la problematicità della definizione di "spazio web italiano", occorre ricordare che questa definizione va poi tradotta in istruzioni che permettono di creare automaticamente la lista di partenza. Ad esempio se si definisce come appartenente allo "spazio web italiano" un sito che contiene uno o più documenti in lingua italiana, per realizzare la lista dei "semi" di partenza si potrebbe analizzare - tramite un software di

riconoscimento della lingua - tutti i documenti presenti su Internet Archive (a oggi intorno ai mille terabyte)

Naturalmente il progetto Crawler deve essere visto come un punto di partenza. Con questo progetto siamo in grado di fare una prima stima dei tempi e dei costi per l'harvesting italiano: un importante e concreto contributo per la sperimentazione che il nuovo Regolamento sul deposito legale prevede per quanto riguarda i "documenti diffusi via rete informatica"²²—.

Tra tutti gli aspetti che possono essere sottolineati quando si parla di conservazione delle risorse quello più rilevante è sicuramente l'aspetto organizzativo. Dal punto di vista delle istituzioni della memoria la conservazione delle risorse digitali può essere vista come un *servizio pubblico*²³ fornito da depositi digitali accreditati (trusted digital repositories)²⁴—.

Per servizio pubblico si intende proprio quel servizio che - ritenuto essenziale e strategico da una determinata comunità - è accessibile indipendentemente dalle possibilità economiche del fruitore. Per depositi digitali accreditati si fa riferimento a standard per la certificazione di affidabilità quali quelli proposti dal RLG²⁵—.

Il servizio pubblico di conservazione delle risorse digitali ha lo scopo di assicurare nel lungo periodo per le risorse digitali depositate:

1. la **vitalità** (viability): le sequenze di bit che compongono i file sono intatte (ogni risorsa depositata può essere rappresentata da uno o più file)²⁶—;
2. la **traducibilità** da parte di un elaboratore (renderability): un determinato hardware e un determinato software sono in grado di gestire le risorse digitali depositate, come ad esempio visualizzare a video un documento in formato PDF;
3. l'**autenticità** delle risorse depositate: intesa come documentazione della identità e della integrità²⁷—;
4. la **fruibilità** da parte delle comunità di riferimento. Ad esempio il deposito rende possibili servizi quali quelli proposti dal modello FRBR: trovare, identificare, selezionare, ottenere una risorsa digitale (o un insieme di risorse)²⁸—.

In molti casi le risorse digitali per essere depositate dovranno essere convertite in formati controllabili e gestibili nel tempo dal servizio di deposito. Da un lato infatti il servizio di deposito non può accettare formati che sono sconosciuti o protetti (si pensi ad esempio ai file protetti da DRM)²⁹—, dall'altro la natura stessa della risorsa digitale rende necessarie una attività di conversione. Molte risorse digitali si presentano infatti come il risultato di una specifica interrogazione effettuata da un determinato utente. Come si è visto precedentemente nel caso del web profondo le risorse digitali che il servizio di deposito prende in consegna non sono necessariamente identiche (a livello di bit) a quelle che si presentano nella realtà.

Un contributo al primo obiettivo (la vitalità - viability - delle risorse acquisite) è offerto dal progetto *Magazzini digitali*³⁰ - finanziato nel 2006 dalla Fondazione Rinascimento Digitale e dalla Biblioteca Nazionale Centrale di Firenze - che si propone di realizzare una sperimentazione³¹ di archiviazione di un ammontare di dati ritenuto significativo (10 TeraByte³²). Proprio come per i magazzini librari le risorse digitali si incrementano per aggiunta (come un libro dopo l'altro sullo scaffale); la cancellazione o la modifica di una risorsa non è di norma prevista. Infine l'accesso a una risorsa archiviata può avvenire in tempi difficilmente prevedibili: dal qualche secondo dopo l'archiviazione a qualche anno dopo; come per i libri di una biblioteca è possibile che nel lungo periodo un libro non venga mai richiesto.

L'architettura del progetto è ispirata a due grandi archivi di risorse digitali esistenti: Google e Internet Archive. Per il primo si fa qui riferimento all'intervento presentato da un gruppo di ricercatori di Google al Convegno sui sistemi operativi SOSP'03 dal titolo *The Google file system*³². Per il secondo si fa riferimento al forum sul Petabox³³ iniziato nel 2004 ma ancora attivo sul sito web di Internet Archive³⁴.

Sia Google che Internet Archive partono dalla considerazione che nel mondo dei sistemi informatici il malfunzionamento non è l'eccezione, ma la regola. Il rischio di perdere i dati può essere affrontato tramite la ridondanza (più copie dello stesso file su macchine differenti e localizzate in luoghi differenti) e la facile sostituibilità dei componenti hardware. Il componente più adatto a questo scopo è proprio il personal computer: poco costoso, facilmente sostituibile e soprattutto non dipendente da un particolare fornitore hardware o software. Oggi un comune personal computer casalingo può arrivare ad archiviare fino a 2.8 TB (4 dischi da 700 GB) con tecnologia SATA (ovvero proprio quella oggi più diffusa).

In concreto Magazzini Digitali sta sperimentato concretamente questa soluzione:

- si è proceduto all'acquisizione di 10 personal computer ognuno di quali dotato di 4 dischi SATA da 500 GB;
- sulle macchine è stato installato un sistema operativo open source (una comune distribuzione Linux) e un software di base - anche questo open source - per la replica automatica dei dati (rsync³⁴): non viene fatto uso di nessun tipo di scheda particolare, di RAID ecc proprio per evitare qualsiasi tipo di dipendenza;
- è stata predisposta una architettura multi-sito (5 macchine sono state installate alla Biblioteca Nazionale Centrale di Firenze, 5 alla Biblioteca nazionale Centrale di Roma)
- è previsto anche un terzo sito che utilizza - per aumentare la sicurezza complessiva del sistema - una tecnologia del tutto differente dai primi due siti: si tratta di una copia dei dati su nastri magnetici del tipo LTO Ultrium3³⁵. Il sito ha le funzioni tipiche dell'*archivio nascosto* (*dark archive*) ovvero di quell'archivio da usare solo nei casi di emergenza³⁶
- in pratica per ogni file, per ogni sequenza di bit, esistono complessivamente 5 repliche (2 a Firenze, 2 a Roma e 1 nel *dark archive*)

Uno dei vantaggi di questa architettura la scalabilità (la capacità di archiviazione può essere facilmente incrementata senza particolari vincoli compreso quello elemento di acquisire i nuovi elementi che si rendono necessari dallo stesso fornitore dei precedenti elementi) e la facilità di manutenzione (sono sufficienti tecnici che conoscono il funzionamento di un personal computer).

Le evoluzioni del progetto Magazzini digitali previste nel 2007 si occuperanno da una lato del secondo obiettivo (la **traducibilità** - *renderability* - da parte di un elaboratore) e dall'altro della *certificazione* del sistema di deposito complessivo.

Nel primo caso si tratta di occuparsi di tecnologie di conservazione (emulazione e migrazione principalmente³⁷) cercando di mettere insieme le soluzioni che in questo campo cominciano a essere disponibili.

Nel secondo caso si tratta verificare che il servizio di deposito raggiunga gli obiettivi indicati al terzo e quarto punto (l'autenticità e la fruibilità di quanto viene archiviato). Il punto di partenza sarà il documento *Audit checklist*³⁸ del RLG per la certificazione di archivio digitale che prevede la verifica degli aspetti che seguono:

- A. Organizzazione (stabilità, capacità economiche ecc.);
- B: Funzioni, processi e procedure;
- C: Comunità di riferimento e uso dell'informazione;
- D: Tecnologie e infrastruttura tecnologica (questo punto equivale alla certificazione della sicurezza dei sistemi informatici prevista dalla norma ISO 27001:2005).

Come abbiamo visto la conservazione e l'accesso alla memoria digitale della società è un servizio pubblico che le istituzioni della memoria sono oggi chiamate ad offrire. Il successo di questo servizio non dipenderà solo dalla disponibilità di strumenti tecnologici, ma anche e soprattutto da adeguate soluzioni organizzative. Il servizio pubblico di conservazione delle risorse digitali non è un servizio centralizzato ma il risultato di una forte cooperazione tra le istituzioni della memoria.

¹ Giovanni Bergamin. Documento di lavoro provvisorio ad uso didattico. Una versione ufficiale di questo documento è pubblicata in *Conservazione e accesso alla memoria digitale della società: esperienze e proposte in Fare storia in rete: fonti e modelli di scrittura digitale per la storia dell'educazione, la storia moderna e la storia contemporanea / a cura di Gianfranco Bandini e Paolo Bianchini. Roma: Carocci, 2007, pp. 153-163.*

² Cammarata, Manlio - Maccarone, Enrico. Introduzione alla firma digitale: 9, La natura del documento informatico. 2000. <http://www.interlex.it/docdigit/intro/intro9.htm>

³ Ciotti, Fabio - Roncaglia, Gino. Il mondo digitale. Introduzione ai nuovi media. Roma [ecc.], Laterza, 2000, p. 348.

⁴ Cammarata, Manlio - Maccarone, Enrico. Introduzione alla firma digitale ... cit.

- 5 Per una rassegna introduttiva sulle strategie si veda il servizio PADI della Biblioteca nazionale australiana <http://www.nla.gov.au/padi/topics/18.html>
- 6 O'Neill, Edward T.- Lavoie, Brian F - Bennett, Rick. Trends in the Evolution of the Public Web 1998 - 2002. <<D-Lib Magazine>> 9(2003) <http://www.dlib.org/dlib/april03/lavoie/04lavoie.html>
- 7 Hoog, Ammanuel. Internet a-t-il une mémoire? <http://www.ac-versailles.fr/pedagogi/ses/vie-ses/hodebas/hog1.htm> Originariamente pubblicato su Le Monde del 16 agosto 2002.
- 8 Grazie al lavoro di Internet Archive <http://www.archive.org> la pagina “fotografata” e archiviata risulta essere la seguente http://web.archive.org/web/20030506213853/http://www.whitehouse.gov/news/releases/2003/05/images/20030501-15_lincoln4-515h.html.
- 9 Come si può vedere anche dalla versione corrente - verificata al 3.4.2006 - della pagina http://www.whitehouse.gov/news/releases/2003/05/images/20030501-15_lincoln4-515h.html
- 10 Il testo è ricercabile, ma per i libri soggetti a copyright è esclusa la consultazione integrale in rete : <http://books.google.it/intl/it/googlebooks/library.html>
- 11 Jeanneney, Jean Noel..Quand Google défie l'Europe, Mille et une nuits, 2005.
- 12 Lavoie, Brian - Connaway, Lynn Silipigni - Dempsey. Lorcan. Anatomy of aggregate collections: the example of Google Print for libraries. << D-Lib Magazine>> 11(2005). <http://www.dlib.org/dlib/september05/lavoie/09lavoie.html>
- 13 Su queste tematiche è fondamentale: Traniello, Paolo La biblioteca pubblica. Il mulino, 1997.
- 14 <http://www.cnn.com/2005/TECH/12/30/poll.results/>
- 15 <http://netpreserve.org/about/index.php>
- 16 Il progetto Pandora della Biblioteca nazionale australiana si basa proprio sulla selezione operata dal bibliotecario con i criteri definiti in <http://pandora.nla.gov.au/selectionguidelines.html>. Attualmente la stessa Biblioteca sta parallelamente sperimentando anche l'harvesting dei siti web.
- 17 <http://www.archive.org>
- 18 <http://www.nla.gov.au/xinq/>
- 19 Masanès, Julien. Towards continuous web archiving, <<D-Lib Magazine>>, 8(2002), <http://www.dlib.org/dlib/december02/masanes/12masanes.html>
- 20 Si tratta di un formato che rappresenta l'evoluzione a cura di IIPC del formato ARC di Internet Archive e che sta per essere standardizzato in ambito ISO: <http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml>
- 21 Informazioni sul crawler Heritrix <http://crawler.archive.org/>
- 22 DPR n. 252 del 3 maggio 2006, art. 37.
- 23 Una buona definizione di servizio pubblico può essere trovata in http://en.wikipedia.org/wiki/Public_services
- 24 Trusted Digital Repositories: Attributes and Responsibilities 2002 <http://www.rlg.org/en/pdfs/repositories.pdf>
- 25 An Audit Checklist for the Certification of Trusted Digital Repositories.(draft 2005 http://www.rlg.org/en/page.php?Page_ID=20769

- [26](http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Future-Proofing%20The%20Web%20What%20We%20Can%20Do%20Today%20-%20John%20Kunze.pdf) La terminologia usata in questa definizione è basata su Kunze, Jhon. Future-proofing the web: what we can do today. 2005 <http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Future-Proofing%20The%20Web%20What%20We%20Can%20Do%20Today%20-%20John%20Kunze.pdf>
- [27](http://www.clir.org/pubs/reports/pub92/lynch.html) Lynch, Clifford. Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust. 2000. <http://www.clir.org/pubs/reports/pub92/lynch.html>
- [28](http://www.ccsds.org/documents/650x0b1.pdf) Per il significato di “comunità di riferimento - designated communities” e in generale per i concetti fondamentali collegati alla conservazione del digitale occorre far riferimento a Reference Model for an Open Archival Information System. 2002. (ISO Standard 14721). <http://www.ccsds.org/documents/650x0b1.pdf>. Per il modello FRBR: Functional requirements for bibliographic records: final report 1998. <http://www.ifla.org/VII/s13/frbr/frbr.pdf>
- [29](http://en.wikipedia.org/wiki/Digital_Rights_Management) http://en.wikipedia.org/wiki/Digital_Rights_Management
- [30](http://www.rinascimento-digitale.it/index.php?SEZ=28) <http://www.rinascimento-digitale.it/index.php?SEZ=28>. Il nome del progetto Magazzini digitali si richiama al progetto europeo NEDLIB dove è stato usato per la prima volta il termine Digital Stacks sul modello dei magazzini librari (Steenbakkers, Johan. The NEDLIB guidelines, NEDLIB Consortium, 2000, p. 3.
- [31](http://labs.google.com/papers/gfs-sosp2003.pdf) Attualmente - dicembre 2006 - la BNCf possiede circa 13 TB di digitalizzazioni e 7.2 TB come risultato della sperimentazione dell'archiviazione del web.
- [32](http://labs.google.com/papers/gfs-sosp2003.pdf) The Google File System 2003. <http://labs.google.com/papers/gfs-sosp2003.pdf>
- [33](http://www.archive.org/web/petabox.php) <http://www.archive.org/web/petabox.php>
- [34](http://samba.anu.edu.au/rsync/) <http://samba.anu.edu.au/rsync/>
- [35](http://en.wikipedia.org/wiki/Linear_Tape-Open) http://en.wikipedia.org/wiki/Linear_Tape-Open
- [36](http://www.dpconline.org/docs/dpctw04-03.pdf) Sull'architettura multi - sito e sulle funzioni del dark archive è fondamentale The large-scale archival storage of digital objects 2005 <http://www.dpconline.org/docs/dpctw04-03.pdf>
- [37](#) Si rinvia alla norma ISO 14721 citata.
- [38](#) An Audit Checklist for the Certification of Trusted Digital Repositories ... cit.