DigItalia

Numero 2 - **2006**

Rivista del digitale nei beni culturali

ICCU-ROMA

La raccolta dei siti web: un test per il dominio "punto it"

Giovanni Bergamin
Biblioteca Nazionale Centrale di Firenze

L'uscita della nuova legge sul deposito legale (106/2004) è stata subito seguita da una vivace discussione soprattutto per quanto riguarda la lettera r) dell'art. 4: l'estensione del deposito legale ai «documenti diffusi tramite rete informatica». Le differenti posizioni che si sono confrontate possono essere riassunte con tre aggettivi: impossibile, inutile, civile.

Il primo aggettivo – impossibile – si riferisce alla posizione che sostiene l'impossibilità tecnica di effettuare un deposito di oggetti che per loro natura non sono legati a un determinato supporto, come ad esempio un sito web. Secondo questa posizione la consegna alla biblioteca depositaria su CD o su altro supporto di copie di siti web risulterebbe essere del tutto impraticabile: alti costi di gestione per avere risultati del tutto discutibili.

Il secondo aggettivo – inutile – si riferisce invece ai contenuti del web. Secondo i sostenitori di questa posizione il web è il mondo dell'effimero e quindi assolutamente non paragonabile alle pubblicazioni tradizionali. Anche se tecnicamente esistono strumenti per il deposito legale del web, l'operazione avrebbe uno scarso risultato dal punto di vista culturale.

L'ultima posizione si riferisce invece al fatto che l'estensione del deposito legale al mondo del digitale è da considerarsi una grande conquista di civiltà: il prendere atto che oggi la conoscenza si trasmette anche attraverso i bit. In questo senso vanno le raccomandazioni dell'UNESCO del 2003 e il recente adeguamento della legislazione in molti paesi[1].

Tenendo presente che nella categoria "documenti diffusi tramite rete informatica" oggi rientrano tipologie molto differenti di risorse digitali (dal periodico elettronico con contenuti scientifici "certificati" al sito web casalingo) questo intervento intende presentare i risultati di una sperimentazione finanziata nel contesto della Biblioteca Digitale Italiana a partire da una discussione del tutto provvisoria e aperta sui presupposti e sui fondamenti delle posizioni sopra richiamate.

La prima posizione molto diffusa deriva da una scarsa conoscenza della natura della risorsa digitale. Nel mondo del digitale, l'informazione non è più legata a un determinato supporto. Si parla di "rappresentazione informatica" ovvero di «una sequen-

[1] *Charter on the Preservation of the Digital Heritage*/UNESCO, agg. 2006 http://portal.unesco.org/ci/en/ev.php-URL_ID=13366&URL_DO=DO_TOPIC&URL_SECTION=201.html.

za 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 materiale»[2]. 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 digitalizzata»[3]: la posta elettronica, le pagine web, la musica su CD, i film su DVD sono esempi di risorse digitali che negli ultimi anni hanno profondamente modificato la nostra vita e le nostre abitudini. Non senza qualche difficoltà, il concetto di documento come "rappresentazione" – non legata a un particolare tipo di supporto – è entrato anche nella legislazione: ad esempio nei meccanismi che regolano la firma elettronica il supporto fisico non è mai determinante.

Del resto la Legge 106/2004 non entra nel merito delle tecnologie utili al deposito. L'art. 5, comma 5 lettera g) rinvia al regolamento per la definizione di «speciali criteri e modalità di deposito» per i documenti h) [manifesti] q) [su supporto informatico] r) [diffusi tramite rete informatica]. Ed è evidente che con questa precisazione il termine deposito non va inteso in senso letterale (fisicamente portare un oggetto in un determinato luogo) ma nel contesto del "deposito legale" come "istituto" dove le procedure operative devono tener conto della natura dell'oggetto.

La seconda posizione si basa sulla considerazione che 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: si tratta di un *mare magnum* dove occorrerebbe procedere con una accurata selezione e non con un deposito indiscriminato.

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 dalle 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 que-

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

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

sta decisione perderemo la possibilità di archiviare diari in rete di autori citati anche nella letteratura scientifica "certificata"[4].

A livello internazionale il Consorzio fra biblioteche nazionali e Internet Archive (International Internet Preservation Consortium – IIPC)[5] è partito nel 2002 proprio dalla terza posizione (l'estensione del deposito legale al digitale come conquista di civiltà). La convinzione del Consorzio 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 Archive[6]. 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"[7] 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.

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") ;

[4] 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.

[5] http://netpreserve.org.

[6] http://www.archive.org.

[7] Per convenzione si parla di *deep web* 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 presentano le pagine HTML 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). 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 HTML. 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 web*.

— 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 dei centri di ricerca;
— criteri usati dai motori di ricerca come la frequenza di aggiornamento del sito, i link in entrata (numero di siti che citano un determinato sito) ecc.[8]

La Biblioteca Nazionale Centrale di Firenze nel corso del progetto Crawler – finanziato dalla Biblioteca Digitale Italiana – ha dato vita a 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) sono stati raccolti 7.22 terabyte di dati in formato WARC[9] (oltre i 2 milioni i server contattati per circa 240 milioni di documenti raccolti). Il software usato per il *crwaler* è Heritrix[10] (open source prodotto da 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).

[8] Julien Masanès. *Towards continuous web archiving*, «D-Lib Magazine», 8 (2002), n. 12,. http://www.dlib.org/dlib/december02/masanes/12masanes.html.
[9] 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.
[10] Informazioni sul crawler Heritrix: http://crawler.archive.org/.

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"[11].

[11] DPR n. 252 del 3 maggio 2006, art. 37.