Web archiving e ruolo della BNCF

Un'iniziativa di carattere esplorativo

Fabio Di Giammarco

Biblioteca di storia moderna e contemporanea, Roma
digiammarco@tiscali.it

In un precedente articolo,[1] occupandomi delle iniziative di *web archiving* intraprese da diverse biblioteche nazionali[2] per conservare i rispettivi spazi web, avevo messo in evidenza l'assenza, nell'elenco, dell'Italia. Infatti, nonostante l'approvazione della nuova legge sul deposito legale (l. 106/2004), in base alla quale le risorse elettroniche, e in particolare i siti web, divenivano, per la prima volta, oggetto di deposito presso le biblioteche centrali, il *web archiving* nostrano si bloccava sul nascere, a causa di una serie di perplessità di carattere amministrativo e organizzativo intorno ai modi e ai tempi della sua attuazione. Tuttavia, proprio all'indomani dell'entrata in vigore della legge, faceva la sua comparsa, all'interno del sito della Biblioteca nazionale centrale di Firenze, un comunicato dal titolo *Archiviazione dei siti web*,[3] nel quale si poteva leggere: "la legge prevede che venga emanato entro sei mesi un regolamento di applicazione, ma si può anticipare che le biblioteche nazionali stanno cooperando a livello internazionale e che concordemente indicano nell'harvesting la modalità più efficiente e sostenibile di deposito". E questo perché la BNCF, precorrendo il legislatore, aveva già da tempo iniziato a muoversi in quella direzione, allacciando rapporti a livello internazionale con l'intenzione di approfondire le conoscenze su questo tipo di tecnologia. In particolar modo nell'ambito dell'IIPC,[4] un consorzio la cui missione è "acquisire, preservare e rendere accessibile la conoscenza e l'informazione disponibile in Internet per le future generazioni"[5] e che, coordinato dalla Bibliothèque nationale de France[6] e costituito da un gruppo di importanti biblioteche,[7] può soprattutto contare sull'apporto dell'Internet Archive,[8] un'associazione non profit americana ideata da Brewster Kahle per l'archiviazione dell'intero web che dal "lontano" 1996 ad oggi ha messo a segno un risultato davvero straordinario: ben 55 miliardi di pagine web catturate. Performance resa appunto possibile dall'harvesting automatico, cioè da quel sistema di acquisizione basato su particolari software chiamati *crawlers*, che instancabilmente setacciano la rete raccogliendo pagine web sotto forma di istantanee, dette *snapshots*.

Mettendo a frutto le esperienze fin qui maturate, nel maggio scorso, la BNCF è passata all'azione: con un breve comunicato indirizzato ai webmaster[9] ha annunciato l'avvio di una prima sperimentazione, vale a dire di una prova di raccolta, affidata all'Internet Archive, per il dominio ".it". Nel frattempo, Giovanni Bergamin, responsabile dei servizi informatici dell'Istituto fiorentino, rilasciava a "Punto informatico", quotidiano di Internet, alcune interessanti precisazioni.[10] Anzitutto che l'operazione andava intesa come "primo assaggio", visto che lo spazio web italiano comprende – nella sua interezza – ovviamente numerosi domini non ".it"; poi, che si trattava di un'iniziativa, con riferimento alle attività previste da IIPC, fondamentalmente tesa "a sviluppare nuove conoscenze e metodologie per mettere a punto tecnologie in sintonia con le esigenze della rete, nonché a stimolare forme di cooperazione e coordinamento tra istituti e biblioteche con l'obiettivo di ottenere modelli di harvesting il più possibile condivisi"; e infine, che il tutto andava inteso come un'estensione delle regole sul deposito legale, tenendo però ben presente il carattere esplorativo dell'iniziativa, dovuto, nella circostanza, all'attesa approvazione del regolamento applicativo per la legge 106.

Eseguita, come da programma, nei mesi di maggio e giugno la raccolta dei siti web ".it", con la metà di agosto sembrava giunto a conclusione anche il sospirato iter normativo del deposito legale delle risorse elettroniche con la pubblicazione del regolamento applicativo.[11] Sennonché la lettura dell'atto smorzava subito i facili entusiasmi, rivelandosi per niente incoraggiante riguardo l'harvesting. Tanto per cominciare, nel primo comma dell'articolo 37[12] si rinviava "la definizione delle modalità di deposito dei documenti diffusi tramite rete informatica" a successivo regolamento! E così l'attesa conclusione dell'iter burocratico si configurava ancora una volta come un "di là da venire". Ma so-

prattutto, alla modalità automatica di raccolta era riservato, nel comma 2, soltanto un limitato cenno: "… gli accordi definiscono le modalità tecniche di deposito prevedendo, ove possibile, anche forme automatiche di raccolta, secondo le migliori pratiche e conoscenze internazionali del settore". Tutto qui. Un po' poco per chi si aspettava un diverso riconoscimento normativo per una tecnologia sempre più incentivata a livello internazionale, perché tra le poche in grado di offrire garanzie contro il rischio della perdita di un patrimonio informativo imprescindibile: la "memoria del web".
Tuttavia, nell'attesa che, con successivo regolamento, la modalità di deposito dei documenti digitali trovi – finalmente – una sistemazione definitiva, un importante risultato è stato nel frattempo ottenuto: la conclusione del primo esperimento di harvesting sul web italiano. Giovanni Bergamin, che dal 2003 segue per conto della BNCF il Consorzio internazionale di biblioteche per la conservazione di Internet e che si è occupato dell'iniziativa di raccolta del dominio ".it", ci ha gentilmente messo a disposizione i primissimi dati disponibili. Il lavoro si è svolto in un intervallo di quattro settimane, durante le quali l'Internet Archive ha setacciato, impiegando il *crawler open source* Heritix,[13] il dominio ".it", ottenendo i seguenti riscontri: più di 2 milioni di host[14] visitati, circa 2 miliardi e mezzo di documenti analizzati, per un totale di informazioni processate pari a 7,22 terabyte, vale a dire più di 7.000 miliardi di byte. Insomma, una porzione web catturata dalle dimensioni non indifferenti che, per dare un'idea, sono quasi equivalenti a quelle della più grande biblioteca del mondo: la Biblioteca del Congresso degli Stati Uniti,[15] che secondo le stime possiede un patrimonio di circa 10 terabyte di informazioni. Certo, quelli sopra elencati sono, per il momento, soltanto dati grezzi: lo "scarno" profilo numerico dell'operazione. Profusione di byte senza contenuti. Tuttavia esprimono, per la prima volta, una parziale istantanea della forma web italiana. Sicuramente un primo passo fondamentale per la BNCF verso l'assunzione, al pari delle altre biblioteche nazionali di importanti paesi, del ruolo di garante del "nostro" futuro da conservare.

Note

[1] Fabio Di Giammarco, *Conservare il futuro*, "Biblioteche oggi", 23 (2005), 2, p. 31-33.
[2] Australia, Svezia, Inghilterra, Francia, Danimarca ecc.
[3] <http://www.bnfc.firenze.sbn.it/notizie/index4.html> (all'ultimo controllo l'indirizzo non è risultato attivo).
[4] Internet Preservation Consortium (http://www.netpreserve.org).
[5] <http://www.netpreserve.org/about/mission.php>.
[6] <http://www.bnf.fr/>.
[7] The British Library, The Library of Congress, National Library of Australia, Library and Archives Canada, National and University Library of Iceland, The Royal Library of Denmark ecc.
[8] <http://www.archive.org>.
[9] Informandoli sulla presenza e sull'attività del *crawler*, e circa la possibilità di essere esclusi dall'operazione di raccolta (<http://www.bncf.firenze.sbn.it/raccolta.txt>; all'ultimo controllo l'indirizzo non è risultato attivo).
[10] <http://punto-informatico.it/servizi/ps.asp?id=1480480&r=PI>.
[11] Dpr n. 252 del 3 maggio 2006, contenente il Regolamento attuativo della legge 106, pubblicato sulla "Gazzetta ufficiale" n. 191 del 18 agosto 2006.
[12] Dal titolo: *Modalità di deposito e acquisizione dei documenti diffusi tramite rete informatica.*
[13] <http://crawler.archive.org>.
[14] Dispositivo o computer connesso direttamente alla rete e che funge da nodo di Internet.
[15] <http://www.loc.gov>.