JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

JLIS.it

"Resource not found": cultural institutions, interinstitutional cooperation and collaborative projects for web heritage preservation

Chiara Storti[(a)]

a) Biblioteca Nazionale Centrale di Firenze

Contact: Chiara Storti, chiara.storti@cultura.gov.it
Received: 31 January 2023; **Accepted:** 02 March 2023; **First Published:** 15 May 2023

ABSTRACT
Awareness of the need to preserve the Web heritage began to spread in the early 1990s, and the first Web archiving initiatives were launched at the National Library of Australia in 1993-1996. However, even today, especially in Italy, it is difficult to find contributions that address the topic of Web and Social media archiving from the perspective of process governance and its relationship with the laws and ethical issues. The purpose of this article is to investigate which actors are involved in these processes, what are the main criticalities or innovations compared to "traditional" cultural heritage management, providing concrete examples where possible.

KEYWORDS
Web archiving; Social Media archiving; Digital legal deposit; Internet Archive.

"Risorsa non trovata": istituzioni culturali, cooperazione interistituzionale e progetti collaborativi per la conservazione del patrimonio web

ABSTRACT
Vedersi restituire l'errore "risorsa non trovata" è un'esperienza comune a tutti coloro che cercano informazioni sul Web. La conservazione del patrimonio web è, infatti, una delle sfide della società contemporanea. Il presente contributo si propone non tanto di analizzare le ormai consolidate tecnologie per il Web e Social media archiving ma piuttosto il piano della governance complessiva dei processi di archiviazione e quello del quadro legislativo di settore non solo a livello nazionale, ma anche nel rapporto con il sovranazionale. Laddove possibile riportando casi reali, si cerca inoltre di definire chi siano i diversi soggetti coinvolti e responsabili della costruzione di un'eredità digitale nazionale che possa essere fruita e compresa compiutamente dalle prossime generazioni.

PAROLE CHIAVE
Web archiving; Social media archiving; Conservazione digitale; Deposito legale.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

Negli Stati Uniti e in Australia una sensibilità verso la necessità di conservare le risorse web ha cominciato a diffondersi già a partire dall'inizio degli anni '90 del '900. Anche in Italia, seppure con evidente ritardo, sono ormai quasi venti anni che si pubblicano contributi che trattano di Web e Social media archiving[1] su riviste e collane specializzate nell'ambito delle Scienze archivistiche e biblioteconomiche, delle Digital Humanities e dell'Informatica applicata. Negli ultimissimi anni anche alcune testate giornalistiche generaliste[2] o divulgative specializzate[3] hanno portato l'argomento all'attenzione di un pubblico più vasto, di frequente partendo da fatti di cronaca quali la chiusura o il cambiamento di management di un sito o servizio online. Tuttavia, almeno in ambito nazionale, è difficile reperire contributi che integrino i diversi punti di vista da cui il Web e Social media archiving può essere compreso, o almeno problematizzato. Nella letteratura specializzata viene esplorato principalmente negli aspetti che riguardano le tecnologie, cercando talvolta di individuare quelle che siano maggiormente in grado di restituirci un prodotto che non risulti "estraneo" alla nostra tradizione storico-culturale e archivistica-biblioteconomica. A questo approccio è legata una prospettiva "istituzionale" o "pubblica". D'altra parte, su testate più generaliste, oltre che sulle tecnologie, ci si concentra sulle implicazioni politiche e sociologiche della perdita o della conservazione di significative moli di dati e informazioni, citando le responsabilità delle grandi aziende del settore tech.

In tutti i casi viene quasi completamente omesso il piano della governance complessiva dei processi di archiviazione e quello del quadro legislativo di settore non solo a livello nazionale, ma anche nel rapporto con il sovranazionale.

L'Italia, rispetto alla maggioranza dei Paesi occidentali, arriva tra gli ultimi a porsi il problema di una strategia nazionale di conservazione del Web ma, proprio per questo, si può avvalere dell'esperienza di quasi tre decenni di iniziative di Web archiving. In particolare, è ormai chiaro come cercare di applicare alle risorse del Web le stesse categorie ontologiche e le stesse procedure di conservazione e accesso finora utilizzate per le risorse tradizionali, o per quelle digitali non diffuse su Web, rischi di farci perdere una grande quantità di informazioni e di restituire alle generazioni future un'immagine distorta o non completa del nostro presente. È pertanto necessario trovare soluzioni che prendano in considerazione non solo le tecnologie ma soprattutto la governance dei processi, e che questi processi siano indirizzati a preservare non solo gli oggetti in sé ma anche il contesto storico, politico e sociologico in cui le risorse sono state create, nel rispetto di vincoli normativi ed etici. Appare altresì inevitabile definire più chiaramente ruoli e forme di collaborazione tra le istituzioni del patrimonio, i proprietari e gestori delle piattaforme, fino ai produttori di contenuti: pubblico o privato, singolo o azienda. In questa sede si prova ad indagare quali siano i diversi soggetti coinvolti, laddove possibile riportando casi reali, per costruire un'eredità digitale nazionale che possa essere fruita e compresa compiutamente dalle prossime generazioni.

[1] Per una parziale bibliografia in italiano si vedano: https://www.bncf.firenze.sbn.it/biblioteca/web-archiving/#Contributi%20online%20in%20italiano e https://site.unibo.it/web-and-social-media-archiving-and-preservation/it/strumenti/pubblicazioni.

[2] Si segnalano in particolare gli articoli de IlPost.it: https://www.ilpost.it/tag/internet-archive/.

[3] Si segnalano in particolare gli articoli di Wired.it. Significativo anche che il tema sia stato trattato di recente in una puntata del podcast "CRASH-La chiave per il digitale" dal titolo "Errore 404, così il web sta perdendo i pezzi": https://open.spotify.com/episode/7iYIin0mYnzkCKporF9pjN?si=Xod5DaFURlihnvb782JzTA.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

"L'invenzione del Web archiviato"[4]

Si è soliti far risalire le origini del Web archiving all'ottobre del 1996, quando gli ingegneri di Internet Archive lanciarono il primo web *crawler* allo scopo di catturare tutto il Web allora esistente.[5] Il poi divenuto celebre "errore 404" (Parlangeli 2017) era infatti già molto diffuso anche se, a causa dell'ancora scarsa familiarità della maggioranza delle persone con Internet, non era ritenuto un vero problema. Brewster Kahle e Bruce Gilliat ne intuirono ugualmente la portata e misero a punto un sistema per copiare e conservare le pagine web prima che sparissero. Dal 2001, le pagine web oggetto di periodica archiviazione non solo da parte di Internet Archive, ma anche di numerosi altri partner pubblici o privati che utilizzano i loro strumenti,[6] sono liberamente ricercabili e consultabili tramite la *Wayback Machine*.[7] Quasi contemporaneamente a quel primo web *crawling* massivo ad opera di Internet Archive, altre istituzioni della memoria, in particolare biblioteche e archivi nazionali, iniziarono *harvesting* selettivi delle informazioni in rete,[8] sulla base della loro appartenenza ad un determinato dominio[9] o gruppo di domini o per argomento. Servizi che nel tempo le comunità nazionali o le comunità di interesse hanno meglio definito per conservare e rendere accessibile il patrimonio web. Il web *crawling* o web *harvesting* è ancora la tecnologia più usata per la copia delle pagine web, così come nel frattempo si sono consolidati il formato WARC[10] per l'archiviazione e la *Wayback Machine* per il *replay* delle copie archiviate. Tuttavia, in un articolo pubblicato su *Internet Histories*, Kieran Hegarty (2022) ha riportato i risultati di una ricerca effettuata alla National Library of Australia (d'ora in poi NLA) sul periodo compreso tra il 1993 e il 1996, ovvero su quella che l'autore definisce la preistoria del Web archiving. In quei primissimi anni '90 del '900, infatti, oltre a singole iniziative di piccoli gruppi amatoriali che provarono a salvare alcune risorse web[11] per specifiche finalità, di solito in file HTML o con screenshot, alla NLA si andava costituendo il gruppo di lavoro che avrebbe portato alla creazione del programma nazionale australiano di Web archiving "PANDORA".[12] Dall'indagine sulle testimonianze documentali e dalle interviste ai protagonisti dell'epoca emerge come il fatto che le prime infrastruttu-

[4] Il titolo del paragrafo è liberamente tradotto da Hegarty (2022).

[5] La video intervista ad un giovane Brewster Kahle, allora impegnato nel primo web *crawling*, è stata pubblicata nel 2021 da Internet Archive in occasione del 25° anniversario dell'evento: https://archive.org/details/wayback-machine-1996.

[6] All'interno della Wayback Machine sono indicizzate anche tutte le url archiviate e definite "pubblicamente accessibili" dalle istituzioni della memoria che utilizzano la suite "Archive-it" https://archive-it.org/, oltre che le url archiviate grazie a iniziative particolari di Web archiving come quelle realizzate dall'Archive Team e di cui si dirà più avanti.

[7] https://archive.org/web.

[8] Un elenco sufficientemente esaustivo di queste iniziative è la *List of Web archiving initiatives* su en.wikipedia: https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

[9] L'unica copia integrale del dominio .it risale al 2006 per iniziativa della Biblioteca Nazionale Centrale di Firenze che l'ha resa di recente disponibile nella propria collezione di Archive-it: https://archive-it.org/collections/15697. Per approfondimenti (Bergamin 2006).

[10] https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.

[11] In questo stesso periodo si colloca anche l'*Occasio project*, uno dei primi progetti pionieri di Social Media archiving, lanciato nel 1994 dall'International Institute of Social History di Amsterdam, allo scopo di conservare le conversazioni a tema sociale e politico postate, tra il 1988 e il 2002, nei gruppi di discussione online. L'archivio non è più disponibile online ma se ne possono recuperare alcune pagine su Internet Archive: https://web.archive.org/web/20020602111206/http://newsarchive.occasio.net/.

[12] https://pandora.nla.gov.au/.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

re[13] per il Web archiving siano state costruite all'interno di biblioteche nazionali abbia fortemente influenzato e ancora influenzi le modalità con cui si sceglie cosa e come archiviare, in ultimo determinando, quasi sempre inconsapevolmente, l'immagine che creiamo del passato recente e come questo potrà essere conosciuto in futuro.
L'assunto da cui parte Hegarty è che nessuna infrastruttura può essere creata dal nulla ma che ogni nuova infrastruttura si innesta su infrastrutture preesistenti, a volte ampliandole e migliorandole, altre semplicemente adattandole a nuovi scopi. E il tipo di infrastruttura su cui un nuovo servizio viene innestato ne determina quasi sempre le caratteristiche fondamentali.

> When considering the history of web archiving, is important to reflect on why most web archiving institutions are major public libraries (Gomes et al. 2011). Indeed, the dominant web archiving institution, the Internet Archive, considers itself a library and its founder Brewster Kahle refers to himself as a librarian. Here, the notion of the public record is critical. Libraries lean on their responsibility to collect, preserve, and provide access to the published output (usually of a nation) when incorporating a selection of publicly available web within their collections (Brügger 2016). By asserting the material they collect from the web is publicly available, libraries position web archiving in terms of continuity with past collecting practices (Rauber et al. 2008). This positioning allows web archiving systems and practices to build on top of, extend, or anticipate frameworks that major public libraries have long relied upon to develop their collections, particularly legal deposit exemptions that call on publishers to deposit a copy of any publications to the library. Positioning web material as "publications" and website producers as "publishers" is therefore a key assumption that animates web archiving infrastructure and sees its actors and materials largely situated within major public libraries (Hegarty 2022).

In altre parole, il fatto che sia stato proprio all'interno di una biblioteca nazionale che si sia iniziato a ragionare di Web archiving come infrastruttura ha ancora oggi conseguenze su cosa vogliamo o possiamo conservare. I siti web infatti sono stati da subito considerati come un mezzo attraverso cui diffondere "pubblicazioni" e i produttori delle informazioni alla stregua di "editori", potendo in questo modo interpretare il Web archiving come un'attività senza soluzione di continuità con il deposito legale. Ciò da una parte ha permesso alle biblioteche di salvare milioni di risorse web, dall'altra ha fortemente influenzato i criteri e le modalità di selezione delle risorse da conservare, escludendo o non analizzando adeguatamente tutto ciò che non era riconducibile all'interno di questo canone riconosciuto. Quando ci muoviamo nell'ecosistema del Web, infatti, i confini tra i concetti di "pubblicato", "pubblicamente accessibile", "destinato all'uso pubblico" o "di interesse pubblico" si fanno sempre più sfumati, ma continuano ad esistere, e collassarli nell'aggettivo "pubblico" può certamente rivelarsi comodo ma anche rischioso. Allo stesso modo decidere che tutti coloro che gestiscono una piattaforma o producono contenuti per il Web siano "editori", con tutti i diritti ma soprattutto i doveri che questo *status* comporta, dovrebbe essere se non altro oggetto di una qualche forma di contrattazione, se non meglio formalmente inquadrato a livello normativo; e non può essere una decisione unilaterale delle istituzioni culturali, pena l'inefficacia di questo as-

[13] Hegarty (2022) chiarisce che in questo contesto con infrastrutture si debbano intendere proprio le modalità con cui sono "assemblate" ed utilizzate le singole tecnologie, anche sviluppate da terzi, in una specifica realtà come quella di una biblioteca nazionale, per perseguire le proprie finalità istituzionali.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

sunto e l'impossibilità di perseguire le finalità pubbliche per cui quelle stesse istituzioni operano. Inoltre, la scelta di definire i siti web come *medium* per le "pubblicazioni", al fine di ricondurli ad una tipologia bibliografica nota ha, almeno per tutti gli anni '90, limitato molto ciò che è stato conservato. L'impossibilità addirittura, per molto tempo, di descrivere attraverso i tradizionali standard catalografici il formato "sito web" ha comportato il sacrificio di una grandissima quantità di informazioni, quasi come se ciò che non potesse essere descritto, in qualche modo non fosse conservabile.

Solo il cambio di paradigma, che ha portato dalla ricerca di "pubblicazioni" alla ricerca di "informazioni" diffuse sul Web, ha permesso di uscire dalla preistoria del Web archiving e di iniziare a pensare a tali risorse come qualcosa di completamente nuovo e per questo meritevole di uno specifico trattamento.

La Library of Congress, a partire dal 2002, definisce all'interno delle *AACR – Anglo-American Cataloguing Rules* i siti web come *"integrating" resource,*[14] continuando ad assimilarli ad oggetti bibliografici noti come le pubblicazioni a dispense o fogli mobili, ma anche potendo finalmente rilevare una delle caratteristiche principali di una risorsa web ovvero il suo, a volte molto frequente a volte meno, aggiornamento.[15] La possibilità di aggiornamento è però solo una delle caratteristiche di queste risorse. Si potrebbe affermare, infatti, che i siti web siano risorse sostanzialmente bibliografiche ma archivistiche nella forma. Riprendendo la terminologia archivistica classica, infatti, un sito web è spesso diretta emanazione del suo soggetto produttore (ente, azienda, persona ecc…), e le sue diverse componenti possono essere individuate singolarmente ma hanno un "vincolo necessario". Allo stesso modo possono essere oggetto di aggiornamenti o modifiche anche minori, che necessitano di un "controllo di versione", ma la cui entità potrebbe non giustificare catalograficamente la definizione di "nuova edizione". D'altra parte i contenuti informativi mediati dai siti possono essere spesso considerati di "interesse culturale" e "destinati all'uso pubblico", riprendendo la definizione che la normativa italiana sul deposito legale[16] fornisce delle risorse destinate alle biblioteche, e agli altri istituti individuati, per la creazione degli archivi (!) nazionali e regionali. Anche l'OCLC – Web Archiving Metadata Working Group (WAM) nel documento *Descriptive*

[14] Nel *RDA Glossary* una *"integrating" resource* viene definita: *a resource that is added to or changed by means of updates that do not remain discrete and are integrated into the whole (e.g., a loose-leaf manual that is updated by means of replacement pages, a website that is updated continuously).* Si veda anche: https://web.library.yale.edu/cataloging/e-resources/updating-websites. In Italia, ancora le recenti norme per il "Trattamento in SBN di risorse pubblicate online" non considerano i siti web come oggetti catalografici autonomi ma solo come media per la diffusione di altre pubblicazioni: https://norme.iccu.sbn.it/index.php?title=Norme_comuni/Ulteriori_indicazioni_e_approfondimenti/Trattamento_in_SBN_di_risorse_pubblicate_online. La motivazione è banalmente da ricondursi alla constatazione che nelle istituzioni del patrimonio non si è ancora sviluppata una sufficiente sensibilità su questi temi e che un'attività istituzionale di Web archiving è portata avanti solo dalla Biblioteca Nazionale Centrale di Firenze. Su questo tema si veda anche (Allegrezza 2023).

[15] Anche la tradizionale modalità di accesso alle risorse archiviate tramite Wayback Machine rispecchia la visione di un archivio web come di una collezione di pubblicazioni in evoluzione: "this imagining of a web archive as a collection of 'evolving' publications endures in the many web archives that use Wayback for the replay of content, where access is given at the level of a single 'title' (webpage) along with a list of 'issues' (snapshots)" (Hegarty 2022).

[16] L. 15 aprile 2004, n. 106 *Norme relative al deposito legale dei documenti di interesse culturale destinati all'uso pubblico*: http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2004-04-15;106!vig=2023-01-11 e D.P.R. 3 maggio 2006, n. 252, Regolamento recante norme in materia di deposito legale dei documenti di interesse culturale destinati all'uso pubblico: http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.del.presidente.della.repubblica:2006-05-03;252!vig=2023-01-11.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

Metadata for Web Archiving[17], contenente alcune indicazioni per la formulazione di metadati descrittivi per le risorse catturate dal Web, propone una soluzione "ponte" tra i due più tradizionali approcci della descrizione archivistica e della catalogazione bibliografica/bibliotecaria. Lo standard di struttura, il Dublin Core, è neutrale rispetto al dominio di applicazione e il set minimo di elementi è stato scelto mettendo a confronto i principali standard in uso per la descrizione degli oggetti digitali di interesse per gli istituti della cultura e della ricerca: MARC21, EAD, MODS, Dublin Core e Schema.org. Inoltre la soluzione proposta può essere utilizzata a tutti i livelli descrittivi, in maniera scalabile, da un dominio di alto livello al singolo documento, in qualsivoglia formato, pubblicato su una pagina web.
Si potrebbe continuare ancora a lungo con esempi che avvalorino il fatto, per altro mai messo in dubbio fuori dalle istituzioni del patrimonio, che il Web sia un media informativo completamente diverso da quelli finora trattati e che la sua gestione resti un'attività di confine tra le competenze degli archivi e delle biblioteche, ma pare maggiormente proficuo provare da ciò a trarre delle proposte operative. Un tale *excursus* storico serve infatti a ribadire quanto siano necessari da una parte una fattiva collaborazione tra istituzioni della memoria, in particolare archivi e biblioteche, e dall'altra una normativa nazionale che tratti in maniera organica il tema della conservazione del patrimonio digitale definendo le responsabilità, anche dei produttori di informazione, gli ambiti di esclusività e quelli di collaborazione. Oltre che definendo in concreto i rapporti tra le necessità di conservazione per finalità storiche, culturali e di ricerca e i sempre più pressanti interrogativi etici riguardanti ad es. il diritto all'oblio.
Infine, come si vedrà meglio di seguito, pare altrettanto importante formare gli utenti ad un uso consapevole degli archivi web:

> Reflecting on web archiving as a story of continuity as well as discontinuity, this article is a call for a greater account of the strengths and limitations of web archiving infrastructure's "installed base"—the major public library—and the myriad artefacts collected throughout these libraries' histories. This can help users of these sources to approach web archives not as a neutral window into the web's past, but a continuation of an inherently partial attempt to mobilise a representation of publics in archival form (Hegarty 2022).

In caso di conflitto

In seguito alla drammatica riconquista dell'Afghanistan da parte dei talebani nell'agosto del 2021, il Partito Repubblicano statunitense ha cancellato un post, pubblicato sul proprio sito il 15 settembre 2020, in cui venivano elogiati i risultati dell'accordo promosso dall'ex presidente Trump con i talebani stessi. Grazie alla *Wayback Machine* di Internet Archive è possibile vedere diverse istantanee del sito antecedenti all'agosto del 2021 e leggere l'articolo successivamente eliminato.[18]
Il ruolo di un archivio web in caso di conflitto non si riduce, però, a quello di mostrare l'incoerenza del politico o del personaggio pubblico di turno che, come spesso accade, modifica o cancella post scomodi. Come è purtroppo noto, infatti, il restaurato governo talebano ha da subito iniziato

[17] https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html.
[18] https://web.archive.org/web/*/https://gop.com/president-trump-is-bringing-peace-to-the-middle-east-rsr/.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

una campagna persecutoria nei confronti di molti cittadini afghani accusandoli di "collaborazionismo" con il governo statunitense, ad esempio perché impiegati all'interno delle istituzioni americane o in associazioni umanitarie sul territorio. I talebani hanno passato al vaglio centinaia di profili social alla ricerca di post, foto ecc…che avvalorassero tali scellerate accuse. Le piattaforme social, per contro, hanno tempestivamente oscurato, su richiesta degli interessati, i profili e ogni altra traccia online (Tuttosport.com. 2021) potenzialmente dannosa per la loro incolumità. Ma ciò rischiava di non essere sufficiente. In particolar modo Twitter[19] ha richiesto la collaborazione di Internet Archive per rendere inaccessibili anche le copie archiviate dei profili degli afghani perseguitati, almeno fino a quando la situazione interna del Paese non ne consentirà il ripristino. L'inaccessibilità di queste informazioni, per ora a tempo indeterminato, è un esempio di come giustamente le ragioni della incolumità fisica delle persone possano e debbano essere poste in primo piano rispetto al diritto di cronaca o alla ricerca storica.

D'altronde che il Web e, di conseguenza, gli archivi web siano oramai da considerarsi, da molteplici punti di vista, infrastrutture strategiche per un Paese alla stregua di strade, ponti o ospedali, lo dimostra anche il drammatico conflitto in corso a seguito dell'invasione russa dell'Ucraina nel febbraio del 2022. Nell'ultimo anno, all'interno di una complessiva strategia di innalzamento globale della tensione, sono aumentati esponenzialmente gli attacchi hacker russi non solo contro le infrastrutture di rete ucraine ma in tutto il mondo, così come gli utilizzi impropri di piattaforme e servizi web da parte delle forze dell'Intelligence del Cremlino.[20] Fortunatamente, sul versante opposto, fin dalle primissime fasi del conflitto, si è costituito un movimento volontario internazionale per la salvaguardia del patrimonio culturale online dell'Ucraina, che ha salvato fino ad ora oltre 50 TB di dati. L'iniziativa denominata "*SUCHO – Saving Ukrainian Cultural Heritage Online*"[21] non è interessante solo per la mole di informazioni immagazzinate in un lasso di tempo molto breve, ma anche perché l'insieme di procedure messe a punto potrebbero costituire uno standard di fatto nella salvaguardia del patrimonio culturale digitale in caso di conflitto o di altro tipo di emergenza.[22] Inoltre, anche se forse è l'aspetto più scontato, le attività di Web archiving servono a salvare il racconto, o meglio, i racconti che di quel conflitto vengono fatti. Per questo tutte le maggiori istituzioni culturali nel mondo hanno creato collezioni di risorse web di interesse nazionale riguardanti la Guerra in Ucraina, come è uso per molti grandi eventi non per forza di carattere catastrofico, come le pandemie, i Giochi Olimpici, le elezioni governative o le scoperte scientifiche. In Italia il deposito legale delle risorse digitali diffuse tramite rete informatica, di cui il Web archiving nelle modalità finora descritte è solo uno delle componenti, non è obbligatorio. Ciò comporta che le istituzioni della memoria debbano ottenere un preventivo consenso alla copia e all'archiviazione delle pagine web. Anche a causa delle scarse risorse a disposizione delle biblioteche, le procedure di richiesta di tali autorizzazioni si rivelano spesso lunghe e infruttuose: non solo i permessi che si riescono ad ottenere sono quantitativamente poco significativi, ma questi

[19] Non solo la società Twitter collabora da anni con le istituzioni del patrimonio per consentire di conservare gli archivi di tweet di rilevanza pubblica ma, a differenza di altri Social, tutti i suoi profili sono "pubblici" almeno per gli iscritti: ciò rende Twitter la piattaforma social più facilmente archiviabile e, di fatto, la più archiviata.

[20] A questo proposito interessante la serie di articoli a tema proposti sul blog Guerre di Rete: https://www.guerredirete.it/.

[21] https://www.sucho.org/.

[22] Con un tweet del 24 agosto 2022 sul proprio profilo, il gruppo SUCHO ha annunciato di stare lavorando alla redazione di un "*Handbook of Emergency Web Archiving*": https://twitter.com/sucho_org/status/1562484110802595841.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

possono arrivare anche con poca tempestività rispetto all'esigenza di archiviare contenuti per loro natura soggetti a continue modifiche.[23] Il risultato sono collezioni di risorse web archiviate non in grado di rappresentare compiutamente la pluralità di aspetti e punti di vista di cui si compone un evento così drammatico come una guerra.
Altre situazioni conflittuali, seppure non riconducibili a scenari di guerra reale, possono infine fungere da esempio di come possa essere fatto anche un uso strumentale dannoso delle url archiviate:

> Archiving services serve a variety of purposes beyond addressing link rot. Platforms like archive.is are reportedly used to preserve controversial blogs and tweets that the author may later opt to delete (Mondal et al. 2016). Moreover, they also reduce Web traffic toward "source URLs" when the original content is still accessible, thus depriving them of potential ad revenue streams (users do not visit the original site, but just the archived copy). In fact, anecdotal evidence has emerged that alt-right communities target outlets they disagree with by nudging their users to share archive URLs instead (Koebler 2014), or discrediting them by pointing at earlier versions of articles (Ralph 2017; Zannettou et al. 2018).

Zannettou et al. (2018) hanno dimostrato che l'utilizzo di url archiviate al posto di quelle originali ancora attive può essere scelto appositamente per diminuire il traffico in entrata verso siti o contenuti che non si vogliono sostenere, o per rimandare volontariamente il lettore a contenuti che l'autore ha scelto di modificare o cancellare, con intenti quindi deliberatamente dannosi. Seppure non sia pensabile poter controllare del tutto l'utilizzo *ex-post* degli archivi, una corretta gestione del patrimonio web dovrebbe considerare anche questi aspetti. La disponibilità di metadati completi in grado di fornire informazioni sul contesto dell'archiviazione, su tutte le fasi del processo di conservazione nonché sull'accesso alle informazioni e sui responsabili delle singole procedure è sicuramente un disincentivo ad un loro uso improprio.

"Archive first, ask questions later"

L'Archive Team si autodefinisce "*a loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritag*e".[24] Dal 2009, spesso in collaborazione con Internet Archive, il collettivo salva siti e servizi web a rischio di chiusura.[25] Tra i suoi progetti che hanno avuto maggiore risonanza mediatica c'è sicuramente quello dell'archiviazione di alcune migliaia di domande[26] e relative risposte fatte dagli utenti di *Yahoo Answers* , lo storico servizio di web forum di Yahoo! chiuso dall'azienda nella primavera del 2021.

[23] La Biblioteca Nazionale Centrale di Firenze, nell'ambito del proprio servizio di Web archiving, ha avviato a marzo 2022 una campagna di raccolta delle risorse web italiane relative al conflitto in Ucraina, tuttavia la generale adesione è stata scarsa: https://www.bncf.firenze.sbn.it/risorse-web-italiane-sul-conflitto-in-ucraina/.

[24] https://wiki.archiveteam.org/.

[25] L'elenco dei progetti dell'Archive Team suddivisi per categoria è visionabile al link: https://wiki.archiveteam.org/index.php/Category:Projects_status.

[26] L'obiettivo dichiarato dell'Archive Team era quello di archiviare le 84 milioni di domande i cui link erano mostrati nella Sitemap del sito ma per ora sono accessibili tramite la Wayback Machine di IA soltanto alcune migliaia di domande: https://web.archive.org/web/*/https://answers.yahoo.com/question/* (Wodinsky and Mehrotra 2021).

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

Il motto dell'Archive Team "*Archive first, ask questions later*", come rilevato in (Ogden 2022), consente a questo gruppo di volontari di effettuare attività di archiviazione con una efficienza e tempestività il più delle volte impossibili da eguagliare per le istituzioni culturali pubbliche, che operano necessariamente all'interno di limiti imposti dalla normativa vigente e da uno specifico mandato:

> For better or worse, this action-oriented "brute force" approach has enabled AT to proceed where institutions like national web archives are subject to their own mandates and legislative environments that constrain the nature of what can be collected, stored and made accessible (Ogden 2022, 120).

Anche Internet Archive non chiede un permesso preventivo all'archiviazione ma, rispetto all'Archive Team, ha un approccio che si potrebbe definire meno aggressivo e un apposito servizio di *opt-out* attraverso il quale chiunque può richiedere la rimozione dei contenuti archiviati in archive.org.[27] L'archiviazione senza l'acquisizione di un preventivo consenso, infatti, non può essere semplicemente liquidata come una sorta di "male necessario", potendosi in talune circostanze considerare un vero e proprio abuso dal punto di vista del diritto d'autore o del trattamento dei dati, come dimostra il caso Tumblr:

> The Tumblr case highlights a mutable AT collective that (though committed to the tenets of practice) is both open to modification and willing to negotiate the often fuzzy boundaries that define what constitutes the "public Web" in web archiving (Ogden 2022, 127).

La piattaforma di microblogging Tumblr, lanciata nel 2007, è sempre stata molto popolare, arrivando a contare 455 milioni di blog, grazie soprattutto alle politiche relativamente poco restrittive in termini di contenuti pubblicabili, che comprendevano anche quelli "per adulti" o "sensibili".[28] Tuttavia, nel 2018, per adeguarsi alle nel frattempo mutate leggi statunitensi, Tumblr ha effettuato una stretta in questo ambito, annunciando la cancellazione dei contenuti non più ritenuti idonei. Il rischio di una massiva perdita di dati era dunque concreto non solo per la loro cancellazione da parte dell'azienda, ma anche per il volontario esodo e conseguente chiusura dei profili da parte di utenti che si erano sentiti traditi dal cambio di rotta della piattaforma. In risposta, nel dicembre dello stesso anno, l'Archive Team avvia il *Tumblr NSFW project*, chiamando a raccolta i volontari per individuare le url da archiviare, ma anche mettendo a disposizione strumenti di facile utilizzo per consentire una più ampia partecipazione all'archiviazione. Il progetto è stato fortemente ostacolato da Tumblr, che ha lamentato un vero e proprio "attacco" ai suoi sistemi. Inoltre l'archiviazione di contenuti che sono oggetto di deliberata cancellazione per cause legali e etiche può, a sua volta, essere una scelta discutibile almeno dal punto di vista morale. L'Archive Team, a questo proposito, si è sempre dichiarata "neutrale" rispetto alla tipologia di informazioni archiviate, sostenendo che tutto ciò che è su Internet almeno teoricamente potrebbe e dovrebbe essere salvato, per fornire un'immagine completa e diversificata dell'esperienza online in un determinato tempo

[27] https://help.archive.org/help/how-do-i-request-to-remove-something-from-archive-org/.

[28] In Italia sotto la definizione di "contenuti sensibili" rientrano comunemente quelli violenti e pornografici, ma anche quelli istigranti al consumo di alcol e droghe o promuoventi comportamenti generalmente dannosi per il singolo o la collettività. In ambito anglofono gli stessi contenuti sono identificati con l'acronimo *NSFW – Not Safe For Work*.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

e contesto (Ogden 2022). A questo punto appaiono però ineludibili alcune domande. Innanzi tutto su cosa si debba intendere per "neutralità". In un contesto come quello della formazione di un archivio, della costituzione di un patrimonio, ogni scelta è comunque una scelta politica. Inoltre può rivelarsi pericoloso perseguire una presunta neutralità durante un processo di conservazione a lungo termine, anche alla luce del fatto che non tutti gli aspetti di questo processo sono realmente controllabili o dipendenti dalla volontà di chi conserva. Forse, più che della neutralità, c'è bisogno di ribadire l'importanza della trasparenza delle scelte e delle procedure adottate dai diversi soggetti, pubblici o privati, che operano nel dominio della costruzione del patrimonio digitale. Garantire cioè che, in ogni momento, chi accede all'archivio possa avere contemporaneo accesso alle informazioni sulla sua costituzione: il contesto in cui le informazioni sono state prodotte, quali sono stati i criteri di selezione e chi li ha determinati e applicati.

> AT is actively shaping access to dead and dying platforms, as well as creating a community of practice centred on the preservation of access and "rogue archiving" strategies for saving the Web. Despite their use of common tools and standards, these practices will be seen in stark contrast to other risk-averse approaches to web archiving taken by conventional memory institutions, and community archives projects that centre an ethics of care for content creators and future users. AT's interventions simultaneously illustrate the possibilities of participatory web archiving at scale and the potential risks of such approaches in the face of platform resistance, rights and privacy concerns. Given the scale of AT's collecting activities and their impact on the coverage of the IAWM, understanding their practices offers insights into how the Web is transformed through web archiving, as well as their critical ethical implications for how these platforms are studied in future (Ogden 2022, 129).

Il caso citato della Biblioteca Nazionale Centrale di Firenze fortemente limitata nella possibilità di archiviare le risorse web nazionali relative al conflitto in Ucraina, a paragone con il progetto *SUCHO*, risulta quindi emblematico e non isolato. Proprio la scarsità di strumenti legali a disposizione, oltre ad una certamente poco diffusa consapevolezza sul tema, è una delle cause principali per cui su una questione tanto importante come la salvaguardia del patrimonio culturale digitale, durante una grave emergenza come quella di un conflitto armato, sia intervenuto per primo un gruppo costituitosi spontaneamente[29] e non gli organismi tradizionalmente a ciò deputati.[30]
È pertanto probabilmente arrivato il momento di ragionare su provvedimenti normativi che prevedano anche per le istituzioni pubbliche italiane la possibilità di procedere unilateralmente all'archiviazione di risorse web di interesse storico-culturale per ben determinate finalità. Risoluzioni se non valide in generale,[31] almeno in casi definiti di emergenza o urgenza, in cui l'intempestività

[29] Il progetto *SUCHO* ha avuto anche il sostegno e la collaborazione di importanti istituzioni e associazioni della cultura, come l'ULA – Ukrainian Library Association: https://www.sucho.org/partners.

[30] La Convenzione per la protezione di beni culturali in caso di conflitto armato, firmata per la prima volta all'AJA nel 1954 e successivamente modificata ed ampliata, ad esempio, non prende ancora in alcuna considerazione il patrimonio culturale digitale: https://www.unesco.beniculturali.it/english-convenzione-dellaja-1954/.

[31] A livello internazionale il complesso di norme entro cui si collocano le attività di Web archiving "istituzionale" è variegato ma riconducibile sostanzialmente a due scenari: il Web e Social media è regolato all'interno del deposito legale oppure è considerato un'attività lecita nell'ambito delle eccezioni al diritto d'autore per finalità culturali (interesse scientifico e storico): https://netpreserve.org/web-archiving/legal-deposit/.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

e quindi l'incompletezza della raccolta arrecherebbe grave danno al patrimonio digitale costituito o in costituzione. Ferma restando la facoltà dei produttori di informazioni, laddove la "pubblicità" di queste ultime non fosse pacifica, di impedire o limitare preventivamente la raccolta, con un utilizzo corretto di strumenti quali i file robots.txt[32]e le sitemap,[33] oppure di richiederne a posteriori una motivata cancellazione dagli archivi.

Risorsa in archivio

Definire un modello di governance per la creazione e la tutela del patrimonio web, valutando adeguatamente le risorse necessarie ad attuarlo, è fondamentale dunque non solo per non perdere la nostra memoria recente ma anche perché come sceglieremo di farlo influenzerà in molti modi l'immagine che di noi avranno nel futuro. Il primo passo, come si è visto, dovrebbe essere quello di definire le responsabilità sull'intero processo, anche e soprattutto con appositi atti normativi: in breve, chi deve conservare e chi è tenuto ad assicurarsi che i propri contenuti siano conservati. L'individuazione delle istituzioni pubbliche deputate alle attività di Web a Social media archiving dovrebbe avvenire abbandonando le tradizionali categorie di "risorsa bibliografica" o "archivistica" o "museale", ridistribuendo gli ambiti di competenza in base alla tipologia di informazione veicolata più che alla forma delle risorse o a chi le ha prodotte. Non dimenticando di delineare gli ambiti in cui la cooperazione interistituzionale si configuri come la soluzione più efficace ed efficiente per il raggiungimento degli obiettivi di conservazione e accesso nel lungo periodo. Cooperazione interistituzionale non solo orizzontale, a copertura delle risorse di diversa natura, ma anche verticale centro-periferia tra istituti appartenenti allo stesso dominio, al fine di intercettare dati di interesse locale o specializzato (Storti 2023). La fase successiva riguarda il coinvolgimento delle comunità, intese sia come gruppi portatori di una storia o di un interesse comune, per individuare e preservare le "memorie particolari", sia come comunità di esperti e attivisti che insieme lavorano per salvaguardare l'infrastruttura cooperativa per eccellenza: il Web. Non trattandosi solo di stabilire una collaborazione, come già avviene da tempo, con enti privati come Internet Archive, ma anche di tentare di ricondurre l'azione di gruppi come l'Archive Team all'interno di una strategia comune. Certo potrebbero esistere dei limiti alla possibilità di stabilire rapporti con gruppi autocostituitisi, sia per ragioni legali che di pura volontà dipendenti da una differente visione di cosa e come dovrebbe essere conservato e reso accessibile, ma anche continuare a progettare infrastrutture che non tengano in alcuna considerazione l'esistenza di queste importanti iniziative non pare producente. Ineludibile, inoltre, l'attivazione di politiche di sensibilizzazione nei confronti dei produttori delle informazioni. Sensibilizzazione sia sul tema generale della archiviabilità[34] delle risorse sia, in particolare per i contenuti diffusi tramite Social media, sull'uso consapevole delle piattaforme terze. Queste ultime, infatti, quasi sempre detengono legalmente l'esclusiva facoltà di decidere se e quando cancellare contenuti, renderli inaccessibili o impedirne l'archiviazione. Per

[32] https://it.wikipedia.org/wiki/Protocollo_di_esclusione_robot.

[33] https://it.wikipedia.org/wiki/Sitemap. È bene precisare che entrambi gli strumenti appena citati possono imporre limiti "logici" ma non "fisici" alla copia, trattandosi di semplici informazioni di tipo testuale che i *crawler* possono o meno tenere in considerazione.

[34] Per approfondimenti: https://www.bncf.firenze.sbn.it/biblioteca/archiviabilita-dei-siti-web/.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

questo motivo, oltre che formare alle pratiche di auto-archiviazione dei profili personali ma anche di quelli istituzionali e aziendali, è necessario parlare dell'importanza della differenziazione dei canali di distribuzione dei contenuti, e della possibilità di pubblicare informazioni con licenze aperte e su piattaforme altrettanto "aperte".[35] In ultimo non bisogna tralasciare una riflessione sulle piattaforme di accesso agli archivi web e l'educazione degli utenti finali al loro utilizzo, nella consapevolezza che la creazione di un patrimonio non è mai un processo del tutto neutrale ma, al contrario, condizionato da numerosi fattori come i vincoli legali ed etici, i limiti e le possibilità delle tecnologie, nonché la sensibilità storico-culturale di chi ne ha la responsabilità.
Nel 2021, in occasione del suo 25° anniversario, Internet Archive ha pubblicato una utopica *Wayforward Machine*[36]che mostra come potrebbe essere il Web del 2046. L'immagine restituita è quella di un mondo in cui avere accesso ad informazioni affidabili in rete potrebbe essere né semplice, né libero, né gratuito:

> Imagining utopia: #EmpoweringLibraries
> We can find a better way forward.
> The Internet Archive is facing a lawsuit by a cartel of corporate publishers that threatens the age-old right of libraries to buy, preserve and lend materials to the public.
> To fight for a world where libraries and learners are empowered through access to information, join our #EmpoweringLibraries campaign.[37]

Scongiurare un futuro che può apparire nel complesso certamente distopico, ma che per molti aspetti rispecchia scenari plausibili, è però nelle possibilità delle istituzioni della memoria purché siano in grado di agire in sinergia con tutti gli attori coinvolti nel processo e all'interno di una strategia complessiva di gestione del patrimonio culturale digitale.

[35] Il caso recente dell'esodo da Twitter verso Mastodon, una piattaforma social decentralizzata, è emblematico in questo senso. Molti utenti, preoccupati dalla parzialità nel trattamento, sia a breve che a lungo termine, dei propri dati da parte dell'azienda recentemente acquisita dal magnate Elon Musk, hanno non solo iniziato a scaricare il proprio archivio ma anche a spostarsi su una piattaforma che sia a livello legale che tecnologico sia resiliente rispetto a singole iniziative potenzialmente dannose. A questo proposito: https://www.ilpost.it/2022/11/21/archivio-dati-twitter/.
[36] https://wayforward.archive.org/.
[37] https://wayforward.archive.org/ia2046/.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

Riferimenti bibliografici

Allegrezza, Stefano. 2023. "Web e social media come nuove fonti per la storia." *Umanistica Digitale*, January, 137-162. https://doi.org/10.6092/ISSN.2532-8816/15665.

Bergamin, Giovanni. 2006. "La raccolta dei siti web: un test per il dominio 'punto it.'" *DigItalia* 1 (2): 170-74. http://digitalia.sbn.it/article/view/306.

Il Post. 2022. "Come richiedere l'archivio di tutto ciò che avete fatto su Twitter." November 21, 2022. https://www.ilpost.it/2022/11/21/archivio-dati-twitter/.

Hegarty, Kieran. 2022. "The Invention of the Archived Web: Tracing the Influence of Library Frameworks on Web Archiving Infrastructure." *Internet Histories*, July, 1–20. https://doi.org/10.1080/24701475.2022.2103988.

International Institute of Social History. n.d. "Occasio Digital Social History Archive." Text. Occasio Digital Social History Archive. Accessed January 6, 2023. https://iisg.nl/occasio/index.php.

"List of Web Archiving Initiatives." 2023. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=List_of_Web_archiving_initiatives&oldid=1132919882.

National Library of Australia. 2023. *PANDORA, Australia's Web Archive*. Accessed January 6. https://pandora.nla.gov.au/.

OCLC – Web Archiving Metadata Working Group (WAM). 2018. 'Descriptive Metadata for Web Archiving'. https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html.

Ogden, Jessica. 2022. "'Everything on the Internet Can Be Saved': Archive Team, Tumblr and the Cultural Significance of Web Archiving." *Internet Histories* 6 (1–2): 113–32. https://doi.org/10.1080/24701475.2021.1985835.

Parlangeli, Diletta. 2017. "Errore 404: la storia della 'Pagina non trovata' più frequente di Internet." *Wired* (blog), December 7, 2017. https://www.wired.it/internet/web/2017/12/07/errore-404-la-storia-not-found/.

Signorelli, Andrea Daniele, and VOIS. 2022. *Errore 404, così il web sta perdendo i pezzi*. CRASH – La chiave per il digitale. https://open.spotify.com/episode/7iYIin0mYnzkCKporF9pjN.

Storti, Chiara. (In press). "Community Webs. Il Web Archiving per La Creazione e l'accesso Permanente Alle Collezioni Di Interesse Locale Nell'ecosistema Del Web." In *Atti Del Convegno "Le Collezioni in Biblioteca: Nuovi Approcci per Un Elemento Di Importanza Strategica", Bolzano – Eurac Research, 21 Ottobre 2022*. Collana Sezioni Regionali AIB. Trentino-Alto Adige. AIB – Associazione Italiana Biblioteche.

Tuttosport.com. 2021. "Ecco come i social network stanno proteggendo gli utenti afghani." August 23, 2021. https://tuttosport.com/news/attualit/cronaca/2021/08/23-84728825/ecco_come_i_social_network_stanno_proteggendo_gli_utenti_afghani.

Wodinsky, Shoshana, and Dhruv Mehrotra. 2021. "We're Archiving Yahoo Answers So You'll Always Know How Babby Is Formed." *Gizmodo*, September 4, 2021. https://gizmodo.com/were-archiving-yahoo-answers-so-youll-always-know-how-b-1846643969.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-533

Yale University Library. n.d. "Cataloging Online Integrating Resources." Accessed January 10, 2023. https://web.library.yale.edu/cataloging/e-resources/updating-websites.

Zannettou, Savvas, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, and Gianluca Stringhini. 2018. "Understanding Web Archiving Services and Their (Mis)Use on Social Media." *Proceedings of the International AAAI Conference on Web and Social Media* 12 (1). https://doi.org/10.1609/icwsm.v12i1.15018.