DigItalia

ISSN 1972-6201 Anno X, Numero 1/2 - **2015**

RIVISTA DEL DIGITALE NEI BENI CULTURALI

ICCU-ROMA

Nuove prospettive per il Web archiving: gli standard ISO 28500 (formato WARC) e ISO/TR 14873 sulla qualità del Web archiving

Stefano Allegrezza
Università degli Studi di Udine

Il Web archiving è un argomento di forte attualità in quanto, come è noto, se non si individuano in breve tempo soluzioni efficaci e sostenibili nel lungo periodo, si rischia di perdere per sempre quello che si è prodotto e pubblicato sul Web negli ultimi venti-trenta anni, dal momento che tale materiale è caratterizzato da un'estrema mutevolezza e dinamicità e spesso interi siti Web cambiano o scompaiono nel giro di poco tempo. Le soluzioni che sono state proposte fino ad oggi sono parziali e non sempre hanno raggiunto l'obiettivo. Tuttavia, recentemente ci sono state due novità che sembrerebbero poter assicurare prospettive migliori: si tratta da una parte della proposta di un formato elettronico specificatamente pensato per l'archiviazione del Web (il formato WARC), dall'altra della pubblicazione di una specifica norma ISO dedicata alla qualità nella conservazione del Web (ISO/TR 14873:2013). La rilevanza dell'argomento per il settore dei beni culturali è tale che è opportuno fare un po' di chiarezza su queste tematiche analizzando sia lo stato dell'arte che le prospettive future.

Introduzione

Le questioni legate al *Web archiving* costituiscono un argomento di grande attualità dal momento che si rischia di perdere per sempre buona parte di quanto il Web ha prodotto negli ultimi decenni se non si individuano in breve tempo soluzioni efficaci e sostenibili nel lungo periodo. Infatti, il Web ha una natura effimera ed è caratterizzato da una grande mutevolezza e dinamicità; non è raro che interi siti cambino radicalmente o scompaiano nel giro di poco tempo[1], spesso senza lasciare alcuna traccia di sé. Alcuni contenuti, specialmente quelli presenti

[1] Diversi studi hanno esaminato la vita media delle pagine Web, con risultati che vanno da 44 giorni a 75 o 100 giorni. Si veda, tra gli altri: Brewster Kahle, *Preserving the Internet*, «Scientific American Special Report», 1998, <http://web.archive.org/web/19980627072808/http://www.sciam.com/0397issue/0397kahle.html>; Rick Weiss, *On the Web, Research Work Proves Ephemeral*, «Washington Post», 23 novembre 2003, <http://faculty.missouri.edu/~glaserr/205f03/Article_WebPub.html>.

sui siti di news o sui social media, mutano con una rapidità tale da mandare in crisi qualsiasi strategia di archiviazione[2].
Coscienti di questa minaccia, le istituzioni – specialmente quelle che operano nel settore culturale – hanno effettuato investimenti notevoli per sviluppare gli strumenti e le tecniche necessarie per supportare soluzioni di *Web archiving* su larga scala, con l'obiettivo di trovare le modalità più adeguate per raccogliere e conservare il materiale pubblicato in maniera sempre più imponente sul World Wide Web. Le soluzioni che sono state proposte fino ad oggi sono parziali e non sempre hanno raggiunto l'obiettivo. Tuttavia, recentemente l'attenzione si è focalizzata su due proposte che sembrerebbero poter fornire nuove soluzioni: si tratta da una parte dell'individuazione di un formato specificatamente pensato per l'archiviazione del Web (il formato WARC), dall'altra della pubblicazione di una specifica norma ISO dedicata alla qualità nell'archiviazione del Web (ISO/TR 14873:2013). La rilevanza dell'argomento per il settore dei beni culturali è tale che è opportuno fare chiarezza su queste tematiche analizzando sia lo stato dell'arte che le prospettive future.

Lo standard ISO 28500:2009 (formato WARC)

Fino a qualche anno fa gli strumenti software utilizzati per effettuare "catture" del Web (i cosiddetti *Web crawler*, detti anche semplicemente *crawler*, *spider* o *robot*)[3] organizzavano le informazioni raccolte codificandole nel modo più disparato ed il più delle volte utilizzando formati proprietari. Per questo si è ben presto sentita l'esigenza di individuare un formato che costituisse un riferimento semplice e nello stesso tempo affidabile da utilizzare per la cattura dei contenuti Web, per la loro gestione e per l'interscambio di tali materiali tra organizzazioni diverse. Il formato WARC (Web ARChive), riconosciuto ufficialmente standard ISO il 15 maggio 2009 con la pubblicazione della norma ISO 28500:2009[4] "*Information and documentation — WARC file format*", risponde in pieno a queste esigenze.
Questo formato costituisce un'evoluzione del precedente formato ARC creato nel 1996 da Brewster Kahle[5] e Mike Burner ed utilizzato fin da allora da Internet

[2] In alcuni casi, i contenuti diventano inaccessibili quando un sito viene riprogettato; in altri casi, il contenuto semplicemente scompare dal Web quando viene sostituito con informazioni più aggiornate, con la conseguenza che l'utente si trova davanti a link non funzionanti o al "classico" errore di tipo "404 page not found".

[3] Il *crawler* è essenzialmente un *browser* con funzionalità minime e privo dell'interfaccia grafica, che ha il compito di "simulare" il comportamento di un utente che vuole visitare, ad esempio, una serie di siti Web. Esso prende in input una lista di URL (Uniform Resource Locator) di pagine Web e le visita una per una seguendone anche i link. Tra i *crawler* più conosciuti vanno citati Heritrix, HTTrack e Googlebot.

[4] L'ultima revisione dello standard risale al 2014.

[5] Brewster Kahle è il fondatore di Internet Archive, la biblioteca digitale *non profit* che ha lo scopo dichiarato di consentire un "accesso universale alla conoscenza". Tra le collezioni di materiali

Archive e da numerose istituzioni culturali per archiviare il materiale – costituito dalle pagine Web e dai link di collegamento tra di esse – raccolto dai *crawler* durante le loro esplorazioni del Web (*Web crawls*). Alla base di questa estensione vi sono motivazioni riconducibili alle esperienze delle organizzazioni che facevano parte dell'International Internet Preservation Consortium (IIPC)[6]; esse stavano sperimentando difficoltà via via maggiori nelle operazioni di archiviazione e di gestione del crescente volume di informazioni prelevate da Internet; per questo il gruppo di lavoro sugli standard dell'IIPC sottomise al Working Group 12 del Sottocomitato 4 del Comitato Tecnico TC 46 dell'ISO (in breve: ISO TC46/SC4/WG12) un progetto denominato "WARC file format". Il progetto venne accolto dall'ISO come New Working Project (NWP) nel maggio 2005. Da allora tutte le fasi del processo di standardizzazione si sono rapidamente succedute fino a giungere, il 15 maggio 2009, alla pubblicazione delle specifiche del formato come standard ufficiale ISO[7].

Le specifiche del formato WARC stabiliscono le modalità attraverso le quali è possibile concatenare gli oggetti digitali che tipicamente costituiscono una pagina Web o un intero sito Web ed i relativi metadati in un unico file d'archivio, il file WARC appunto. In questo senso il formato può essere utilizzato per creare applicazioni per la "raccolta" (*harvesting*), la gestione, l'accesso e lo scambio dei contenuti Web tra le varie istituzioni archivistiche. Inoltre, i file WARC possono essere utilizzati per archiviare qualsiasi genere di contenuto digitale, recuperato sia attraverso il protocollo Hypertext Transfer Protocol (HTTP) che mediante altri protocolli, come il File Transfer Protocol (FTP).

Rispetto al precedente formato ARC, il formato WARC possiede specifiche definite in maniera più puntuale, è più flessibile ed offre nuove funzionalità, come la memorizzazione delle richieste HTTP, la possibilità di inserire metadati arbitrari, l'assegnazione di un identificativo per ogni file trovato, la gestione dei duplicati e delle migrazioni dei record e la segmentazione dei contenuti raccolti su più record.

digitali include anche i siti Web. L'interfaccia Web utilizzata da Internet Archive per la navigazione sui siti Web che sono stati archiviati nei vari istanti della loro vita (mediante altrettanti *snapshot*) dal software di acquisizione prende il nome di *Wayback Machine*.

6 L'International Internet Preservation Consortium è un consorzio internazionale il cui obiettivo principale è quello di acquisire, rendere disponibili e conservare per le generazioni future le informazioni raccolte da Internet. Costituitosi nel luglio 2004 per coordinare gli sforzi necessari per preservare i contenuti Internet per il futuro, conta oggi un numero elevato di istituzioni (biblioteche, archivi, musei e altre organizzazioni culturali) di vari paesi del mondo. Il sito Web del consorzio è raggiungibile all'indirizzo: <http://netpreserve.org>.

7 Il formato è attualmente mantenuto dal Sottocomitato 4 – Interoperabilità tecnica del Comitato tecnico ISO/TC 46 – Informazione e documentazione.

La struttura di un file WARC

Il formato WARC appartiene alla categoria dei formati di tipo "contenitore" e la struttura di un file WARC è piuttosto elementare: è costituito da una semplice successione di record (Fig. 1) il primo dei quali di solito è una semplice descrizione dei record seguenti.

Ciascun record è composto da un *text header* seguito da un *content block*, che costituisce il contenuto vero e proprio. Il *text header* è costituito da una prima linea che dichiara la conformità a una specifica versione del formato (ad esempio, WARC/1.0) seguita da un certo numero di campi del tipo "nome:valore" che servono per fornire varie informazioni sul record (come l'URI del sito oggetto della cattura, la data di cattura ecc.); il tutto è concluso con una riga vuota che serve da separatore prima del blocco del contenuto. Come separatore tra un record e il successivo vengono utilizzate due linee vuote. Il *content block* è costituito dal risultato di un tentativo di recupero dal Web (pagine Web, immagini, informazioni di reindirizzamento degli URL, risultati della ricerca degli host DNS ecc.) o da altri dati (ad esempio, metadati o contenuti trasformati) che forniscono ulteriori informazioni sui contenuti archiviati.

A seconda del tipo di contenuto, i record possono appartenere ad una delle seguenti otto categorie: *warcinfo, response, resource, request, metadata, revisit, conversion e continuation*; il tipo di record viene specificato nel campo WARC-Type. Ad esempio, viene utilizzato un record di tipo *request* per memorizzare le informazioni inviate al server Web che ospita un determinato sito nella fase di richiesta di una determinata pagina Web; le risposte ottenute vengono memorizzate

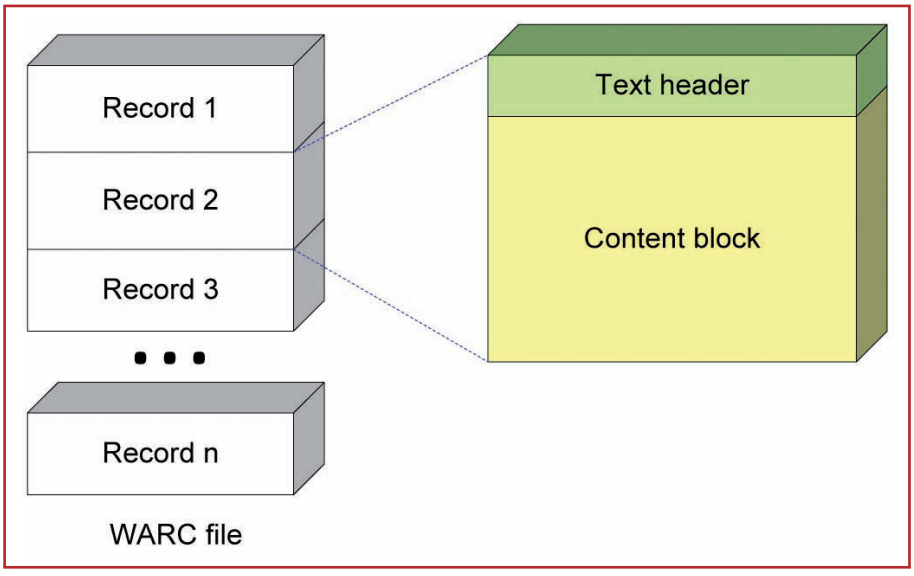

Figura 1: *La struttura di un file WARC*

in un successivo record di tipo *response*. Una immagine catturata dal Web costituisce il *content block* di un record di tipo *resource* (Fig. 2).

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://www.digitalia.org/images/immagine.jpg
WARC-Warcinfo-ID: <urn:uuid:d7ae5c10-e6b3-4d27-967d-34780c58ba39>
WARC-Date: 2015-06-12T16:15:20Z
WARC-Block-Digest: sha1:UZY6ND6CCHXETFVJD2MSS7ZENMWF7KQ2
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-IP-Address: 193.141.213.58
WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: application/http;msgtype=response
WARC-Identified-Payload-Type: image/jpeg
Content-Length: 1902

HTTP/1.1 200 OK
Date: Tue, 19 Sep 2015 17:18:40 GMT
Server: Apache/2.2.31
Last-Modified: Mon, 16 Jun 2013 22:28:51 GMT
Accept-Ranges: bytes
Content-Length: 1662
Connection: close
Content-Type: image/jpeg

[image/jpeg – dati in binario]
```

Figura 2: *Esempio di record di tipo "response" (adattamento dallo standard ISO 28500:2009, Annex C "Examples of WARC records")*

La dimensione del file che si ottiene varia, evidentemente, a seconda dell'estensione del sito da catturare; la dimensione massima consigliata è intorno ad 1 GB, più che sufficiente per buona parte dei siti di piccole-medie dimensioni oggi esistenti. Nel caso in cui un sito Web abbia una dimensione maggiore è possibile suddividerlo in due o più file WARC.
Per il formato WARC è stato anche coniato un identificatore di tipo MIME; infatti, i file che si ottengono possiedono come MIME type la stringa *application/warc*.
Non rientra tra le finalità del presente articolo approfondire ulteriormente i dettagli tecnici del formato[8]; tuttavia, è importante metterne in evidenza la semplicità e la trasparenza[9], caratteristiche che costituiscono certamente un elemento a favore della possibilità di conservazione a lungo termine.

[8] Per un approfondimento sugli aspetti tecnici del formato si rimanda alla lettura della norma ISO 28500:2009.

[9] Per "trasparenza" di un formato elettronico si intende il grado di semplicità con cui è possibile fruire degli oggetti digitali codificati secondo quel formato, ad esempio utilizzando semplici strumenti di base.

Analisi del formato WARC ai fini della conservazione a lungo termine

Come evidenziato da numerosi studi[10], un formato compatibile con un processo di conservazione digitale deve possedere innanzitutto le caratteristiche della non proprietà, dell'apertura, della standardizzazione e della trasparenza (Fig. 3).

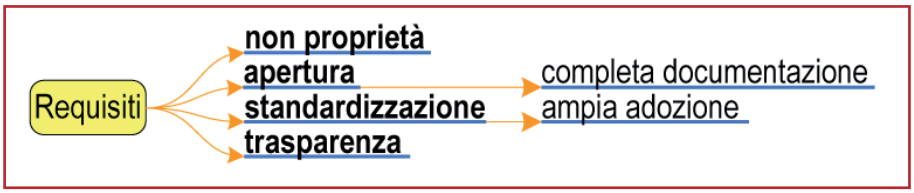

Figura 3: *I requisiti di base che un formato deve possedere per essere compatibile con un processo di conservazione a lungo termine*

Il formato WARC rispetta pienamente questi requisiti; infatti:

- è "non proprietario" in quanto è stato sviluppato sotto gli auspici dell'International Internet Preservation Consortium (IIPC); attualmente il gruppo di lavoro ISO responsabile del suo mantenimento è il TC46/SC4/WG12, che ha sede presso la Biblioteca nazionale francese[11];
- è "aperto" dal momento che le sue specifiche sono pubblicamente disponibili (si possono acquistare sul sito dell'International Standard Organization (ISO); le specifiche del formato nella versione *final draft* (ISO/FDIS 28500) che ha costituito la base per il processo di approvazione da parte dell'ISO sono, invece, liberamente disponibili[12]); inoltre, il formato è completamente documentato tramite le sue specifiche;
- è "standard" dal momento che è stato riconosciuto standard ISO 28500:2009; inoltre è ad oggi ampiamente adottato;
- è "trasparente", trattandosi di un formato che si limita a offrire un contenitore per gli oggetti digitali raccolti dal Web (che possono essere ovviamente nei formati più disparati).

Oltre a queste caratteristiche di base, il formato WARC gode di ulteriori interessanti proprietà[13]: non è sottoposto ad alcuna restrizione (in termini di licenze o brevetti); non vi sono meccanismi tecnici di protezione che potrebbero compro-

[10] Per un approfondimento si veda, tra gli altri, Stefano Allegrezza, *Requisiti e standard dei formati elettronici per la produzione di documenti informatici*, «Archivi & Computer», 19 (2009), n. 2-3, p. 42-82.

[11] Il sito Web di riferimento per le attività del gruppo è: <http://bibnum.bnf.fr/WARC/>.

[12] Il *final draft* è liberamente disponibile sul sito della Biblioteca nazionale francese all'indirizzo: <http://bibnum.bnf.fr/ WARC/WARC_ISO_28500_version1_latestdraft.pdf>.

[13] Si veda la sezione del sito della Library of Congress dedicata al formato WARC: <http://www.digitalpreservation.gov/ formats/fdd/fdd000236.shtml>.

metterne l'utilizzo in un prossimo futuro; infine, è auto-documentato (ciascuna risorsa – HTML, JPG, GIF ecc. – contenuta all'interno di un file WARC è corredata dalle informazioni su di essa).

Ormai tutti i più diffusi sistemi per il *Web archiving* hanno adottato il formato WARC per l'archiviazione dei materiali "catturati". Ad esempio, Heritrix[14], il *crawler* utilizzato da Internet Archive e da numerose istituzioni culturali (Fig. 4), adotta il WARC come formato predefinito per le sue "catture".

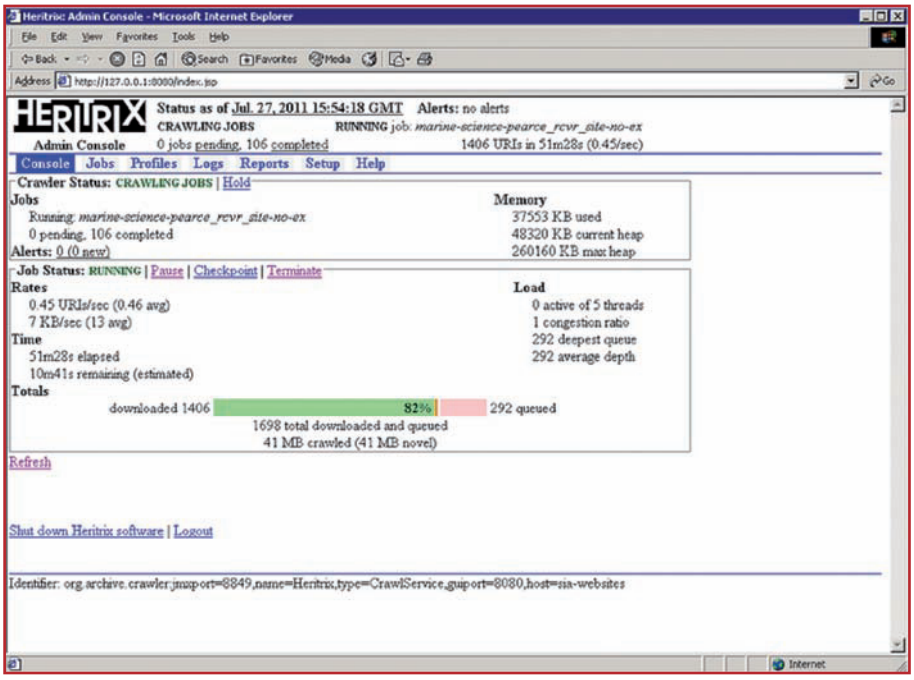

Figura 4: *La schermata di amministrazione di Heritrix*

Per la fruizione dei siti Web archiviati nel formato WARC sono disponibili estensioni per i più comuni browser e numerosi applicativi[15]. Tra questi merita sicuramente una segnalazione WERA (Web ARchive Access)[16], una soluzione il cui sviluppo è stato sponsorizzato dall'International Internet Preservation Consortium (IIPC) ed è

[14] Heritrix è il *crawler open source* sviluppato nell'ambito del progetto gestito da Internet Archive; è utilizzato per la raccolta dei siti Web da parte di numerose istituzioni, come Austrian National Library, Bibliothèque Nationale de France, British Library, Web Archiving Service della California Digital Library, Internet Memory Foundation, Library and Archives Canada, Library of Congress, National Library of New Zealand, Koninklijke Bibliotheek- Nationale Library van Nederland.

[15] Oltre all'applicativo WERA citato nel seguito, vanno segnalati i WARC tools per la gestione e lo scambio dei file WARC, la Wayback Machine di Internet Archive, NutchWAX e altri strumenti di ricerca che consentono l'accesso ai materiali raccolti.

[16] Il sito Web di riferimento è: <http://archive-access.sourceforge.net/projects/wera>.

liberamente disponibile per chiunque voglia utilizzarla. Essa consente la navigazione e la ricerca nelle collezioni di siti Web archiviati e funziona in maniera analoga alla *Wayback Machine* dell'Internet Archive (ma a differenza di quest'ultima consente anche la ricerca *full-text* all'interno dell'archivio Web).

Lo standard ISO/TR 14873:2013

Per rispondere alla necessità di disporre di linee guida affidabili sulla gestione e valutazione delle attività di *Web archiving*, il 1 dicembre 2013 è stata pubblicata la norma ISO/TR 14873:2013 *Information and documentation — Statistics and quality issues for Web archiving*[17].


TECHNICAL REPORT

ISO/TR 14873

First edition 2013-12-01

Information and documentation — Statistics and quality issues for web archiving

Information et documentation — Statistiques et indicateurs de qualité pour l'archivage du web

ISO

Reference number ISO/TR 14873:2013(E)

© ISO 2013

Figura 5: *Il frontespizio della norma ISO/TR14873:2013*

Si tratta di un notevole passo avanti nel tentativo di standardizzare le metodiche e le tecniche che sono alla base del *Web archiving*. La norma si concentra essenzialmente sui principi e metodi del *Web archiving* e ne stabilisce la terminologia, gli indicatori statistici da rilevare e i criteri di qualità, tenendo in debita considerazione le necessità di una vasta gamma di utenti (biblioteche, archivi, musei, centri di ricerca ed istituzioni culturali). Ad esempio, la terminologia utilizzata tenta di riflettere la vasta gamma di interessi e competenze, cercando un equilibrio tra archivistica, informatica, management e biblioteconomia. È destinata in primo luogo ai professionisti direttamente coinvolti nell'archiviazione del Web ma è utile anche agli organismi di finanziamento delle istituzioni culturali e altri *stakeholders*.

La norma ISO/TR 14873:2013, composta da 54 pagine, è suddivisa in sei sezioni. Dopo le consuete sezioni iniziali costituite dal campo di applicazione (*Scope*) e dalle definizioni (*Terms and definitions*), la terza sezione (*Methods and purposes of Web archiving*) prende in esame i metodi e gli obiettivi del *Web archiving*, analizzandone i meccanismi di raccolta, le modalità di accesso e di descrizione, le stra-

[17] Il sottocomitato tecnico dell'ISO che ha curato e mantiene lo standard è l'ISO/TC 46/SC 8 *Quality - Statistics and performance evaluation*. La norma andrebbe letta in congiunzione con le altre due norme ISO 2789:2013 *Information and documentation – International library statistics* e ISO 11620:2014 *Information and documentation – Library performance indicators*.

tegie di conservazione, gli aspetti legali e le motivazioni.
La quarta sezione (*Statistics*) prende in esame tutta una serie di indicatori statistici che possono essere utilizzati per effettuare misurazioni quantitative e qualitative. Ad esempio, per misurare la dimensione di un archivio Web la norma suggerisce tre possibili indicatori: il conteggio degli URL; il conteggio dei domini o degli host; il conteggio dei byte.
Vi sono poi indicatori di tipo statistico che affrontano altri aspetti del problema; di particolare interesse sono quelli relativi alla *digital preservation*, che prendono in esame questioni come la percentuale di risorse che sono state duplicate, la percentuale di risorse perse o deteriorate, la percentuale di risorse con un formato di file ben identificato, la percentuale di formati per i quali è stata individuata una ben precisa strategia di conservazione ed, infine, la percentuale di risorse che sono state sottoposte ad un controllo antivirus.
Infine, l'ultima sezione (*Usage and benefits*) mette in evidenza i benefici che possono derivare dall'applicazione della norma a tutte le possibili figure professionali coinvolte a vario titolo nella attività connesse con il *Web archiving* (archivisti, *digital curators*, informatici, manager, etc.).
Il vantaggio innegabile che si deve riconoscere a questo standard è quello di essere riuscito a conferire sistematicità a tutto un insieme di informazioni spesso eterogenee e, fino ad ora, male organizzate.

Prospettive future

Sebbene la fase di archiviazione del Web sia di importanza fondamentale in una strategia di conservazione a lungo termine, è evidente che da sola non esaurisce il problema della conservazione, dal momento che occorre anche mantenere nel tempo l'accessibilità alle risorse raccolte e quindi affrontare tutta una serie di problemi che riguardano l'obsolescenza dei formati, quella dei supporti dove le informazioni raccolte vengono archiviate, il mantenimento delle relazioni tra i vari oggetti digitali ecc.
Oltre a ciò occorre tenere presente alcune questioni che rimangono tutt'ora irrisolte:
– innanzitutto, la sterminata estensione del Web rende di fatto impossibile effettuare una archiviazione esaustiva e, tantomeno, conservarlo nella sua interezza (anche facendo ricorso a sistemi informatici dotati di capacità di memorizzazione straordinarie);
– in secondo luogo, il Web è per sua stessa natura estremamente dinamico (i contenuti delle pagine Web dei siti cambiano continuamente e spesso vengono generati *on the fly* a seconda delle richieste che giungono da parte degli utenti) ed è impensabile ricorrere a una soluzione che riesca a "scattare" – conservandone copia – continui *snapshot* di ciascun sito Web che si intende conservare e che muta istante dopo istante[18]; per questo è evidente che occorre necessariamente proce-

[18] Anche *Internet Archive* è in grado, con i suoi *crawler*, di raccogliere delle "istantanee" dei siti Web sulla base di un programma temporale che tiene conto di diversi fattori, ma certamente non di

dere in modo selettivo, con criteri sia di tipo tematico che legati all'importanza che determinati siti possono avere per le future generazioni a seconda delle varie chiavi di lettura (storica, sociologica, politica, economica ecc.);
– infine, vi è l'oggettiva difficoltà – se non addirittura l'impossibilità – di accedere al cosiddetto *deep Web*[19], cioè a quella porzione del Web comunemente non indicizzata dai motori di ricerca e che, secondo alcune stime è 4.000-5.000 volte più grande del cosiddetto *surface Web*, cioè della parte accessibile, per un volume totale di oltre 10 Petabyte.

Nonostante tutte queste difficoltà, si comprende quanto la conservazione di queste informazioni sia di importanza fondamentale per la futura ricerca in ogni campo delle scienze e numerose organizzazioni a livello internazionale hanno cominciato ad occuparsi di *Web archiving*[20], ottenendo risultati incoraggianti anche grazie all'impiego del formato WARC.

Conclusioni

Per molto tempo il *Web archiving* ha costituito una sfida impari per gli archivisti e i *digital curators* e ciò ha portato alla perdita di innumerevoli contenuti Web, oggi non più accessibili né conservati da alcuna istituzione.
Il formato WARC, che è aperto, non proprietario, standard, ben documentato ed ampiamente adottato rappresenta indubbiamente un notevole passo avanti verso la soluzione del problema dell'archiviazione, consentendo la raccolta, la gestione e l'interscambio delle risorse "catturate" nel Web. Pur non potendo essere additato come la soluzione definitiva a tutte le criticità, costituisce comunque una eccellente risposta al problema dell'obsolescenza dei formati elettronici dal momento che consente di ricondurlo a quello di un unico formato. Inoltre, la sua standardizzazione offre una garanzia di durata nel tempo, assicura l'interoperabilità tra il patrimonio raccolto da istituzioni archivistiche diverse e contribuisce a far sì che il *Web archiving* diventi una delle attività che possono essere svolte con efficacia dalle organizzazioni che si occupano di conservazione del patrimonio culturale.

conservare una copia di ciascun sito Web istante per istante.

[19] Il *deep Web* è costituito da vari tipi di pagine, tra cui in particolare: *pagine dinamiche*, cioè costruite al momento a seguito di valori forniti dall'utente (ad esempio, il prezzo di un biglietto del treno determinato in un certo istante); *pagine scollegate*, cioè non accessibili tramite un percorso che parta dalla home del sito, ma solo a chi ne conosca l'indirizzo; *pagine private*, accessibili solo attraverso l'inserimento di credenziali di autenticazione (*username* e *password*); *pagine ad accesso limitato*, per esempio quelle il cui accesso è protetto dai CAPTCHA. Cfr. la sezione dedicata al Web archiving sul sito Conservazione Digitale del Centro di eccellenza italiano sulla conservazione digitale <http://www.conservazionedigitale.org/wp/approfondimenti/ web-archiving-2/>.

[20] A questo proposito occorre riconoscere che la situazione delle istituzioni culturali in Italia è in netto ritardo rispetto a quanto avviene negli altri Paesi, in special modo rispetto a quelli del mondo anglosassone.

Infine, la definizione della norma ISOTR 14873:2013 costituisce una risposta alla mancanza di linee guida affidabili sulle attività di gestione e valutazione del *Web archiving* e rappresenta un notevole passo avanti nel tentativo di standardizzare le metodiche e le tecniche che ne sono alla base.
Nel complesso si tratta di due novità che contribuiscono ad alimentare la speranza di riuscire a conservare per le generazioni future l'imponente mole di informazioni presenti sul Web che altrimenti rischierebbe di andare irrimediabilmente perduta.

Web archiving is a topic of great relevance because, if we do not identify quickly effective and sustainable solutions, we may lose what has been produced and published on the Web in the last twenty/thirty years. In fact, this material is characterized by extreme changeability and dynamism and often entire Websites change or disappear in a short time. The solutions that have been proposed up to now are partial and have not always achieved the goal. However, recently there are two developments that seem to provide new tools: the proposal of a file format specifically designed for Web storage (the WARC format) and the publication of a specific ISO standard devoted to quality in the preservation of the Web (ISO/TR 14873:2013). The relevance for the cultural heritage sector is such that it is important to go into depth on these news.

L'ultima consultazione dei siti Web è avvenuta nel mese di dicembre 2015.

Riferimenti bibliografici

Burner – Kahle 1996 Mike Burner - Brewster Kahle. *ARC File Format*, 15 September 1996, <http://www.archive.org/Web/ researcher/ArcFileFormat.php>.

Hanzo Web Archiving 2014 Hanzo Web Archiving. ISO *28500 WARC The open standard for Web content preservation for eDiscovery*, disponibile previa registrazione sul sito: <http://www.hanzoarchives.com/learn/whitepapers/Web-archiving-iso-28500-warc-white-paper>.

ISO 28500: 2009 2013 *ISO 28500:2009. A new standard for the WARC file format*, «Engineer Live Magazine», 21st February 2013 <http://www.engineerlive.com/content/22130>.

Kahle 2008 Brewster Kahle. *Preserving the Internet*, «Scientific American Special Report», 1998, <http://Web.archive.org/Web/19980627072808/ http://www.sciam.com/0397issue/0397kahle.html>.

Kim – Ross 2012 Yunhyong Kim – Seamus Ross. *Digital Forensics Formats: Seeking a Digital Preservation Storage Container Format for Web Archiving, «International Journal of Digital Curation», 7(2012), n.2 <http://www.ijdc.net/index.php/ijdc/article/view/217/286>*.

Kunze 2005 John Kunze. *WARC: an archiving format for the Web*, 5th International Web Archiving Workshop (IWAW05), <http://www.iwaw.net/05/kunze.pdf>.

Niu 2012 Jinfang Niu. *Functionalities of Web Archives*, «D-Lib Magazine», 18 (2012), n. 3-4 <http://dlib.org/dlib/march12/niu/03niu2.html>.

Niu 2012 Jinfang Niu. *An Overview of Web Archiving*, «D-Lib Magazine», 18 (2012), n. 3-4 <http://dlib.org/dlib/march12/niu/03niu1.html>.

Pennock 2013 Maureen Pennock. *Web-Archiving: DPC Technology Watch Report 13-01*, Digital Preservation Coalition. 2013 <http://www.dpconline.org/component/docman/doc_download/865-dpctw13-01pdf-dpctw13-01pdf>.

Sinibaldi-Buongiorno 2012 Alessandro Sinibaldi – Paolo Bartolomeo Buongiorno. *Manuale di conservazione digitale*, Milano, Franco Angeli, 2012.

Thompson 2008 *Dave Thompson. Archiving Web Resources*, DCC Digital Curation Manual, [a cura di] Seamus Ross, Michael Day, (December 2008), <http://www.dcc.ac.uk/resource/curation-manual/chapters/Web-archiving>.

Uk National Archives 2011 Uk National Archives. *Web Archiving Guidance*, 2011, <http://nationalarchives.gov.uk/documents/ information-management/Web-archiving-guidance.pdf>.

Web archiving 2006 *Web archiving, [a cura di] Julien Masanès. Berlin:* Springer, 2006.

Weiss 2003 *Rick Wesiss. On the Web, Research Work Proves Ephemeral*, Washington Post, 23 novembre 2003, <http://faculty.missouri.edu/~glaserr/205f03/Article_WebPub.html>.

Sitografia

*The WARC File Format (ISO 28500) - Information, Maintenance, Drafts;*sezione del sito della Bibliothèque Nationale de France dedicata la formato WARC, <http://bibnum.bnf.fr/WARC/>.

WARC File Format specifications; pagina dedicata agli Internet Archive ARC access tools su Sourceforge <http://archive-access.sourceforge.net/warc>.

Heritrix 3.0 and 3.1 User Guide, pagina dedicata all'utilizzo del formato WARC da parte del *crawler* Heritrix sul sito del progetto *open-source* Heritrix; si veda, in particolare, la sezione *: Output/WARC files* <https://Webarchive.jira.com/wiki/display/Heritrix/Heritrix +3.0+and+3.1+User+Guide>.

WARC (Web ARChive), pagina dedicata al formato WARC sul sito del progetto *open-source* Heritrix; si vedano in particolare gli esempi di file WARC <https://Webarchive.jira.com/wiki/pages/viewpage.action?pageId=4817>.

OpenWayback: General Overview (OpenWayback è un'applicazione open source Java finalizzata alla ricerca e all'accesso al material Web archiviato) <https://github.com/iipc/openwayback/wiki/General-overview>.

Sezione della Library of Congress Sustainability of Digital Formats Planning for Library of Congress Collections dedicata al formato WARC, <http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml>.

Sezione dedicata al Web archiving" sul sito Conservazione digitale del Centro di eccellenza italiano sulla conservazione digitale <http://www.conservazionedigitale.org/wp/approfondimenti/Web-archiving-2/>.