Archiviare la rete: strumenti e servizi
Osservazioni a margine del 6° Workshop sul documento elettronico

Giovanni Bergamin - *(Biblioteca Nazionale Centrale di Firenze)*
Augusto Cherchi - *(Associazione Nazionale Archivistica Italiana)*
M. Alessandra Panzanelli Fratoni - *(University of Oxford)*[1]

Biblioteche e archivi sono sempre più consapevoli che l'archiviazione del web è una componente essenziale della loro missione. Mentre il numero di documenti (tra questi anche fonti primarie) disponibili sul web è in crescita esponenziale, molta di questa documentazione è già persa per sempre. La conservazione digitale non è solo un problema tecnologico, ma anche il risultato della collaborazione tra i produttori, gli utenti e le istituzioni della memoria.
"Conservare il digitale" è il messaggio del Workshop sul documento elettronico che l'Anai organizza a Torino dal 2010, che nella sua 6. edizione è stato dedicato al tema della conservazione del web, inteso nella sua interezza e complessità come un patrimonio informativo da selezionare e preservare per il futuro. Al centro della giornata l'esperienza delle biblioteche nazionali di Francia e Regno Unito, che da anni portano avanti un'attività di web-archiving, interpretata come conseguenza naturale del loro mandato: conservare la memoria identitaria del Paese. Due casi di studio che permettono di confrontare approcci e soluzioni differenti a partire da istanze comuni, e che nel loro insieme rappresentano un modello autorevole cui guardare da parte dei Paesi (tra cui l'Italia) che non hanno ancora messo in pratica un'archiviazione istituzionale del web.
Il panorama odierno offre un crescente sviluppo di strumenti e servizi di webarchiving, accessibile ad un numero sempre maggiore di utenti, e le istituzioni della memoria possono contare su un'infrastruttura di base. In Italia, ai sensi del D.P.R. 252/2006, è stata attuata una sperimentazione del deposito legale delle pubblicazioni digitali, cui si auspica segua in tempi brevi la realizzazione del regolamento specifico previsto dalla legge.

[1] La responsabilità del contributo è divisa tra gli autori come segue: l'Introduzione e la parte finale (dal paragrafo intitolato "Il caso Wikipedia" in poi) a Giovanni Bergamin; "Il Workshop sul documento elettronico" e "Le sfide" ad Augusto Cherchi, "Il modello francese", "Il modello inglese" a M. Alessandra Panzanelli Fratoni. Gli autori ringraziano Sophie Derrot e Helen Hockx-Yu per la disponibilità allo scambio di informazioni successiva al workshop, di grande aiuto nella stesura del presente contributo; ringraziano inoltre Paola Puglisi per i suggerimenti forniti in fase redazionale. Tutti i riferimenti a siti internet sono stati controllati il 3 ottobre 2016.

Introduzione

Secondo i risultati di una ricerca apparsa nel 2014 su *Harvard law review*, oltre il 50% dei riferimenti a risorse web (URL) presenti nelle sentenze della Corte Suprema degli Stati Uniti non sono più utilizzabili[2]: le pagine citate non esistono più oppure sono state modificate. Dalla fine del 2015 una pagina del sito della Corte Suprema – intitolata *Internet sources cited in opinions*[3] - riporta tutti i link, distinti per anno, alle risorse web citate nelle sentenze.

Se attiviamo questi link saremo indirizzati a una stampa in formato pdf della risorsa originale[4]. Si tratta di una soluzione di archiviazione del web che non fa uso degli standard correnti (il più importante dei quali è WARC[5]) ed è realizzata con tecnologie che qualcuno ha definito «inelegant but a start»[6].

Il caso sopra richiamato ci può aiutare a capire meglio il significato di "archiviazione del web" (espressione che traduce l'inglese *web archiving*). Secondo Wikipedia[7] l'archiviazione del web è il processo di raccolta di parti del World Wide Web con l'obiettivo di garantire che le informazioni siano conservate in un archivio e siano accessibili nel lungo periodo per ricercatori, storici e utenti in generale. In altre parole oltre alle tecnologie di raccolta del web (*harvesting*) occorre tenere presente che l'archiviazione del web ha (o dovrebbe avere) come obiettivo un servizio di conservazione per il lungo periodo. Questo servizio può essere definito come un servizio pubblico offerto da depositi digitali affidabili (*trusted*) in grado di garantire per una comunità di riferimento:

1. la "conservazione a livello di bit" (procedure di salvataggio);
2. la "interpretabilità" dei formati di file archiviati da parte di applicazioni informatiche;
3. l'"autenticità" (informazioni o metadati che assicurano l'identità e l'integrità di una determinata risorsa);

[2] Jonathan Zittrain, Kendra Albert, Lawrence Lessig. *Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. How to make legal scholarship more permanent,* <<*Harvard law review*>>, V. 127 (2014), n. 4, <http://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations/>.

[3] https://www.supremecourt.gov/opinions/Cited_URL_List.aspx.

[4] Questa è ad esempio relativa alle fonti citate nel 2005: <https://www.supremecourt.gov/opinions/cited_urls/05>. In questa pagina troviamo ad esempio: <https://www.supremecourt.gov/opinions/URLs_Cited/OT2005/04-473/04-473.pdf> (puntamento alla risorsa archiviata) e < http://www.fda.gov/cder/career/default.htm> puntamento alla risorsa originale che se cliccato fornisce il fatidico errore 404 (pagina non trovata).

[5] Sullo standard WARC si veda Stefano Allegrezza. *Nuove prospettive per il Web archiving: gli standard ISO 28500 (formato WARC) e ISO/TR 14873 sulla qualità del Web archiving,* <<Dig*italia*>>, A. X (2015), n.1/2, p. 49-61. <http://digitalia.sbn.it/article/view/1473/981>.

[6] Jeff John Roberts. *Google Link in Supreme Court Case Shows Struggle on Citation*, <<Fortune>>, 22.6.2016, <http://fortune.com/2016/06/22/google-link-supreme-court-citation/>.

[7] https://en.wikipedia.org/wiki/Web_archiving.

4. l'"accessibilità" da parte degli utenti (la presenza di metadati che permettono l'accesso)[8].

Se applichiamo le definizioni appena richiamate al servizio offerto dalla Corte Suprema a partire dalla pagina *Internet sources cited in opinions* possiamo notare che può presentare qualche problema il requisito della "interpretabilità": infatti, la scelta di non usare formati di archiviazione standard fa aumentare nel tempo il rischio di non reperire più applicazioni in grado di trattare altri formati.
Naturalmente l'esigenza di archiviare il web non riguarda solo la Corte Suprema degli Stati Uniti. Per un punto di vista aggiornato sull'argomento può essere utile il recente *Web Archiving Environmental Scan*[9]. Il rapporto – del gennaio 2016 – esamina dettagliatamente i programmi in campo da parte di 23 istituzioni a livello mondiale, includendo anche le esperienze degli utenti che accedono alle risorse archiviate. Le considerazioni di partenza del rapporto sono ormai generalmente condivise: sta aumentando sempre più il numero di documenti (tra questi anche fonti primarie) che si trovano *solo* sul web. La raccolta e la conservazione di questi contenuti – in continua crescita e in continuo cambiamento – è diventata ormai una necessità. Biblioteche ed archivi stanno sempre più riconoscendo che l'archiviazione del web è una componente essenziale della loro missione. Una stima risalente al 2014[10] parla di 17 Petabyte archiviati (1 *peta* è una grandezza composta da 1 seguito da 15 zeri).

Il Workshop sul documento elettronico

Conservare il web vuole dire accettare (almeno per ora) una perdita inevitabile – spesso grave – di contenuti, operare in un contesto in cui non valgono le regole di selezione tradizionali ma se ne dovranno creare di nuove, che dovranno essere trasformate in "pratiche" condivise, sostenibili dal punto di vista della applicabilità e dei costi, supportate da risorse adeguate e applicate con costanza di impegno. A chi spetta questa attività, chi sono i soggetti incaricati di archiviare il web? E quale parte del web merita di essere archiviata? Infine, cosa si sta facendo ora, quanto non è stato ancora fatto e quanto dunque è irreparabilmente perduto?

[8] Per questa definizione si veda Giovanni Bergamin, Maurizio Messina. *Magazzini digitali: dal prototipo al servizio*, <<Digitalia>>, A. V (2010), n. 1, p.115-122, <http://digitalia.sbn.it/article/view/246/157>.

[9] Gail Truman, *Web Archiving Environmental Scan. Harvard Library Report*. <https://dash.harvard.edu/bitstream/handle/1/25658314/HL_web_archiving_env_scan_2006.pdf?sequence=1>.

[10] Miguel Costa, Daniel Gomes, Mário J. Silva, *The evolution of web archiving*, «International Journal on Digital Libraries» 2016, DOI: 10.1007/s00799-016-0171-9.

A questi temi è stato dedicato il Workshop sul documento elettronico che si è tenuto il 26 maggio 2015 a Torino con il titolo *Web archiving. La rete come* universitas rerum*: selezionare, descrivere, conservare*[11].

Organizzata, come ogni anno, dall'associazione nazionale degli archivisti, questa edizione del workshop torinese ha affrontato il tema della conservazione del web dal punto di vista più ampio, che riguarda non solo la documentazione ufficiale prodotta dalla pubblica amministrazione bensì tutto quanto prodotto e trasmesso tramite la rete. La giornata ha così visto protagoniste le due biblioteche nazionali di Francia e Regno Unito, che hanno presentato i risultati di esperienze di archiviazione attive da anni[12].

In questo panorama la situazione italiana appare indubbiamente arretrata, anche se i vari interventi dei relatori che si sono susseguiti (bibliotecari e archivisti) hanno mostrato come, dal punto di vista teorico, il problema sia tutt'altro che disatteso[13].

Introducendo alla consultazione dei documenti proposti nel corso della giornata (disponibili sul sito dell'evento), le pagine che seguono cercano di fare il punto della situazione, collocando il tema in questione in quello più ampio della conservazione del digitale. Per questa ragione un cenno è dedicato alla presentazione

[11] http://www.documento-elettronico.it/workshop/workshop-2015.

[12] L'attività si svolge così in parallelo a quella organizzata presso i rispettivi archivi nazionali e finalizzata alla conservazione del documento in senso stretto. Un *Vademecum de l'archivage des documents électroniques* è disponibile nel sito delle Archives Nationales, insieme ad altri documenti che guidano alla corretta conservazione del documento elettronico (strumenti messi a disposizione degli archivi delle amministrazioni pubbliche: <http://www.archives-nationales.culture.gouv.fr/web/guest/archives-des-administrations>). Ai National Archives britannici spetta il compito di curare la conservazione della documentazione prodotta dal governo britannico e pubblicata in rete: lo UK Government Web Archive: <http://www.nationalarchives.gov.uk/webarchive/>. Per la situazione italiana si rinvia alle generiche considerazioni sulla conservazione del documento digitale proposte sul sito dall'Agenzia per l'Italia digitale (AGID) <http://www.agid.gov.it/>, constatando che a oggi non molto è cambiato rispetto a quanto scriveva Federico Valacchi in un articolo del 2002: *Il web per gli archivi e gli archivi nel web*, in «Archivi & computer», A. 12 (2002), n. 3, p. 7-16, in cui sottolineava il ritardo del mondo archivistico italiano in tema di *web archiving*, là dove la dottrina e la prassi archivistica avrebbero potuto invece offrire un fondamentale apporto metodologico. Sul tema Valacchi sarebbe poi tornato con *Archiviare il web? Verso l'obbligo della conservazione delle sedimentazioni documentarie telematiche*, in *1. gennaio 2004: pronti, attenti e via! La "nuova" gestione degli archivi delle pubbliche amministrazioni: Atti del 4. incontro di lavoro, Perugia, 26 novembre 2002: Atti del 5. incontro di lavoro, Terni, 2-3 dicembre 2003*, a cura di Giovanna Giubbini, Perugia: Soprintendenza archivistica per l'Umbria, 2005, p. 224-233. Per un inquadramento del tema in un'ottica archivistica si vedano inoltre i contributi di Stefano Vitali: *Una memoria fragile: il Web e la sua conservazione*, in *La storiografia digitale*, a cura di Dario Ragazzini, Torino: UTET Libreria, 2004, p. 101-127, e *La conservazione a lungo termine degli archivi digitali dello Stato*, in *Conservare il digitale*, a cura di Stefano Pigliapoco, Macerata: Edizioni Università di Macerata, 2010, p. 35-61.

[13] Si vedano in particolare le riflessioni proposte in quella sede da Alberto Salarelli e Stefano Allegrezza. Di quest'ultimo il recente contributo dedicato agli standard ISO, a cui rinviamo per approfondimenti relativi alle questioni tecniche qui menzionate: S. Allegrezza, *Nuove prospettive per il Web archiving, cit.*

del workshop torinese, del contesto in cui l'iniziativa è nata e si è consolidata, dimostrando fin dagli esordi di essere aperta alla riflessione e alla collaborazione tra professionisti di settori contigui. I casi di Francia e Regno Unito vengono quindi ripresi, per essere confrontati ed esposti con le proprie peculiarità così da aprire la strada verso l'ultima parte del contributo, che propone un resoconto aggiornato di alcune delle maggiori iniziative in corso a livello internazionale, per concludersi focalizzando l'attenzione sulla situazione italiana.

Il contesto

Il workshop sul documento elettronico è nato nel 2010 per iniziativa dell'Associazione nazionale archivistica italiana (Anai) - Sezione Piemonte e Valle d'Aosta, è realizzato in collaborazione con il Politecnico di Torino e con l'Archivio di Stato e sostenuto dal contributo della Compagnia di San Paolo. A segnare la continuità scientifica all'interno del workshop, è la direzione affidata a Maria Guercio (Digilab - Università Roma La Sapienza); mentre un supporto imprescindibile è sempre stato garantito dall'Università di Macerata, rappresentata da Stefano Pigliapoco coordinatore del Master dedicato alla formazione, gestione, conservazione degli archivi digitali (Fgcad)[14].

La scelta compiuta fin dal momento ideativo è stata quella di interpretare la giornata non come una ricognizione teorica e di informazione passiva su aspetti normativi, ma come un momento di confronto in cui, partendo da situazioni concrete, si illustrano e discutono i problemi affrontati, le soluzioni adottate, le questioni aperte su specifici aspetti del passaggio dal documento tradizionale a quello elettronico: per queste ragioni il workshop si caratterizza ormai come un appuntamento atteso per approfondire la riflessione sulla transizione al digitale[15].

Oltre le norme per condividere buone pratiche è il titolo di testa dell'evento, a segnare il taglio pratico e l'inquadramento problematico che si vuole dare al tema scelto di anno in anno, che viene affrontato con la presentazione di *case studies* selezionati tra le esperienze più significative del momento.

Alla base dell'esperienza torinese c'è la convinzione che le profonde trasformazioni che hanno attraversato, nel corso degli ultimi vent'anni, tutti i processi di gestione delle informazioni e di veicolazione della conoscenza abbiano modificato in profondità le tecniche e i modi in cui si forma, si gestisce, si conserva la documentazione. Di qui l'esigenza di prevedere dei momenti di analisi e di interpretazione dei cambiamenti intervenuti nei sistemi produttivi, nei modelli organizzativi, nei contesti amministrativi, negli scenari legislativi finalizzati a mettere in evidenza sia i punti di forza e di debolezza dei processi in atto, sia i risultati acquisiti e i nodi

[14] FGCAD: Formazione, gestione e conservazione di archivi digitali in ambito pubblico e privato: <http://masterarchividigitali.unimc.it/>.

[15] Si vedano in proposito i discorsi posti in apertura della prima edizione: <http://www.documento-elettronico.it/workshop/workshop-2010/atti-della-giornata/56-introduzione-ai-lavori>.

problematici, i vantaggi competitivi e le criticità del mondo digitale. Il tutto nell'ottica di contribuire a creare un ecosistema al quale partecipino, con ruolo da protagonisti, tutti gli attori impegnati ad affrontare i nuovi scenari della gestione della documentazione: professionisti a vario titolo coinvolti, agenzie formative, aziende, pubblica amministrazione nelle sue diverse articolazioni ed espressioni.
In questa prospettiva, i temi affrontati dal workshop, dopo una prima edizione (2010) di inquadramento metodologico, sono stati:

- 2011: *L'archivio ibrido*: la più tipica situazione della transizione, dove documento tradizionale e documento digitale si sovrappongono, duplicano, moltiplicano creando situazioni spesso difficili da governare;
- 2012: *L'email come documento*: uno dei nodi più problematici della gestione documentale, dove il tema della conservazione si confronta con problemi di quantità, sicurezza, privacy, standard;
- 2013: *Lo scarto in ambiente digitale*: dal Massimario di selezione al Piano di conservazione, che ancora più di prima impone di costruire l'archivio a partire dal momento della produzione del documento, nell'operatività corrente;
- 2014: *I data base*: intesi nella doppia funzione di contenitori di conservazione di documenti e di documenti in quanto tali dunque da conservare.

Nell'arco delle prime cinque edizioni l'appuntamento ha raccolto l'adesione di oltre 600 partecipanti. Tra i relatori, a illustrare come stanno affrontando la transizione al digitale, si sono presentate importanti aziende private produttive e di servizi (Alenia, Intesa Sanpaolo, Reale Mutua Assicurazioni, Registro italiano navale), enti pubblici di diversa natura, da organi di sorveglianza (come la Banca d'Italia), a enti economici (come le Camere di Commercio), agenzie sanitarie, enti territoriali (Regioni e Comuni), università, uffici dell'amministrazione archivistica del Ministero dei beni e delle attività culturali e del turismo. Di tutte le edizioni, il sito www.documento-elettronico.it pubblica gli atti in forma di registrazione video dell'intera giornata. Dal 2014 il panel dei relatori si è inoltre aperto a contributi internazionali. La prima occasione è stata la presentazione da parte dell'Archivio federale svizzero della piattaforma SIARD Suite dedicata alla conservazione nel tempo delle banche dati relazionali.
Ma è soprattutto nel 2015 che il centro della scena è stato occupato dal confronto con esperienze straniere. E a imporre questa situazione è stato il tema stesso del *web archiving*, che vede il nostro Paese scontare una posizione di grave ritardo; di qui l'interesse e l'esigenza di confrontarsi con le esperienze di due paesi europei, Francia e Regno Unito, che hanno ormai da qualche anno definito cornici normative e prassi che si possono assumere come punti di riferimento.

Il modello francese

Ad illustrare il modello francese è intervenuta Sophie Derrot, della Bibliothèque Nationale de France (BnF), dove è curatrice all'interno della sezione dedicata al deposito legale (Département du Depôt légale)[16]. Dal titolo stesso del suo intervento emerge la forte continuità con la tradizione, che inserisce l'archiviazione del web nel solco della conservazione della memoria fondato sul deposito legale: *Le dépôt légale de l'internet à la Bibliothèque Nationale de France: de l'URL au patrimoine national*.

La cornice legislativa

Il quadro normativo che costituisce la base di una sistematica archiviazione del web risponde a un bisogno culturale, chiaramente esplicitato nel sito internet della BnF. «Perché archiviare il web?» è la domanda che accoglie il visitatore della sezione dedicata. E le risposte fornite sono chiarificatrici dello spirito con cui la questione è affrontata. «Perché abbiamo bisogno di conservare questa forma nuova di comunicazione che è oramai pervasiva, presente in ogni spazio di interesse per la conoscenza e per la vita sociale». E ancora: «Archiviare il web facendo leva sul deposito legale estende la missione storica di raccogliere l'eredità culturale francese»[17].

Punto di partenza è la legge sul deposito legale, che la Francia vanta dal 1537, quando Francesco I introdusse l'obbligo di deposito di una copia di ogni libro stampato e messo in vendita all'interno del Regno. Il deposito avveniva allora presso il castello di Blois, ma lo spostamento a Parigi, e poi le varie trasformazioni che portarono la biblioteca reale a diventare l'attuale biblioteca nazionale, non hanno prodotto soluzioni di continuità. Così l'odierna formulazione del deposito legale si inscrive in una tradizione di lungo periodo, allargando e definendo meglio le diverse forme di pubblicazioni che si sono via via manifestate, in particolare i cambi di supporto che dalla carta si sono estesi a comprendere l'audio, il video e il digitale. Sicché oggi la Bibliothèque Nationale de France «ha il compito di raccogliere, a titolo di deposito legale, dal momento in cui essi sono messi a disposizione del pubblico, i documenti a stampa e tutti i materiali che veicolano informazioni e conoscenza, siano essi fotografici, sonori, audiovisivi, multimediali, qualunque sia il procedimento tecnico con cui sono prodotti, le modalità di edizione o diffusione, inclusi i software e le banche dati, di qualunque natura e supporto. Il deposito interessa altresì qualunque segno, messaggio, suono, immagine che sia oggetto di comunicazione pubblica per via elettronica (Internet)»[18].

[16] Non sembra secondario come alla sua attuale posizione Sophie Derrot arrivi con un curriculum formativo in cui molto spazio hanno le discipline storiche: diploma di archivista paleografo (École Nationale des Chartes) e dottorato di ricerca all'École Pratique des Hautes Etudes, scienze storiche e filologiche.

[17] http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html.

[18] Traduzione, a cura degli autori, del testo disponibile alle pagine dedicate al deposito legale nel sito della Bibliothèque Nationale.

Dunque la rete non fa eccezione: ogni forma di comunicazione concepita per essere pubblicata, ovvero rivolta a un pubblico, ricade sotto questa norma e deve essere conservata. L'estensione della normativa, con un inevitabile adattamento, è stata prodotta nel 2006, e successivamente (2009) inserita nel *Code du Patrimoine* (articoli L131-1 a L133-1 e R131-1 a R133-1).
Ma se per ogni pubblicazione digitale l'obbligo si applica estendendolo, ovvero chiedendo il deposito della copia presso la biblioteca nazionale, la comparsa della rete ha necessariamente spostato sull'istituto conservatore anche l'onere della raccolta dei documenti: «Contrairement au dépôt légal traditionnel (des publications imprimées ou sur support audiovisuel, par exemple), le dépôt légal des sites web n'implique aucune démarche active de la part de l'éditeur. Les collectes se font de manière automatique […] la BnF est susceptible de prendre contact avec l'éditeur au cas par cas pour trouver des solutions techniques afin d'améliorer la collecte du site»[19].
Se il contesto normativo fa sì che la conservazione del web si inscriva perfettamente nella tradizione e sia stata inserita tra i compiti e nelle pratiche della biblioteca nazionale senza problemi di carattere amministrativo, la particolare natura del web produce la necessità di un cambiamento importante negli obiettivi tradizionalmente legati ai depositi librari nazionali. Da un lato il perimetro nazionale costringe alla individuazione di un "internet francese" entro il quale la norma si può, e si deve, applicare. Da un altro punto di vista, l'ampiezza della rete obbliga a rinunciare all'*esaustività*, ovvero a raccogliere tutto quanto pubblicato sul territorio nazionale, puntando invece alla *rappresentatività*.

Gli attori

In totale il Départment du depôt légale della Bibliothèque Nationale occupa 142 bibliotecari. Non tutti si occupano di tutto: se la legge crea l'obbligo alla raccolta di ogni pubblicazione, indipendentemente dal supporto, diversi sono i servizi allestiti per gestire le differenti tipologie di materiali. Il digitale e il web ricadono sotto le cure del Service du depôt legale numérique, che impiega 7 unità, con le quali collaborano gli ingegneri della sezione informatica (Département des Systèmes d'information).
Importanti poi le collaborazioni con altre istituzioni, a livello nazionale (biblioteche titolate a raccogliere il deposito legale a livello regionale, associazioni, ricercatori) e internazionale (lo International Internet Preservation Consortium, altre biblioteche nazionali o universitarie, istituzioni culturali, fondazioni non a scopo di lucro, Internet Archive, e imprese).

Cosa si conserva? Raccogliere e selezionare

Di fatto il deposito legale si applica a ogni pubblicazione indipendentemente dal

[19] http://www.bnf.fr/fr/professionnels/depot_legal/a.dl_sites_web_mod.html.

formato. Ci si trova pertanto di fronte alla necessità di governare documenti digitali che si presentano nei formati più vari: dalle pagine html, ai pdf, all'epub, alle diverse estensioni che veicolano contenuti multimediali (immagini, video, audio). Si conserva però ciò che ricade nella categoria delle pubblicazioni, vale a dire quei contenuti che sono stati concepiti per essere diffusi; sono dunque escluse le reti interne (intranet) e la posta elettronica. L'obiettivo è la rappresentatività di quanto pubblicato nel Paese: vi è perciò l'ambizione di conservare tutto ciò che compare con estensione di dominio .fr, indipendentemente dal contenuto. La selezione invece si applica a tutti gli altri domini (.org, .com etc.) per i quali si chiama in causa l'expertise dei professionisti (bibliotecari) specializzati, che hanno il compito di monitorare il web e individuare i siti di interesse.

La raccolta è affidata a Heritrix[20], il robot (*crawler*) sviluppato da Internet Archive (https://archive.org/index.php) e utilizzato da altre istituzioni che si occupano di archiviare il web. In BnF sono al lavoro 70 installazioni di Heritrix che percorrono il web monitorando una lista di indirizzi loro assegnati, copiando i singoli elementi di cui si compongono le pagine in modo da permettere, a valle di questa scansione, la ricostruzione del sito. L'intero percorso effettuato dai robot è tracciato attimo dopo attimo.

"Cogli l'attimo" potrebbe essere adottato come slogan. Si nutre infatti l'ambizione di individuare e conservare quei siti che, per loro natura, sono destinati ad avere un ciclo di vita assai breve oppure sono cancellati dai loro stessi autori poco dopo la messa online: "Archiver l'instant sur le temps long". L'obiettivo «è costruire l'archivio nel momento stesso in cui si forma». Vengono così archiviati anche i messaggi che denunciano la perdita dei legami di navigazione, così come le pagine che non esistono più (codice errore 404). Solitamente i *crawler* si scontrano con blocchi di accesso alle pagine di interesse (file robots.txt), determinati dagli stessi editori. Questi vengono superati grazie a software in grado di valicarli che la BnF può utilizzare in virtù della legge.

Volendo costruire un archivio davvero rappresentativo del web in lingua francese, la BnF si è data alcuni obiettivi prioritari: da una parte rendere conto della *specificità della rete* (e qui rientra la sfida del voler catturare quanto ha vita brevissima); dall'altra parte, con *habitus* proprio del conservatore, prestare grande attenzione alla *continuità delle collezioni*, evidenziando la necessità di mantenere il legame con le raccolte tradizionali pubblicate su supporto materiale.

Discorso a sé poi meritano le pubblicazioni native digitali, tra le quali ad esempio i siti ufficiali dei principali organi dello Stato. Vengono archiviati i social media e i contenuti legati a singoli eventi di interesse pubblico. Non da ultimo, i siti dell'editoria d'informazione, al fine di creare una continuità appunto con le collezioni cartacee. E volendo perseguire l'obiettivo di costruire strategie di reperimento di

[20] Si veda in proposito il già citato S. Allegrezza, *Nuove prospettive per il Web archiving*, p. 7.

contenuti intellettuali, si cercano accordi con gli editori per tutti i casi di accessibilità dei siti ad accesso oneroso.

Come si conserva?

Nella storia del *web archiving* in Francia (come pure nel Regno Unito) si distinguono due fasi: prima e dopo la legge che affida alle biblioteche nazionali il compito (e l'obbligo) di costruire la memoria del web.
Nell'archivio attuale confluiscono: collezioni retrospettive dal 1996 al 2005, acquisite da Internet Archive; e raccolte costruite internamente a partire dal 2010 col risultato che, al 1 gennaio 2016, l'archivio conteneva 668,05 Terabytes e 25,9 miliardi di url.
Questa quantità enorme (e sempre crescente) di dati viene archiviata in tre copie, su due supporti diversi: in hard disk per averne la disponibilità immediata e quindi l'accesso; su supporto magnetico, per una durata maggiore, avviene il salvataggio finalizzato alla conservazione perenne. Il sistema scelto è SPAR, conforme allo standard OAIS ed adottato non solo per il web, ma per ogni pubblicazione digitale; ciò ha l'ulteriore vantaggio di costruire un deposito unico per le pubblicazioni digitali e il web.

L'accesso

La consultazione dei depositi del web è regolata strettamente: avviene a fronte di accreditamento e solo *in loco*, nelle biblioteche incaricate del *web archiving*. La restrizione dell'accesso è legata alla protezione dei diritti.
Da sottolineare che l'accesso non è diretto ai contenuti: al momento la ricerca si può effettuare sul nome del sito o per data. L'indicizzazione a testo completo non è ancora attiva. L'interfaccia utilizzato per la consultazione è la Wayback Machine sviluppata da Internet Archive.
Il pubblico principale sono i ricercatori; l'archivio del web è visto naturalmente come fonte di informazione, ragion per cui ad ogni sito si applica un link permanente che ne consente la corretta citazione; ma l'archivio del web può costituire *di per sé* oggetto della ricerca, e questo uso, puntualizza Derrot, si accresce gradualmente.

Il modello britannico

A trattare del modello adottato nel Regno Unito è Helen Hockx-Yu, fino al settembre 2015 responsabile del progetto Web Archiving della British Library (BL)[21]. Il titolo del suo intervento – *Meeting the challenges of preserving the UK Web* –

[21] Da settembre 2015 Helen Hockx-Yu è *Director of Global Web Services* di Internet Archive, l'organizzazione statunitense che è tra le più importanti agenzie al mondo specializzate in web archiving. Da essa, come si legge anche qui, sia la British Library che la Bibliothèque Nationale de France hanno acquisito i materiali con cui hanno avviato la propria attività di archiviazione del web.

mette in chiaro la tensione verso un obiettivo non facile, insieme alla forte volontà di perseguirlo, senza alcun riferimento al vincolo di legge del deposito legale.
D'altronde presso la British Library l'archiviazione del web è iniziata ben prima che una legge intervenisse ad estendere il deposito legale delle pubblicazioni non a stampa: il *Non-print Legal Deposit Regulation* è del 6 aprile 2013. La costruzione di un deposito dei siti era già iniziata da più di un decennio.
Inizialmente l'archiviazione era basata su un preventivo accordo con i proprietari dei siti, ed ha condotto alla costruzione di un open archive di 72.000 *snapshots*. La legge, estendendo l'obbligo di deposito anche alle pubblicazioni non a stampa, e dunque includendo il digitale e l'online, ha non solo esteso la conservazione, ma anche cambiato i criteri di selezione e raccolta. Da una parte si è creata la necessità, come in Francia, di creare il deposito completo di quanto va a formare l'eredità culturale del Paese; dall'altra è stata superata la necessità di stringere accordi preventivi con i proprietari dei siti. E ciò ha avuto effetti nell'accessibilità della raccolta.

Chi conserva?

Nel Regno Unito, la funzione di conservazione è concepita sulla base di un modello che tiene insieme istituzioni governative e strutture accademiche in una bilanciata forma di collaborazione. La conservazione è assicurata in 4 copie: presso la British Library, nelle due sedi di St. Pancras (Londra) e di Boston Spa; presso la National Library of Scotland e la National Library of Wales. L'accesso in consultazione è inoltre consentito in tre sedi universitarie: la Bodleian Library di Oxford, la University Library di Cambridge e la biblioteca del Trinity College di Dublino.

Selezione e raccolta

La raccolta interessa la totalità di quanto pubblicato da siti con nome a dominio .uk, che è tra più grandi al mondo con oltre 10.000.000 di siti. A questi si aggiungono i siti .org e .com, costituiti nel Regno Unito. Il *non-print legal deposit* interessa tutto il cosiddetto web aperto, con estensione del nome a dominio .uk, nonché tutti gli altri siti pubblicati nel Regno Unito non .uk, che siano resi disponibili da una persona o da una organizzazione la cui attività si svolga all'interno del Paese. La strategia di raccolta mira a catturare il dominio .uk con il maggior grado di completezza possibile. I materiali così raccolti sono organizzati in collezioni speciali.
La selezione dei siti non .uk si basa su tre criteri di priorità: siti di organizzazioni o personalità di particolare rilievo per la nazione; siti di informazione da agenzie di stampa; siti dedicati ad eventi politici, culturali o sociali di particolare rilevanza (ad esempio, per il 2015: le elezioni politiche, l'epidemia di ebola, il centenario della prima guerra mondiale).
Di conseguenza anche le modalità di scansione e raccolta delle pagine web varia in funzione della tipologia di sito:per tutti i domini .uk la raccolta è automatica, mentre

per gli altri siti si procede verificando la rilevanza, secondo quanto stabilito dalla legge che fa riferimento alla localizzazione della erogazione del sito o ai contenuti trattati).
Questa impostazione ha fatto sì che nel 2014 si siano scaricati 2.500.000 siti non .uk. Tutte queste strategie, messe in atto dal momento della pubblicazione della legge, hanno prodotto un'archiviazione in crescita costante: nel 2014 sono stati raccolti siti per un totale di 57 TB (l'anno precedente erano 31), di cui 2,5 non .uk.
La selezione e conservazione dei social media pone problemi particolari. Dal punto di vista dei riferimenti normativi essi sono considerati alla stregua delle altre risorse web. Nei fatti è impossibile archiviare per intero Twitter e Facebook, mentre Youtube non è nemmeno considerato perché ricade sotto un'altra tipologia di materiali. In questi contesti un ruolo decisivo lo svolgono i curatori, cui è affidata la selezione delle fonti e delle pagine da conservare.

Accesso

Come in Francia, l'accesso ai contenuti del web archivi UK è possibile solo *in loco*, nelle sale di consultazione delle sette biblioteche titolari del *Legal Deposit*.
L'accesso è ristretto per questioni attinenti i diritti degli editori, giacché la raccolta dei siti non avviene più a fronte di un accordo. Inoltre, alcune restrizioni sono state introdotte per rispettare quanto richiesto dal riconoscimento di un diritto emerso solo di recente, proprio per effetto della visibilità che, con le teconologie web, possono assumere vicende di natura personale: il diritto all'oblio. Ricorda infatti Hockx-Yu il pronunciamento della Corte di giustizia europea che ha riconosciuto il diritto del singolo di chiedere la cancellazione di tracce relative alla propria esistenza. Si rileva come questo diritto, da un canto, sia in concorrenza con due istanze: il "diritto ad essere ricordato", che è quanto la legge sul deposito legale protegge; e il diritto del cittadino di una società democratica ad accedere nella forma più ampia e libera possibile alle informazioni e alla conoscenza, sfruttando gli strumenti più evoluti messi a disposizione dal livello di sviluppo della società in cui vive. D'altra parte nel Regno Unito il diritto all'oblio non è tutelato: la normativa di riferimento è il *Data Protection Act* del 1998, che sposta sulla persona interessata l'onere della prova: questi deve cioè dimostrare l'esistenza di un danno o di una forma di sofferenza, prodotta dalla diffusione di informazioni relative alla sua persona.
La ricerca nell'archivio web britannico si può effettuare anche tramite il catalogo generale della British Library[22] ed è una ricerca full text, non limitata al nome del sito o alla data. Per consentire una navigazione più efficace, da parte della BL i siti vengono non solo indicizzati nel catalogo generale, ma riorganizzati in cosiddette collezioni speciali, ovvero gruppi di siti legati da un tema.

[22] http://explore.bl.uk/primo_library/libweb/action/search.do?vid=BLVU1.

Le sfide: la perdita di dati e il web archiviato come oggetto di indagine

Accanto alla contraddizione che emerge tra la tensione verso un archivio più completo possibile, e la sua navigabilità, si affianca un altro paradosso: nell'archiviare il web, si vanno a spezzare i legami tra i siti conservati e il resto della rete, in certo senso intaccando la natura stessa della documentazione pubblicata sul web.

E comunque la vera sfida è rappresentata dalla perdita di dati che si verifica comunque, anche quando si pone una adeguata attenzione ai criteri e ai metodi di conservazione; mentre la ricerca dell'*Harvard Law Review* citata nell'introduzione analizzava un singolo caso esemplare, uno studio riferito più in generale al web britannico (fig. 1) mostra come in dieci anni si sia persa una percentuale molto alta dei contenuti archiviati.

Per fronteggiare questo problema la BL adotta un sistema di archiviazione che offre, per ogni sito, quante più immagini possibile, e conserva i link interni. Inoltre, per i siti che sono preventivamente selezionati e considerati di particolare interesse sono stati sviluppati sistemi di conservazione e di ricerca ad hoc particolarmente raffinati. A fronte della perdita di dati, grazie al lavoro di selezione ed organizzazione di ciò che si conserva il web archiviato si presenta come un oggetto di indagine in grado di fornire al ricercatore signficativi spunti di riflessione.

Uno degli aspetti del *web archiving* così come interpretato in BL consiste nell'attivazione di progetti in collaborazione con istituti di ricerca. Uno di questi vede coinvolti lo Institute for Historical Research (prof. Jane Winters, Digital History), la

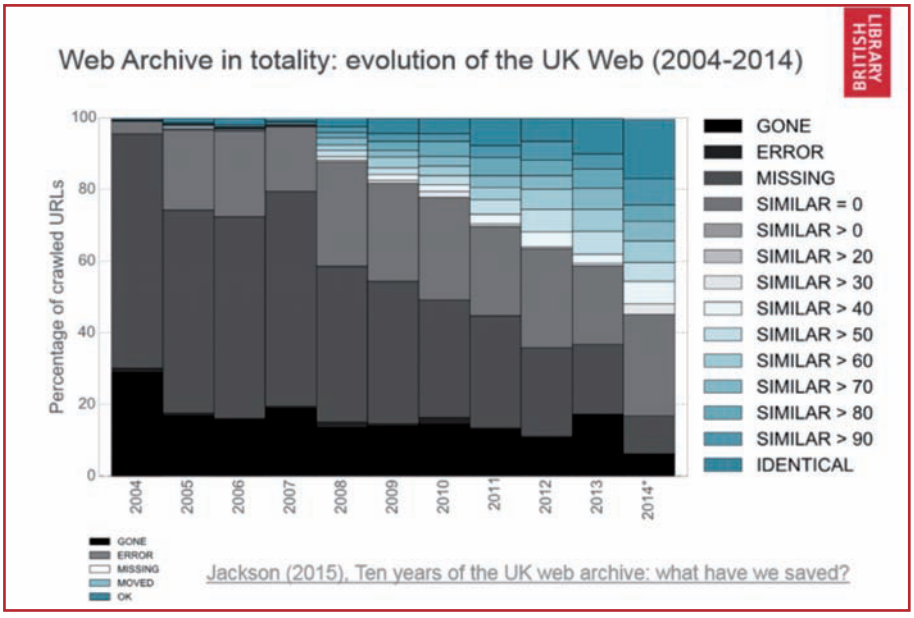

Figura 1.

University of Oxford (Josh Cawls, Research Assistant, Oxford Internet Institute), la stessa British Library e la Aarhus University. Oggetto di ricerca sono i dati raccolti nel periodo 1996-2014, con i seguenti obiettivi:

- mettere in evidenza il valore del *web archiving* per la ricerca scientifica;
- sviluppare un modello teorico e metodologico finalizzato all'analisi degli archivi web;
- esplorare le implicazioni etiche della ricerca basata sui big data;
- contribuire allo sviluppo delle raccolte e dell'accesso presso la British Library;
- formare i ricercatori all'uso degli archivi web.

Di seguito alcuni risultati del progetto:

- miglioramento dell'interfaccia allestita alla BL per i dati del periodo 1996-2013;
- una monografia ad accesso aperto dal titolo *The Web as History: Using Web Archives to Understand the Past and the Present* (UCL Press);
- un modulo online di formazione all'uso dell'archivio web;
- due corti animati che spiegano al grande pubblico cosa sia l'archivio del web;
- una serie di casi di studio che illustrano la ricerca basata sull'archivio web.

Il progetto ha coinvolto 11 ricercatori del settore umanistico, sostenuti con borse di studio. Dai titoli dei singoli progetti si evince bene la varietà degli approcci e dei temi di ricerca, che vanno dalla sociologia alla storia culturale, dalla politologia alla letteratura contemporanea[23]. Lo sviluppo di un'interfaccia per l'accesso ai dati del dominio UK per il periodo considerato (1996-2013: 2,5 bilioni di url per 30 TB) ha generato il prototipo Shine (<http://webarchive.org.uk/shine>): realizzato con un meccanismo di ridefinizione ad ogni stadio del progetto (dalla elencazione dei requisiti necessari al caso in studio, al collaudo con l'utenza finalizzato ad implementare immediatamente il prototipo in funzione delle esigenze evidenziate dall'utente), Shine consente:

- ricerca a tutto testo, con opzioni di prossimità e la possibilità di escludere specifiche stringhe di testo;
- possibilità di esclusione di precise risorse o di interi hosts dai risultati della ricerca;

[23] Tracing notions of heritage (Rowan Aust); Beat literature in the contemporary imagination (Rona Cran); Revealing Euroscepticism in the UK web domain and archive (Richard Deswarte); The UK Parliament Web Archive (Chris Fryer); An ethnosemantic study of London French habitus as displayed in blogs (Saskia Huc-Hepher); Capture, commemoration and the citizen-historian: Digital Shoebox archives relating to P.O.W. in the Second World War (Alison Kay); Digital barriers and the accesible web disabled people information and the internet (Gareth Millward); A history of the online presence of UK companies (Marta Musso); The Ministry of Defence's online development and strategy for recruitment between 1996 and 2013 (Harry Raffal); Public archaeology a digital perspective (Lorna Richardson); Do online networks exist for the poetry community? (Helen Taylor).

– filtri a faccette multiple, come tipologie di contenuto, suffissi, dominii, anno di raccolta dei dati;
– salvataggio ed esportazione dei risultati della ricerca.

Importanti insegnamenti sono derivati da questa esperienza. Si è innescato un processo di apprendimento per tutti gli attori coinvolti, chi fa ricerca e chi produce gli strumenti per fare ricerca. Si sono evidenziate le potenzialità e i limiti del web-archive, e chiarite le sfide di carattere metodologico implicate. Si è capito che non si tratta di scegliere tra grandi e piccoli depositi di dati ("big data" vs "small data") ma di sapersi muovere intelligentemente tra i due, realizzando strumenti che consentano, partendo da un punto preciso, una raccolta di dati particolari; e di visualizzarla nel contesto generale dei bilioni di altri punti così come di focalizzarsi fino al dettaglio minuto. L'esperienza della British Library mette così in evidenza l'importanza del contesto, come sottolinea la stessa relatrice («voi archivisti sarete d'accordo con me»), con riferimento allo specifico della dottrina archivistica che riconosce il valore instrinseco e il significato del singolo documento appunto quando inserito nel suo contesto di produzione[24].
In conclusione, per quanto riguarda l'esperienza della BL le emergenze, i punti su cui continuare a riflettere, sono sicuramente le dimensioni e le quantità con cui ci si deve confrontare, e la conseguente difficoltà di monitorare, selezionare, raccogliere, organizzare uno spazio web in continua crescita; inoltre, l'uso ancora limitato dell'archivio da parte della ricerca scientifica (anche per la permanenza di problematiche metodologiche e tecniche non ancora risolte); meno scontato ma di grande interesse, trovare una chiave di apprendimento per la prossima generazione di web archive, con attenzione ad aprire una finestra sull'archivio esistente, già molto ricco, e studiare il web UK storico nel contesto più generale.

Il caso Wikipedia

Un caso interessante di archiviazione del web – e che potremmo definire in controtendenza – è quello offerto da Wikipedia. Si tratta del settimo sito a livello mondiale per numero di accessi (mezzo miliardo di visitatori unici al mese[25]) che prevede un sistema molto efficiente di "auto-archiviazione" delle pagine pubblicate. In pratica tutte le varie revisioni delle pagine vengono memorizzate con un identificativo (*oldid*). Il riferimento a una voce di Wikipedia può essere fatto quin-

[24] Questa osservazione di Helen Hockx-Yu, accostata a quella di Sophie Derrot, sull'importanza di costruire l'archivio del web nel momento in cui esso si forma, offrono, ci sembra, interessanti spunti di riflessione circa una necessaria permeabilità delle discipline (l'archivistica da una parte, la bibliografia e la biblioteconomia dall'altra) quando le metodologie proprie dell'una offrano all'altra strumenti più utili nella organizzazione di particolari depositi informativi.
[25] http://www.unacitta.it/newsite/intervista.asp?id=2443.

di con due modalità. La prima è l'uso di quello che viene chiamato link permanente[26] e che indirizza ad una determinata revisione di una voce: ad esempio it.wikipedia.org/w/index.php?title=Matteo_Renzi&oldid=82905253 è la pagina di Wikipedia dedicata al Presidente del Consiglio così come è stata pubblicata il 30 agosto 2016, mentre it.wikipedia.org/w/index.php?title=Matteo_Renzi&oldid=1452812 è la versione del 26 ottobre 2005 della stessa pagina, ma che ci fornisce informazioni sull'allora Presidente della Provincia di Firenze. La seconda modalità – quella comunemente usata – ci porta alla versione corrente della pagina (it.wikipedia.org/wiki/Matteo_Renzi), che può cambiare nel tempo e presentare così significative differenze con la pagina alla quale ci si voleva riferire.

Soluzioni per il *link rot*

Il servizio di "auto-archiviazione" offerto da Wikipedia è un'eccezione: normalmente un sito non offre sistemi così efficienti per il riferimento alle varie versioni di una determinata pagina. La regola di chi pubblica le informazioni sembra essere quella di non prendersi cura di quello che avviene dopo la pubblicazione. La ricerca apparsa su *Harvard law review*, richiamata all'inizio, ci informa che il fenomeno del mancato o non corretto funzionamento dei link interessa circa il 70% dei riferimenti che compaiono in riviste accademiche relative a discipline legali e il 20% dei riferimenti in quelle di ambito tecnologico-scientifico[27].

Ci sono iniziative che cercano di offrire una soluzione generalizzata al fenomeno noto come *link rot* o *reference rot*[28] e che hanno l'obiettivo di archiviare la pagina nel momento in cui questa viene citata. Tra queste: Webcite[29] risalente al 1997 e Archive.is[30] attiva dal 2012. L'ultima in ordine di tempo è una iniziativa che viene dal mondo delle biblioteche accademiche. Si tratta del servizio *perma.cc*[31] attivo dal 2013, che offre al ricercatore strumenti per proteggere dal rischio dell'inusabilità i riferimenti a risorse in rete (URL) che compaiono nelle sue pubblicazioni. In pratica quando vogliamo proteggere un riferimento è sufficiente, dopo essersi registrati, digitare su http://perma.cc l'indirizzo (URL) che vogliamo proteggere (ad esempio: http://www.bncf.firenze.sbn.it/). Il servizio – gestito dalla *Harvard Law School Library* con altre biblioteche associate – provvede ad archiviare il contenuto della pagina referenziata e a fornirci un link permanente alla risorsa archiviata (nel nostro caso: https://perma.cc/VK9H-WFMH). Il link permanente ci indirizza

[26] https://en.wikipedia.org/wiki/Help:Permanent_link.
[27] Jonathan Zittrain, Kendra Albert, Lawrence Lessig, *Perma: Scoping and Addressing, cit.*
[28] Ivi: «Link rot refers to the URL no longer serving up any content at all. Reference rot, an even larger phenomenon, happens when a link still works but the information referenced by the citation is no longer present, or has changed».
[29] http://www.webcitation.org.
[30] http://archive.is/.
[31] http://perma.cc.

alla pagina citata così come si presentava al momento della citazione («captured September 2, 2016 6:55 a.m.»); viene offerta anche la possibilità di visualizzare la pagina web in formato immagine (*screenshot view*) (https://perma.cc/VK9H-WFMH?type=image).

Il servizio *perma.cc* – che a oggi conta oltre 300.000 risorse archiviate – si propone la conformità agli standard dell'archiviazione del web. Occorre precisare che l'archiviazione della pagina citata è puntuale: i collegamenti presenti nella pagina referenziata non vengono presi in conto. Inoltre la pagina che si vuole archiviare deve essere accessibile all'applicazione (*crawler*) di *perma.cc* che archivia le pagine: non si possono archiviare pagine di siti protetti da password o di siti che applicano il "protocollo di esclusione robot"[32].

Internet Archive: raccolta automatica del web e problemi giuridici[33]

Quando si parla di archiviazione del web il servizio sicuramente più noto è Internet Archive[34] con oltre 505 miliardi di pagine web raccolte. Qui l'archiviazione non è puntuale, ma l'ambizione è quella di fotografare, con l'utilizzo di una applicazione denominata *crawler*, periodicamente tutto il web accessibile. Come è noto il *crawler* (o *spider* o *harvester*) può essere visto come un browser (ad esempio Chrome, Firefox, Safari o Internet Explorer) che viene attivato in maniera automatica. In pratica vengono forniti in input uno più indirizzi di rete (URL); il *crawler* provvede quindi alla chiamata automatica di tali indirizzi e registra (raccoglie) le risposte ottenute. Se la riposta ottenuta (tipicamente una pagina HTML) contiene altri indirizzi (URL) il *crawler* – compatibilmente con le istruzioni ricevute – provvede ad attivare iterativamente nuovi cicli di raccolta.

Normalmente tutti i *crawler* rispettano le regole del "protocollo di esclusione robot". Ogni sito ha infatti la possibilità di limitare in tutto o in parte la raccolta automatica esponendo le regole di accesso in un file apposito (robots.txt) posto nella radice del sito, oppure pagina per pagina in opportuni campi HTML contenenti meta-informazioni sulla pagina stessa (campi META). Inoltre, tutti i *crawler* di norma accompagnano la propria richiesta fornendo anche il proprio nome o *User-agent*. In questo modo un sito può limitare selettivamente la raccolta automatica.

Occorre essere consapevoli che dal punto di vista del diritto d'autore la raccolta automatica si configura come una copia: le pagine vengono infatti trasferite – di solito mediante il protocollo HTTP – dal server web alle memorie di massa gestite dal *crawler*. La copia è esplicitamente vietata dalla normativa sul diritto d'autore.

[32] https://it.wikipedia.org/wiki/Protocollo_di_esclusione_robot>. Sul protocollo di esclusione robot si veda anche il prossimo paragrafo.

[33] Riporto in questo paragrafo anche mie considerazioni apparse in: <http://www.aib.it/aib/cg/gbdigd05.htm3>.

[34] http://archive.org/web/.

Solo l'autore ha infatti «il diritto esclusivo di autorizzare o vietare la riproduzione diretta o indiretta, temporanea o permanente, in qualunque modo o forma, in tutto o in parte» (art. 2 Direttiva 2001/29 CE del 22 maggio 2001). Per evitare sanzioni all'utente che semplicemente naviga tra i siti web, ma che per vedere deve comunque copiare (in pratica navigare è utilizzare una tecnologia di raccolta automatica – sia pure in maniera limitata e temporanea), la normativa comunitaria prevede eccezioni per gli «atti di riproduzione temporanea privi di rilievo economico proprio» e che siano «parte integrante e necessaria e essenziale di un procedimento tecnologico», allo scopo di consentire la «trasmissione in rete» e «l'uso legittimo» (art 5, c. 1 Direttiva 2001/29 CE del 22 maggio 2001, o per le banche dati art. 6 Direttiva 96/9 CE dell'11 marzo 1996).
I motori di ricerca come ad esempio Google quindi non godrebbero, a rigore, di questa eccezione. Il fatto di rispettare le scelte di chi pubblica (protocollo di esclusione robot) non risolve il problema. Il detentore dei diritti non è tenuto ad applicare queste tecniche e il silenzio non può essere interpretato come assenso alla raccolta automatica. Tuttavia nessuno fino a oggi ha seriamente sostenuto l'illegalità dei motori di ricerca. Il web senza motori di ricerca perderebbe gran parte del suo valore.
Internet Archive è una fondazione senza scopo di lucro che archivia il risultato della raccolta automatica dei siti web dal 1996 e li mette a disposizione – di solito dopo un anno dalla raccolta – attraverso la sua Wayback Machine. Internet Archive oltre ad avere le caratteristiche tipiche dei motori di ricerca (in particolare, rispetto del protocollo di esclusione robot) offre anche la possibilità di decisioni successive alla raccolta: chi non volesse più essere archiviato e ripubblicato può chiedere la rimozione delle proprie pagine già pubblicate. Internet Archive offre un importante servizio di salvataggio dello spazio web ed è a oggi tra i primi 400 siti consultati al mondo[35].

Archive-it e altri servizi di raccolta del web

Internet Archive offre nel campo della raccolta automatica altri due servizi di sicuro interesse. Il primo – proposto nel 2013 – si chiama *Save the page now*[36] e si presenta come una soluzione alternativa per il *link rot*. Il secondo servizio è *Archive-It* ed è attivo dal 2006. Si tratta di un servizio che viene offerto a pagamento. Attraverso un'applicazione web fornita da Internet Archive ai sottoscrittori (istituzioni, ma anche privati) è possibile raccogliere, catalogare, indicizzare siti

[35] Pagina archiviata <https://perma.cc/5QSZ-Z8T5> e pagina corrente <http://www.alexa.com/siteinfo/archive.org>. La posizione del sito in Alexa potrebbe cambiare nel tempo.
[36] http://archive.org/web/.

web e, nel corso di 24 ore, accedere interamente all'archivio e ottenere una copia dei dati. Con *Archive-It* Internet Archive mette a disposizione tutta la sua esperienza e la sua infrastruttura tecnologica per gli obiettivi specifici di una istituzione. Si possono creare collezioni "private" (non disponibili su Internet Archive, ma disponibili per gli utenti delle istituzioni che curano quelle determinate collezioni). Molte biblioteche nazionali, ad esempio, usano il servizio per la cattura selettiva dello spazio web nazionale[37].

Archive-it offre strumenti aggiuntivi che migliorano la qualità della raccolta automatica. Del resto lo spazio web è oggi profondamente cambiato rispetto al 1996 e le tradizionali applicazioni di raccolta automatica non riescono a catturare tutta la complessità delle pagine. Dal 2004 Archive-It fornisce ai sottoscrittori insieme al *crawler* tradizionale (*Heritrix*) anche una nuova applicazione chiamata *Umbra*[38] in grado di migliorare la raccolta automatica di pagine complesse, comprese quelle dei social network (Facebook, Twitter ecc).

A differenza di Internet Archive, dove la ricerca può essere fatta solo conoscendo l'indirizzo web (o parte dell'indirizzo), le collezioni archiviate con Archive-it sono indicizzate anche a livello di full-text.

Altri due servizi in questo campo possono essere visti come alternative ad Archive-It. Il primo viene offerto dalla Internet Memory Foundation (IMF)[39], una fondazione senza scopo di lucro con sede ad Amsterdam e a Parigi attiva dal 2004 (prima del 2010 si chiamava European Archive Foundation). Nel corso della sperimentazione della raccolta dei siti con dominio di primo livello .it effettuata dalla Biblioteca Nazionale Centrale di Firenze nel 2006, IMF è stata il partner tecnologico del progetto[40]. Un altro servizio da ricordare in questo contesto è quello offerto da Hanzo Archives, che si occupa dal 2005 di raccolte web "difendibili" dal punto di vista legale[41].

La diffusione degli strumenti e dei servizi per il web archiving

Recentemente si assiste a una forte diffusione di strumenti che facilitano l'archiviazione del web e la rendono accessibile a un numero di utenti sempre più vasto. Qui di seguito alcuni tra quelli ritenuti più significativi:

1. *wget*[42]: si tratta di un'applicazione a linea di comando presente nei sistemi operativi Linux e Mac che permette di catturare risorse web e che a partire

[37] https://archive-it.org/explore?fc=organizationType%3AnationalInstitutions.
[38] https://webarchive.jira.com/wiki/display/ARIH/Introduction+to+Umbra.
[39] http://internetmemory.org/en/.
[40] Giovanni Bergamin, *La raccolta dei siti web: un test per il dominio "punto.it"*, «Digitalia», 2006, n. 2, p.170-174, <http://digitalia.sbn.it/article/view/306/199>.
[41] http://www.hanzoarchives.com/.
[42] https://en.wikipedia.org/wiki/Wget.

dalla versione 1.12 (del 2012) permette anche la creazione di raccolte in formato WARC;
2. *webrecorder.io*[43]: si tratta di un servizio che permette la creazione di WARC mentre si sta esplorando un sito e che si propone registrazioni interattive ad alta fedeltà. Anche qui come con *Umbra* – vista precedentemente – l'obiettivo è quello di far fronte a pagine web diventate sempre più complesse. *Webrecorder* – il cui motto è «web archiving for all» – offre agli utenti registrati anche un servizio di ospitalità sui propri server delle collezioni create, che possono essere dichiarate pubbliche (aperte alla fruizione di tutti) oppure private (accessibili solo a chi le ha create).
3. *warcreate*[44]: si tratta di un'estensione per il browser Chrome che permette il salvataggio in formato WARC della pagina che si sta visualizzando.

Una segnalazione particolare deve essere fatta per il progetto Memento[45], che propone un servizio generalizzato per informare l'utente sull'esistenza di versioni differenti di una determinata risorsa web. Il servizio prevede che gli archivi del web rendano disponibili in modalità standard (RFC 7089[46]) l'indice delle versioni delle pagine archiviate. Il protocollo previsto da Memento può essere applicato non solo ad archivi di pagine web, ma anche a siti che conservano versioni differenti dello stesso documento (ad esempio, come abbiamo visto, Wikipedia o gli standard del World Wide Web Consortium[47]). Un esempio di servizio che usa il protocollo Memento e che estende un'unica interrogazione su più archivi web si può trovare in http://timetravel.mementoweb.org/.

Prospettive in Italia

Come si è cercato di illustrare, sono ormai molte le esperienze di archiviazione del web: da quelle di ambito sopranazionale a quelle specificamente legate a un particolare servizio; da quelle focalizzate sull'archiviazione del "dominio" nazionale, a quelle progettate in funzione della legislazione sul deposito legale (non necessariamente coincidenti in tutto e per tutto con le precedenti).
In Italia, come testimonia la pagina archiviata da Internet Archive (fig. 2)[48], la Biblioteca Nazionale Centrale di Firenze ha partecipato alla costituzione nel 2003 dell'IIPC (International Internet Preservation Consortium)[49]. In particolare, le po-

[43] https://webrecorder.io/.
[44] http://warcreate.com/.
[45] http://www.mementoweb.org/guide/quick-intro/.
[46] https://tools.ietf.org/html/rfc7089.
[47] https://www.w3.org/blog/2016/08/memento-at-the-w3c/.
[48] http://web.archive.org/web/20060214202634/http://netpreserve.org/about/members.php.
[49] http://www.netpreserve.org/.

Figura 2. *Pagina catturata il 14.2.2006 dal sito IIPC*

tenzialità dell'archiviazione del web nel campo del deposito legale sono state fin da subito studiate e applicate. È del 2006 la sperimentazione della raccolta dei siti appartenenti al dominio di primo livello .it, con oltre 7 Terabyte archiviati e 240 milioni di pagine web raccolte[50].

Invece, quello che è sicuramente mancato in Italia in questi ultimi anni è stato un forte investimento nel campo del dell'archiviazione del web sul modello francese e

[50] Giovanni Bergamin, *La raccolta dei siti web, cit.* La collezione è fruibile in: <http://collections.europarchive.org/bncf>.

inglese. La legge sul deposito legale (L.106/2004) estende l'obbligo di deposito anche ai «documenti diffusi tramite rete informatica»; d'altra parte, per una sfortunata serie di circostanze, all'epoca della sua emanazione non si riuscì ad evitare che nella legge venisse inserita la cosiddetta "clausola costo zero"[51] – clausola che ha comportato non pochi problemi, alla prova dei fatti, riguardo alla gestione del deposito dei documenti di tipo tradizionale. Il successivo regolamento applicativo della legge (D.P.R. 252/2006), all'art. 37, ha previsto per i documenti diffusi in rete un periodo di sperimentazione su base volontaria, e al termine di questa un successivo decreto volto a disciplinarne i criteri e le modalità di raccolta. La sperimentazione è stata chiusa a fine 2015, e al momento della stesura di questo articolo è in preparazione il regolamento dedicato al deposito del digitale nativo. Una sfida da non sottovalutare, in questa circostanza, sarà quella avviare un adeguato processo di comunicazione, anche al fine di individuare opportune risorse e, a regime, garantire la sostenibilità dei processi di archiviazione.

Sul piano tecnico invece, nell'ambito della sperimentazione da poco conclusa le tecnologie condivise in ambito IIPC (Heritrix[52] per la raccolta automatica e OpenWayback[53]) sono state largamente usate per tipologie di oggetti selezionati (pubblicazioni ad accesso aperto). L'art. 37, c. 2 del citato D.P.R. 252 prevede come modalità di deposito anche la raccolta automatica. Durante la sperimentazione quest'ultima è stata effettuata con successo per quanto riguarda le tesi di dottorato in formato digitale (sono ad oggi 35 le università aderenti), e per gli articoli di riviste ad accesso aperto (80 titoli).

Tra i punti che il nuovo regolamento dovrà prevedere sarà essenziale la facoltà per le istituzioni depositarie di archiviare il web anche senza tener conto del già descritto "protocollo di esclusione robot". A questo proposito, è opportuno ricordare che già l'art. 38 del D.P.R. 252 prevede che in ogni caso l'accessibilità dei documenti depositati e raccolti avvenga «nel rispetto delle norme sul diritto d'autore e sui diritti connessi».

Non si ritiene opportuno – e obiettivamente non sarebbe realistico – che le istituzioni depositarie italiane progettino di replicare quelle raccolte del web su larga scala che vengono già effettuate da altri soggetti come Internet Archive: con Internet Archive sarebbe opportuno invece prendere accordi, sull'esempio di altre biblioteche nazionali. I programmi di archiviazione del web dovrebbero puntare su una lista di priorità pubblicamente definita. Lo stesso art. 37 – nel regolare la sperimentazione – prevede la seguente lista di priorità, che potrebbe essere assunta come base di partenza:

[51] Ovvero che dalla legge non dovessero derivare nuove spese a carico dello Stato: cfr. Anna Maria Mandillo, *Il difficile percorso della nuova legge sul deposito legale*, «AIB Notizie», A. 16 (2004), n. 6, p. III-IV.

[52] https://webarchive.jira.com/wiki/display/Heritrix.

[53] https://github.com/iipc/openwayback/wiki.

- documenti che assicurano la continuità delle collezioni delle biblioteche depositarie;
- documenti concernenti la produzione scientifica delle università, dei centri di ricerca e delle istituzioni culturali;
- documenti elaborati e messi in rete da soggetti pubblici;
- siti che si aggiornano con più frequenza, ovvero contenuti in siti maggiormente citati da altri siti.

Con l'approvazione da parte del Ministero dei beni e delle attività culturali e del turismo (D.M. 28 gennaio 2016) per gli anni 2016-2018 del Programma triennale (previsto dall'articolo 1, comma 9, della legge 23 dicembre 2014, n. 190) è stato finanziato anche un progetto presentato dalla Biblioteca Nazionale Centrale di Firenze, per la raccolta dei siti web delle istituzioni culturali[54]. Potrebbe essere questa l'occasione per l'Italia di consolidare l'esperienza di archiviazione del web in vista dell'attuazione del deposito legale del digitale, e di tornare a partecipare attivamente ai lavori dell'International Internet Preservation Consortium[55].

Libraries and archives are increasingly aware that web archiving is an essential component of their mission. The number of documents (including primary sources) available on the web is growing continually; on the other hand, most of them are already lost for ever. Digital preservation is not simply a technological problem, but also requires cooperation between producers, users and memory institutions.

"Preserving digital documents" is the message which underpins the program of an annual workshop that the National Association of Italian Archivists has been organising since 2010. The sixth Workshop was dedicated to web-archiving. Preserving the web is possibly the most urgent challenge that institutions in charge of the preservation of memory have to tackle today. The organisers were keen to approach the subject from a wide perspective: the web as a whole, as part of each country's national memory. Web-archiving as it is being carried out in two national libraries (France and UK) was presented as two case studies, allowing a comparison of different approaches and solutions yet relating to common issues, but the studies also served as a model and a starting point for an analysis of the current Italian situation.

Web archiving tools and services have been increasingly developed in recent times: web archiving is now accessible to an ever larger number of users, and memory institutions can rely on a core infrastructure. A new regulation for the legal deposit of digital publications, as required by D.P.R. 252/2006, will hopefully give Italian institutions the opportunity to take advantage of recent developments and the results from successful trials.

[54] http://www.beniculturali.it/mibac/multimedia/MiBAC/documents/1460630545725_ALLEGATO_ALLO_SCHEMA_DI_DECRETO.pdf.

[55] http://www.netpreserve.org/.

L'ultima consultazione dei siti Web è avvenuta nel mese di dicembre 2016.