Conservare *il futuro*

Modelli e progetti di web archiving

Fabio Di Giammarco

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

World Wide Web, Giano bifronte del sapere mondiale: fonte imprescindibile di informazioni e, nello stesso tempo, strumento dalla natura effimera, evanescente, sfuggente, come dimostrano recenti stime che valutano in soli 44 giorni la durata media di una pagina web.[1] Ciò ha indotto, tra coloro che si occupano del trattamento delle informazioni, uno stato di tale incertezza da far scattare l'allarme per "un'emergenza web". Preoccupazione comune è scongiurare che un immenso patrimonio informativo storico, qual è quello riversatosi in questi anni nella rete, vada perduto e sia, in qualche modo, conservato a beneficio non solo dei contemporanei ma soprattutto delle future generazioni. Di conseguenza, l'auspicio di tutti è che si faccia, al più presto, qualcosa. Ma cosa?

Un esempio "ragguardevole" arriva dall'iniziativa September 11 Archive,[2] messasi in moto, grazie alla collaborazione tra l'Internet Archive Project[3] e la Library of Congress,[4] subito dopo l'attacco, con aerei suicidi, alle Twin Towers di New York l'11 settembre 2001. È risultato, infatti, immediatamente chiaro come di fronte a un evento definito "assoluto", cioè in grado di cambiare il corso della storia contemporanea, il web rappresentasse una fonte di testimonianza unica, eccezionale, da preservare assolutamente. Per cui, in tempo reale, è stata avviata una gigantesca memorizzazione off-line di circa 30 mila siti, attivi nel periodo 11 settembre – 31 dicembre 2001, provenienti sia dai media che da agenzie ufficiali, governative, imprese, associazioni, ma anche da singoli cittadini e contenenti discorsi ufficiali, analisi geopolitiche, fotografie, racconti strazianti messi online da scampati all'attentato o da testimoni, siti di consigli psicologici, avvisi di ricerca di persone date per disperse ecc.

Appare evidente che il caso September 11 Archive, per la sua eccezionalità cui ha fatto riscontro una mobilitazione, a tutti i livelli, fuori del comune, esprime un modello impossibile da replicare, ma che merita grande attenzione, in quanto il suo significato, in termini generali, è trasferibile dallo straordinario all'ordinario. In altre parole, contribuisce decisamente a mettere in risalto come il web, da un punto di vista storico, sociale, intellettuale e culturale, è ormai da considerare una risorsa "ordinariamente irrinunciabile", soprattutto per gli enti deputati alla conservazione e diffusione della conoscenza e della memoria. A cominciare da quelle biblioteche nazionali di diversi paesi del mondo che hanno, appunto, intrapreso delle iniziative tese a salvaguardare i rispettivi spazi web (nazionali). Ciò, a prima vista, potrebbe sembrare in contrasto con l'assioma che vuole il web spazio informativo globale, assolutamente non circoscrivibile. Eppure il fatto che si esplichi un'azione attraverso vari progetti nazionali di preservazione del web, oltre a risultare forse al momento la strada più praticabile, potrebbe determinare, rispetto alla troppo variegata

offerta totale, una selezione delle risorse di notevole interesse.
Battistrada in questa difficile sfida è la National Library of Australia (NLA),[5] un'autorità internazionale in fatto di web archiving. Risale, infatti, al lontano 1996 l'inizio della sua attività d'archiviazione delle risorse online del web australiano, attività poi estesa all'intero campo della preservazione del digitale. Risultato di questo impegno è stato il portale web PADI (Preserving Access of Digital Information),[6] strumento di supporto online che si propone come punto di raccordo, anche con l'ausilio di un forum specializzato, per informazioni e iniziative che hanno lo scopo di facilitare lo sviluppo di strategie e linee guida per la preservazione dell'accesso alle informazioni digitali.
Le attività di web archiving riguardanti le biblioteche nazionali rappresentano l'ultima novità nell'offerta informativa disponibile su PADI. Si tratta di una serie di programmi avviati in sedici paesi,[7] che prospettano ricerche e soluzioni di vario genere, tra cui fare conto sul mandato ricevuto per il deposito legale delle collezioni elettroniche o puntare sulla possibilità di stabilire partnership internazionali attraverso le quali esplorare e testare un più ricco ventaglio d'ipotesi per la progettazione e costruzione di archivi web. Tutte queste elaborazioni teoriche e pratiche, con relativi sviluppi, sono riconducibili a una serie di modelli, che è possibile cominciare a definire nei loro tratti essenziali.
Ma prima di procedere sono necessarie alcune informazioni sui criteri e le tecniche impiegate nella raccolta e archiviazione dei siti web. Fondamentalmente, oggi si utilizzano tre sistemi:
– la selezione manuale;
– l'harvesting automatico senza selezione;
– l'harvesting automatico con parametrizzazione manuale.
Nel primo caso, l'intervento umano è totale; nel secondo esso è invece assente, a vantaggio di software chiamati *crawler,* che setacciano instancabilmente la rete e raccolgono pagine web sotto forma di istantanee (*snapshots*); nel terzo i *crawler* sono impostati per puntare su siti in precedenza selezionati come rilevanti.
Il modello che applica integralmente l'harvesting automatico viene definito a "dominio completo". È un approccio che mira a scorrere il proprio spazio web nazionale collezionando tutto il possibile: ne è un esempio il progetto Nordic Web Archive[8] (implementato da Svezia e Finlandia). Un'estensione del medesimo criterio fino alla scommessa di poter raccogliere e preservare l'intera sfera web è l'ambizioso obiettivo del progetto USA Internet Archive, che viaggia, automaticamente, a ritmi di crescita di 20 terabyte al mese.
Se invece lo scopo è quello di archiviare, secondo specifici criteri, definite porzioni dello spazio web o particolari risorse, si ha il modello selettivo. La selezione può basarsi sul significato, sulla qualità delle risorse oppure su particolari argomenti, o anche individuando un insieme specifico di siti web. È selettivo l'archivio australiano delle pubblicazioni online Pandora,[9] realizzato appunto dalla National Library of Australia nel 1996 e poi sviluppato, adoperando la selezione manuale, in collaborazione con altre biblioteche e istituzioni culturali del paese. Anche l'archivio del web britannico[10] utilizza lo stesso modello. I siti sono archiviati sulla base dei settori d'interesse afferenti alle sei istituzioni culturali che aderiscono all'iniziativa.[11] Ad esempio, la Wellcome Library[12] si occupa dei siti di medicina, la Biblioteca nazionale del Galles[13] colleziona, invece, siti che rispecchiano la vita contemporanea del proprio paese, mentre la British Library[14] attua una raccolta più generale basata su siti che rivestono una particolare importanza culturale, storica e politica. Il progetto USA Minerva[15] (Mapping the Internet Electronic Resource Virtual Archive) nato per la conservazione di particolari contenuti online, quali le tornate elettorali americane 2000 e 2002, nonché i giochi olimpici invernali del febbraio 2002, offre un esempio di web archiving tematico, cioè di una forma d'archivio, solitamente basato su materiali *born digital*, centrata su un particolare argomento o evento. Altro caso assai evidente di modello tematico, già trattato, è ovviamente il September 11 Archive.
La questione del deposito legale delle pubblicazioni online, poc'anzi accennata in termini generali, finisce anch'essa per costituire una categoria nell'ambito degli archivi web. Sono sempre di più i paesi dove si sperimentano diverse soluzioni per il deposito di contenuti web proprietari. Interessanti a questo proposito risultano le iniziative concernenti il deposito volontario. Un esempio in tal senso viene dall'Olanda,[16] grazie a un accordo tra e-journal e editori. La Royal Library di Svezia[17] è stata, invece, una delle prime biblioteche nazionali ad essere stata autorizzata, con una legge del 2002, a conservare i siti del proprio spazio web nazionale. Si segnala inoltre l'esperienza della National Library of Norway presentata nel numero scorso di "Biblioteche oggi".[18] Anche in Italia, almeno in quest'ambito, le cose si muovono. Con la nuova legge sul deposito legale (l. 106/2004) le risorse elettroniche, in particolare i siti web, sono diventati oggetto di deposito presso le biblioteche centrali. Pur tuttavia, dopo una lunghissima attesa (la precedente legge sul deposito legale risaliva al 1939) si è, per adesso, entrati in una fase di stallo giacché la nuova normativa sta suscitando, per quanto riguarda la sua ap plicazione alle risorse elettroniche, una serie di perplessità di carattere amministrativo e organizzati-

vo che attendono di essere risolte. In conclusione, il numero dei modelli fin qui descritti ("dominio completo", "selettivo", "tematico", "a deposito legale o volontario" ecc.) oltre a evidenziare diversi, possibili e interessanti approcci di web archiving, mettono altresì in luce un problema di fondo: attualmente nessuno di essi, per quanto sviluppato, è in grado di soddisfare in pieno tutte le esigenze riguardanti la preservazione dei vari patrimoni nazionali online. Questo è uno dei motivi per cui paesi come Francia e Danimarca perseguono anche altre strade, esplorando la fattibilità di sistemi combinati per l'archiviazione dei contenuti web. Il fatto è che ogni modello, per quanto raffinato, presenta inevitabilmente vantaggi e svantaggi al momento in cui si applica alle difformi tipologie di *web contents*. Evidentemente si tratta di un problema che, al di là del settore archivistico-bibliotecario, coinvolge il mondo delle università e degli istituti di ricerca. È il caso dell'Università di Heidelberg, che sta portando avanti l'interessante progetto Digital Archive for Chinese Studies,[19] oppure della Cornell University con un altro approccio assai innovativo definito VRC[20] (Virtual Remote Control) che ha lo scopo di monitorare, nel tempo, i cambiamenti dei siti web catturando quelli a rischio di perdita d'informazioni. Da uno sforzo comune, su più fronti, si attendono al più presto ulteriori buone notizie per la preservazione dei patrimoni sul web.

Note

[1] Cfr. UK Archiving Consortium, <www.archiving.org.uk>.
[2] <http://web.archive.org/collections/sep11.html>.
[3] <http://www.archive.org/>.
[4] <http://www.loc.gov/>.
[5] <http://www.nla.gov.au>.
[6] <http://www.nla.gov.au>.
[7] Australia, Austria, Canada, Repubblica Ceca, Danimarca, Finlandia, Francia, Germania, Giappone, Lituania, Olanda, Nuova Zelanda, Norvegia, Svezia, Regno Unito, USA.
[8] <http://nwa.nb.no/>.
[9] <http://pandora.nla.gov.au/index.html>.
[10] <http://www.webarchive.org.uk/>.
[11] British Library (*lead partner*), The National Archives, National Library of Wales, National Library of Scotland, JISC, Wellcome Trust.
[12] <http://www.wellcome.ac.uk/>.
[13] <http://www.llgc.org.uk/>.
[14] <http://www.bl.uk/>.
[15] <http://www.loc.gov/minerva/>.
[16] <http://www.kb.nl/kb/resources/frameset_kenniscentrum-en.html>.
[17] <http://www.kb.se/ENG/kbstart.htm>.
[18] *Paradigma Project: il deposito legale delle risorse remote nell'esperienza norvegese*, 13 (2004), 1, p. 17-28.
[19] <http://www.sino.uni-heidelberg.de/dachs/>.
[20] <http://irisresearch.library.cornell.edu/VRC/>.