3 Una nota polemica sul web archiving in Italia

Federico Mazzini
Università degli Studi di Padova

DOI: 10.25430/pupb-9788869384394-10

L'Italia è uno dei pochi paesi occidentali a non avere in essere un sistematico sforzo di archiviazione del web (web archiving). In questa nota l'autore riflette sulle cause di questo ritardo e sulle sue ripercussioni sulla storiografia e su altri aspetti della cultura digitale, sottolinea l'imprescindibilità dell'intervento istituzionale e dell'interessamento accademico nel web archiving e suggerisce alcuni semplici accorgimenti che possono essere intrapresi a livello individuale.

Italy is one of the few Western countries that does not have in place a systematic web archiving effort. In this note, the author reflects on the causes of this delay and its repercussions on historiography and other aspects of digital culture. The author also stresses the importance of institutional intervention and academic interest in web archiving, and suggests some simple steps that can be taken by individuals.

Archiviazione del web, Wayback Machine, Internet Archive, Fonti nate digitali
Web archiving, Wayback Machine, Internet Archive, Born-digital sources

Gli articoli scientifici sul web archiving (la pratica di archiviare i siti web attraverso software che salvano il codice HTML e lo restituiscono in una simulazione del web "*live*") sono stati spesso accomunati da argomentazioni ricorrenti: 1) la spiegazione del perché il web archiving è importante (se tanta parte della nostra cultura si esprime online, va da sé che il web sarà una fonte storica imprescindibile); 2) la spiegazione di diverse tecniche di raccolta (prima tra queste, sempre, quella "a strascico" proposta da Internet Archive, seconde, spesso, quelle più mirate condotte dalle Biblioteche Nazionali); 3) l'appello ad archivisti e storici perché si facciano coinvolgere nella conservazione e nell'uso di fonti web, e perché lo facciano presto, data la loro natura effimera e il continuo crescere della produzione di artefatti culturali in forma esclusivamente digitale. Io stesso ho scritto, qualche anno fa, un articolo introduttivo che segue grossomodo questa

struttura, e a questo rimando per chi volesse dettagli sulla storia dell'archiviazione web e sui metodi principali fino ad allora adottati[1].

Recentemente, tuttavia, il tono e gli obiettivi del dibattito internazionale sono cambiati[2]. La sfida del web archiving non ha perso la sua urgenza, ma gli iniziali successi e la diffusione della pratica portano a interrogarsi su questioni più complesse, che riguardano la metadatazione, la coerenza dei corpora e il rapporto tra la natura delle fonti e la ricostruzione storica. Gradualmente si sta abbandonando il tempo futuro ("il web archiving sarà fondamentale…") in favore di quello passato, che riconosce l'esistenza di studi storici basati su fonti web già pubblicati e parte di un crescente dibattito sulla storia del web. Iniziano a vedere la luce libri di testo e breviari di introduzione alla pratica di raccolta, e persino approcci sperimentali e artistici alla conservazione[3]. Il campo ha raggiunto, insomma, una certa maturità e legittimità.

Non così in Italia. Il tema del web archiving è virtualmente assente dal dibattito storiografico, anche nei luoghi, come le associazioni degli storici contemporaneisti o di quelli dei media, dove il tema dovrebbe essere maggiormente sentito. I pochi meritevoli interventi di studiosi italiani provengono da archivisti e bibliotecari[4]: le opere storiche in italiano che si basano, anche solo in parte, su fonti "nate digitali" si contano sulle dita di una mano. Il passato digitale nazionale, per quanto possa essere fatto risalire perlomeno alla metà degli anni ottanta, con la diffusione dei primi Bullettin Board Systems e la nascita di un precoce attivismo telematico, è una terra incognita nella quale solo occasionalmente si avventurano memorialisti, giornalisti e volenterosi laureandi.

Non sarebbe la prima volta che la storiografia italiana si muove in ritardo o in divergenza rispetto a quella internazionale, con motivazioni non sempre illegittime e con risultati non sempre necessariamente negativi. Ma in questo caso la natura effimera delle fonti, denunciata dalla storiografia internazionale fin dalla fine degli anni novanta, fa la differenza. Le mancanze di oggi non potranno essere colmate dall'impegno di domani, o da approcci alternativi. Alcune porzioni del nostro passato digitale sono già irrimediabilmente perse, e dovranno essere ricostruite con mezzi indiretti. Altre si sono salvate per caso, generalmente grazie al fatto di essere rimaste impigliate nelle raccolte "a strascico" di enti internazionali. Altre ancora sono ancora online, o negli hard disk di privati e istituzioni, e sono a continuo rischio di scomparsa. Se la salvaguardia del nostro web storico era urgente all'inizio del millennio, è ora una vera e propria emergenza.

[1] Federico Mazzini, *I semi e il raccolto. Archiviazione del web e ricerca storica*, in *La storia in digitale. Teorie e metodologie*, a cura di Deborah Paci, UNICOPLI, Milano 2019, pp. 145-159.

[2] Si vedano ad esempio i saggi contenuti in Daniel Gomes, Elena Demidova, Jane Winters, Thomas Risse (a cura di), *The Past Web: Exploring Web Archives*, Springer International Publishing, Cham 2021.

[3] Marijn Josephien Bril, *Performatively Archiving the Early Web: One Terabyte of Kilobyte Age*, «VIEW Journal of European Television History and Culture», 5 settembre 2023, 12, fasc. 23, pp. 69-85.

[4] Lorenzana Bracciotti, *Il Web Archiving. Conservazione e uso di una nuova fonte*, «Officina della Storia», 2018, fasc. 19, <https://web.archive.org/web/20190522100050/https://www.officinadellastoria.eu/it/2019/01/10/il-web-archiving-conservazione-e-uso-di-una-nuova-fonte/>; Chiara Storti, *"Resource not found": cultural institutions, interinstitutional cooperation and collaborative projects for web heritage preservation*, «JLIS.it», 15 maggio 2023, 14, fasc. 2, pp. 39-52; Stefano Allegrezza, *Web e social media come nuove fonti per la storia*, «Umanistica Digitale», 2022, fasc. 14, pp. 137-162.

L'Italia è infatti tra i pochi paesi occidentali che non ha in essere un progetto di web archiving del proprio dominio nazionale (.it) o di importanti parti di esso – un compito in altri paesi affidato alle Biblioteche Nazionali sulla base delle leggi sul deposito legale. Tale strada è stata tentata anche in Italia: già nel 2006 un decreto del Presidente della Repubblica apriva una fase di sperimentazione per il deposito «dei documenti diffusi tramite rete informatica», con particolare attenzione alle pagine prodotte da istituzioni culturali e pubbliche (ivi comprese le università) e ai «documenti relativi a siti che si aggiornano con più frequenza, ovvero contenuti in siti che sono maggiormente citati da altri siti»[5] (vale a dire quelli più popolari secondo i criteri adottati dai motori di ricerca). Il risultato di questa sperimentazione, mai chiusa ma forse mai veramente avviata, è, a più di quindici anni di distanza, desolante. La Biblioteca Nazionale Centrale ha effettuato un singolo salvataggio nel 2006 dell'intero dominio .it[6]. Oggi la Biblioteca offre la possibilità di aderire volontariamente a un programma di archiviazione che fa uso del servizio Archive-it messo a disposizione da Internet Archive. La possibilità è aperta soltanto alle istituzioni pubbliche e culturali, che devono fare richiesta di adesione. Tra il 2018 e il 2020 la BNCF ha salvato, secondo gli stessi promotori, soltanto 250 siti web (su circa 3.300.000 domini registrati nel dominio .it nel 2020, quasi 3.500.000 oggi)[7]. Questo è in netto contrasto con le pratiche più diffuse del web archiving internazionale, che vedono le istituzioni archivistiche raccogliere attivamente milioni di pagine web senza chiedere l'adesione, inevitabilmente sporadica soprattutto se il servizio non è adeguatamente pubblicizzato, dei gestori dei siti web. Non è dato peraltro sapere quali siano i dati raccolti dopo il 2020 (o quali dati siano disponibili in accesso limitato), ma un'occhiata alla sezione Archive-it curata dalla BCNF[8] sembra suggerire che lo sforzo si sia arenato.

Ovviamente gli archivisti e i bibliotecari della BCNF non hanno altra colpa che quella, forse, di un eccessivo ottimismo nella presentazione del progetto[9]. Le responsabilità sono della politica e dell'accademia. Della politica, perché non ha saputo andare oltre ambigue e isolate indicazioni di principio, verso studi seri di esperienze estere, iniziative adeguatamente finanziate e un regolamento tecnico che, per quanto previsto dal DPR del 2006, ancora manca[10]. Dell'accademia, perché non ha saputo provare (o forse ancora percepire) l'importanza della salvaguardia del web, attraverso la formulazione di domande sul nostro passato digitale che solo negli archivi web possono trovare risposta. Una soluzione strutturale e di lungo periodo non può dunque che venire dai legislatori e dalle discipline storiche.

5 <https://www.normattiva.it/eli/id/2006/08/18/006G0272/ORIGINAL>.

6 Giovanni Bergamin, *La raccolta dei siti web: un test per il dominio "punto it"*, «Digitalia», 2006, fasc. 2, pp. 171-174.

7 <https://web.archive.org/web/20231004155808/https://stats.nic.it/domain/growth>.

8 <https://archive-it.org/home/BNCF>.

9 <https://web.archive.org/web/20231004115759/https://www.bncf.firenze.sbn.it/biblioteca/web-archiving/>.

10 Chiara Storti, *Web archiving: il servizio della Biblioteca Nazionale Centrale di Firenze*, «FPA (blog)», 12 giugno 2019, <https://www.forumpa.it/pa-digitale/gestione-documentale/web-archiving-sfida-culturale-il-servizio-della-biblioteca-nazionale-centrale-di-firenze/>.

Ma è anche possibile, per chiunque sia responsabile di una ricerca o di un insegnamento di storia, di un database o di un semplice sito web, agire in prima persona e immediatamente. In primo luogo, insegnando agli studenti e ai colleghi la sensibilità verso la longevità dei materiali nati digitali. Se le regole editoriali imponessero che ogni sito citato in uno scritto scientifico, in una tesi o in una semplice esercitazione in classe appaia non nella sua forma "*live*" ma in una forma archiviata (come i link in questo testo), ben presto avremmo una imponente presenza di siti italiani o di interesse per l'accademia italiana conservati in archivi storici. Dato che il salvataggio di singole pagine all'interno di servizi come Wayback Machine di Internet Archive[11] è gratuito (e anche automatizzabile attraverso apposite estensioni del browser) non vi è alcun motivo per il quale non si debba cominciare già da ora a richiederlo come parte di ogni testo che ci viene consegnato.

In secondo luogo, ogni progetto scientifico dovrebbe dedicare parte del proprio budget alla longevità dei propri prodotti digitali, esplicitando fin dal principio quale sarà loro destino una volta che il progetto e i suoi fondi si siano esauriti. Questo è molto meno complicato o dispendioso di quanto possa a prima vista apparire. Un salvataggio completo di un sito web attraverso Archive-it costa poche centinaia di euro e non richiede alcuna competenza da parte del responsabile del progetto. Ma anche raccolte più estese richiedono un impegno finanziario relativamente limitato, nell'ordine di alcune migliaia di euro per centinaia di giga conservati. È così possibile immaginare che un progetto che intenda investigare, ad esempio, il dibattito ambientalista in Italia tra gli anni '90 e oggi includa nei propri "*output*" un archivio web dedicato ai siti che in quegli anni sono individuati come più significativi[12]. Vero è che un approccio minimalista di questo genere lascia aperte molte questioni fondamentali: la metadatazione, la formazione degli storici sull'uso qualitativo e quantitativo di fonti web, la ricreazione imperfetta del web storico, le inevitabili lacune di una raccolta anche limitata a un singolo argomento. Problematico è anche l'affidarsi a una fondazione privata, per quanto animata dalle migliori intenzioni, come Internet Archive, e non a una istituzione pubblica. Ma questo minimo sforzo permette perlomeno di salvare una parte del codice HTML che racconta il nostro presente e recente passato, e lo apre a future contestualizzazioni e all'analisi via software.

Occorre in ultimo sottolineare che, sebbene io mi sia in questo breve intervento concentrato sulla storiografia, l'importanza del web archiving non è ad essa limitata. In un ecosistema informativo nel quale i testi e le immagini appaiono e scompaiono senza posa, il web archiving costituisce un elemento di verificabilità e tracciamento imprescindibile. La Wayback Machine (o servizi analoghi)[13] può essere usata nella navigazione quotidiana, per recuperare pagine che ricordiamo di aver visto, ma che sono or-

[11] Wayback Machine è l'interfaccia che permette di navigare le pagine web "storiche" salvate da Internet Archive. Per aggiungere una pagina di proprio interesse è sufficiente inserirla all'indirizzo <https://archive.org/web/>, sezione "Save Page Now". Si potrà poi usare il link fornito in luogo di quello "*live*" nella citazione.

[12] <https://archive-it.org/collections/21520>.

[13] Ad esempio <https://perma.cc>; <https://archivebox.io>; <https://archive.is> e tanti altri. Nessuno di questi, a mia conoscenza, ha la facilità d'uso e l'ambizione archivistica di Internet Archive.

mai scomparse (chi non ha mai incontrato la pagina di errore 404?) o che sono state radicalmente modificate. Il suo uso è uno dei modi migliori per contrastare il *link rot*, il fenomeno che vede la maggior parte dei link esterni citati nei siti web o negli scritti scientifici non essere più funzionanti (poiché spostati o modificati) a distanza di pochi anni. Le pagine archiviate sono sempre più frequentemente usate come elemento probatorio in tribunale, tanto che Internet Archive offre procedure di autentificazione legale del codice HTML salvato. Non è infine difficile immaginare che l'enorme quantità di testo contenuta negli archivi web sia stata e sarà usata per il training dell'intelligenza artificiale basata su modelli linguistici (ChatGPT e i suoi equivalenti). I ritardi e le lacune della conservazione del web in lingua italiana hanno inevitabili ripercussioni in tutti questi campi.