Umanistica Digitale – ISSN: 2532-8816 – n. 14, 2022
S. Allegrezza – *Web e social media come nuove fonti per la storia*
DOI: http://doi.org/10.6092/issn.2532-8816/15665

Web e social media come nuove fonti per la storia

Stefano Allegrezza

stefano.allegrezza@unibo.it
Department of Cultural heritage, University of Bologna, Bologna, Italy

Abstract

Il contributo intende mettere in evidenza come negli ultimi anni l'interesse verso i temi dell'archiviazione e conservazione del web e dei social media sia cresciuto enormemente, parallelamente alla consapevolezza dell'importanza di tali "risorse" come fonti privilegiate per ricostruire la storia della nostra epoca. Come faranno gli storici del futuro a ricostruire il periodo storico che stiamo vivendo se le istituzioni della memoria non saranno capaci di archiviare e preservare i siti web e social media di enti pubblici, partiti, associazioni, organi di governo, personaggi politici, personaggi illustri in genere, dal momento che ormai tutto viene veicolato attraverso tali canali? La fragilità del web, poi, imporrebbe di agire subito ed avviare senza indugio iniziative di "web and social media archiving", pena la scomparsa di quanto è stato reso disponibile online negli ultimi anni, ma su questo punto la situazione in Italia – salvo poche eccezioni – appare molto in ritardo rispetto agli altri paesi europei ed enormemente in ritardo rispetto ai paesi dell'area anglosassone. È urgente, quindi, avviare iniziative di sensibilizzazione su questi temi e di formazione delle competenze e delle professionalità necessarie per condurre progetti di archiviazione e conservazione del web e dei social media.

This paper aims at highlighting how interest in the issues of web and social media archiving and preservation has grown enormously, in parallel with the awareness of the importance of these 'resources' as privileged sources for reconstructing the history of our era. How will the historians of the future be able to reconstruct the historical period we are living through if memory institutions are not able to archive and preserve the websites and social media of institutions, public bodies, parties, associations, government bodies, political figures, and famous people in general, given that everything is now conveyed through these channels? The fragility of the web, then, would require immediate action and the launch of 'web and social media archiving' initiatives without delay, on pain of the disappearance of all that has been made available online in recent years, but on this point the situation in Italy - with a few exceptions - appears to lag far behind other European countries and enormously behind the Anglo-Saxon countries. There is therefore an urgent need to launch initiatives to raise awareness on these issues and to train the skills and professionalism required to conduct web and social media archiving and preservation projects.

Introduzione

L'interesse verso i temi dell'archiviazione e conservazione del web e dei social media è andato via via crescendo nel tempo e in particolar modo negli ultimi anni, ovvero da quando è emersa sempre più distintamente la consapevolezza della loro importanza come fonti insostituibili per la ricostruzione della storia e della civiltà contemporanee. Alcuni esempi saranno sufficienti a chiarire il concetto. Si pensi alla pandemia da COVID-19 ancora non del tutto terminata: le fonti web saranno fondamentali per ricostruire gli avvenimenti occorsi in questo periodo [6], dalle decisioni delle autorità sanitarie alle preoccupazioni della popolazione, fino alle tensioni susseguenti a certe scelte; senza di esse e solo a partire dalle fonti tradizionali sarà molto difficile riuscire a raccontare in maniera oggettiva e compiuta quanto successo in questi due anni. Oppure si pensi al conflitto, tutt'ora in corso, tra Russia ed Ucraina: la maggior parte delle notizie relative alla guerra vengono fornite, da ambo le parti, facendo ampio uso di interazioni sui social media (Facebook, Twitter, Telegram, Instagram, etc.), quasi come fossero uno strumento di comunicazione ufficiale. È sui social media che vengono annunciate le intenzioni delle parti e le possibili conseguenze, così come le risoluzioni, le decisioni e le scelte (cfr. Figura 1).[1] È sempre sui social media che vengono diffuse le riprese audiovisive delle azioni di guerra, sia dall'una che dall'altra parte, così come con gli stessi mezzi vengono rese note le testimonianze dei combattenti, dei superstiti agli attacchi missilistici, delle famiglie distrutte dai bombardamenti.

[1] «Ci fu un tempo in cui la guerra veniva dichiarata convocando l'ambasciatore del paese che si intendeva attaccare. Poi fu la volta delle Nazioni Unite, che autorizzavano le cosiddette "missioni di pace". Ora tutto è cambiato: la guerra si dichiara su Twitter. Il presidente americano Donald Trump ha usato il social network per avvisare la Russia e il mondo intero della decisione di bombardare la Siria». Cfr. Mauceri, A. *Siria. Donald Trump, dichiara guerra su Twitter, 'ho missili carini e intelligenti'*, 11 Aprile 2018, https://www.notiziegeopolitiche.net/siria-donald-trump-dichiara-guerra-su-twitter-ho-missili-carini-e-intelligenti.

Figura 1. Tweet del presidente ucraino Vladimir Zelensky che informa dei bombardamenti avvenuti il 26 settembre 2022

Anche nel recente passato l'utilizzo dei social media per veicolare le intenzioni dei capi politici e di Stato ha avuto ampia diffusione: perfino le dichiarazioni di guerra che nel passato sarebbero state comunicate attraverso documenti ufficiali, oggi vengono comunicate a tutto il mondo mediante un tweet o un post (cfr. Figura 2)


Donald Trump a un passo dalla dichiarazione di guerra all'Iran via Twitter

di HuffPost

Il consigliere di Rohani minizza: "Minacce per farci accettare le sanzioni". Ma il generale Hajilou promette una "risposta devastante"

20 Maggio 2019

Figura 2. La notizia del tweet con cui Donald Trump aveva annunciato la sua 'dichiarazione di guerra'

Allo stesso modo i social media vengono sempre più spesso utilizzati per diffondere comunicazioni ufficiali raggiungendo il maggior numero possibile di persone. Come non ricordare che la notizia della morte della regina Elisabetta II è stata data, prima ancora che attraverso i canali ufficiali di comunicazione del Parlamento d'Inghilterra, mediante un tweet (cfr. Figura 3)?

Figura 3. Il tweet con cui è stata comunicata la morte della regina Elisabetta II il giorno 8 settembre 2022

Questi sono solo alcuni esempi, ma sono utili a comprendere la portata del fenomeno. Il web e i social media stanno diventando fonti insostituibili per la ricostruzione della storia e la storiografia non potrà fare a meno di essi. La domanda che ci si pone quindi è la seguente: come potranno gli storici del futuro ricostruire il periodo storico che stiamo vivendo o le vicende belliche che stiamo vivendo se le istituzioni della memoria non saranno in grado di archiviare e conservare le risorse sul web e i social media? Ancora: come sarà possibile, in futuro, studiare gli aspetti della vita quotidiana attuale – come la cultura popolare, gli usi ed i costumi, l'alimentazione, la salute e il benessere o i viaggi – per ricostruire la civiltà presente prescindendo dalla conservazione dei social media, da Facebook a Twitter ad Instagram, che costituiscono una 'finestra' aperta sulla vita di tutti i giorni? È per questo motivo che negli ultimi vent'anni si è molto sviluppato il web archiving, cioè: [5]

> «il processo finalizzato alla 'cattura' e conservazione sistematica di porzioni del web a cura di istituzioni della memoria, come archivi e biblioteche nazionali, istituzioni universitarie, fondazioni.»

Diverse istituzioni della memoria si sono attivate con iniziative e progetti di *web archiving* e, recentemente, si è cominciato anche a sviluppare il *social media archiving*, considerato un versante ancora poco esplorato ma verso cui si sta dirigendo sempre più l'interesse non solo delle istituzioni ma anche delle aziende e dei cittadini.

Stato dell'arte

Le prime riflessioni sul tema del web archiving risalgono alla fine degli anni Novanta, periodo in cui il web si era ormai affermato come strumento di condivisione di contenuti ed iniziava a porsi concretamente il problema della sua archiviazione. Già nel 1996, sei anni dopo lo sviluppo del World Wide Web ad opera di Tim Berners Lee, prese avvio il progetto di Internet Archive,[2] voluto fortemente da due ingegneri statunitensi, Brewster Kahle e Bruce Gilliat, a capo di una organizzazione senza scopo di lucro con l'obiettivo di creare una digital library di siti internet e così salvaguardarne e garantirne l'accesso permanente. I fondatori ebbero la brillante idea di archiviare i siti web mediante la cattura di 'istantanee' delle pagine che costituiscono un sito web effettuate da un *crawler*.[3] Oggi Internet Archive vanta più di 25 anni di cronologia web, per un totale di oltre 625 miliardi di siti web, resi accessibili tramite la *Wayback Machine*,[4] l'interfaccia pubblica che consente di ricercare e visualizzare le versioni archiviate dei siti web. Internet Archive mette a disposizione anche Archive-It, un servizio in abbonamento disponibile fin dal 2006, che consente alle istituzioni della memoria di costruire raccolte di contenuti nativi digitali. Attraverso una applicazione web di facile utilizzo, i partner di Archive-It possono raccogliere, aggiungere metadati, gestire e generare una copia delle proprie raccolte digitali, le quali vengono archiviate ed ospitate nel data center[5] di Internet Archive e rese accessibili al pubblico con ricerca full-text.

L'importanza della conservazione dei siti web è stata riconosciuta fin dal 2003 anche dall'Organizzazione delle Nazioni Unite per l'educazione, la scienza e la cultura (UNESCO), che nel "Charter on the Preservation of Digital Heritage" ha inserito le pagine web tra i materiali digitali che costituiscono il 'digital heritage': [28]

> «The digital heritage consists of unique resources of human knowledge and expression. It embraces cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other kinds of information created digitally, or converted into digital form from existing analogue resources. Where resources are "born digital", there is no other format but the digital object. Digital materials include texts, databases, still and moving images, audio, graphics, software and web pages, among a wide and growing range of formats. They are frequently ephemeral, and require purposeful production, maintenance and management to be retained. Many of these resources have lasting value and significance, and therefore constitute a heritage that should be protected and preserved for current and future generations. This ever-

[2] Raggiungibile all'indirizzo https://archive.org.

[3] Un *crawler* (detto anche web crawler, spider o robot) è un software che acquisisce una copia di tutti gli oggetti presenti in una o più pagine web creando un indice che ne permetta, successivamente, la ricerca e la visualizzazione.

[4] Cfr. Internet Archive, https://archive.org.

[5] I data center di Internet Archive sono ubicati a San Francisco, a Redwood City e a Mountain View, in California; per garantire la sicurezza dei dati archiviati, l'intera collezione ha un mirror nei server della Bibliotheca Alexandrina ad Alessandria d'Egitto.

> growing heritage may exist in any language, in any part of the world, and in any area of human knowledge or expression.»

Nel 2003 venne fondato l'International Internet Preservation Consortium (IIPC),[6] un'organizzazione internazionale che raccoglie biblioteche, archivi, musei e altre istituzioni allo scopo di coordinare gli sforzi finalizzati alla preservazione dei contenuti Internet; svolge attività di promozione e sviluppo di strumenti, tecniche e standard comuni per la creazione di archivi web internazionali. Attualmente partecipano all'IIPC organizzazioni di oltre 35 paesi, tra cui biblioteche e archivi nazionali, universitari e regionali.

Sebbene dal 2018 non sia più attiva, merita una citazione anche l'Internet Memory Foundation. Fondata nel 2004 e conosciuta, fino al 2010, come European Archive Foundation, è stata coinvolta in diversi progetti di ricerca finanziati dalla Commissione Europea, volti a migliorare le tecnologie di web crawling, estrazione dati, text mining e conservazione degli archivi web delle istituzioni europee [32].

Oltre ad iniziative e strumenti per raccogliere e tenere traccia delle risorse sul web, all'interno della comunità internazionale è sorta anche l'esigenza di individuare uno specifico formato contenitore che consentisse di archiviare più risorse web in un unico file. Un grosso passo avanti in questa direzione è stata la pubblicazione, nel 2009, dello standard ISO 28500[7] che ha definito il formato WARC (Web ARChive) – una revisione del formato ARC [12] utilizzato inizialmente da Internet Archive per archiviare le catture del web – che oggi rappresenta il formato standard per l'archiviazione del web [2] insieme al formato WACZ (Web Archive Collection Zipped) recentemente proposto [9]. Nel 2013 venne pubblicato anche lo standard ISO/TR 14873:2013 che definiva i principi, i metodi e gli standard di qualità per le istituzioni culturali che si occupano di *web archiving* [9]. Dagli anni Novanta del secolo scorso sono state promosse in ambito internazionale numerose iniziative di archiviazione del web. PANDORA, avviato nel 1996, è stato il primo progetto di *web archiving* sviluppato da un'istituzione pubblica, i National Libraries of Australia. Molto attiva l'area del Nord Europa, con i progetti di Svezia (1996),[8] Norvegia (2001),[9] Islanda (2004),[10] Danimarca (2005).[11] Il progetto americano della Library of Congress prende avvio nel 2000, quello della Bibliotèque Nationale de France nel 2006. Nel Regno Unito si occupano di web archiving sia i National Archives che le biblioteche incaricate del deposito legale, costituendo così un modello di riferimento

[6] Cfr. International Internet Preservation Consortium (IIPC), https://netpreserve.org.

[7] La versione corrente è la ISO 28500:2017 "Information and documentation — WARC file format", https://www.iso.org/standard/68004.html.

[8] Cfr. Kulturarw3 – The web archive of the national library of Sweden, https://dhnb.eu/projects/kulturarw3-the-web-archive-of-the-national-library-of-sweden.

[9] Cfr. Nettarkivet, https://www.nb.no/samlingen/nettarkivet.

[10] Cfr. The Icelandic Web Archive, https://vefsafn.is/index.php?page=English.

[11] Cfr. Netarkivet, http://netarkivet.dk/in-english.

internazionale. L'elenco più esaustivo ed aggiornato relativo alle iniziative internazionali di archiviazione del web è quello redatto sulla base dell'indagine condotta dal team di Arquivo.pt, l'archivio web del Portogallo. I risultati, che sono stati resi disponibili su Wikipedia nella pagina "List of web archiving initiatives" [32] indicano non solo i progetti ed i relativi paesi di appartenenza, ma anche le tecnologie di *web archiving* utilizzate ed il personale dedicato, distinguendo anche tra incarichi full-time o part-time.

Nel 2013 la Digital Preservation Coalition ha pubblicato un report specifico sul tema del web archiving nella serie delle DPC Technology Watch Publications [10]. Il rapporto affronta le questioni chiave che le organizzazioni impegnate in iniziative di archiviazione del web debbono risolvere ed offre una panoramica dei principali software e strumenti attualmente disponibili.

In questo quadro di grandi sforzi a livello internazionale nel tentativo di trovare le strategie per preservare una risorsa che è di per sé estremamente effimera, l'Italia si distingue per l'enorme ritardo rispetto agli altri paesi, anche europei. L'unica iniziativa a livello nazionale meritevole di menzione è stata avviata nel 2018 dalla Biblioteca Nazionale Centrale di Firenze con il progetto di raccolta e archiviazione di siti web di 'interesse culturale' per la storia e la cultura italiana, secondo i principi della legge nazionale sul deposito legale (L. 106/2004 e il suo Regolamento attuativo D.P.R. 252/2006) [27]. Oltre alla raccolta, la Biblioteca si fa carico anche dell'organizzazione e della metadatazione 'manuale' dei siti archiviati, avvalendosi della piattaforma Archive-it di Internet Archive per l'accesso e la conservazione. Il deposito legale dei documenti diffusi tramite reti informatiche, tuttavia, non è ancora obbligatorio perché lo stesso D.P.R. 252/2006 all'art. 37 prevedeva che il deposito di tali documenti fosse subordinato alla redazione di uno specifico regolamento tecnico che non è ancora stato emanato. Pertanto, l'adesione al programma è su base volontaria da parte dei gestori dei siti, i quali possono manifestare il proprio interesse compilando il *form* online disponibile sul sito della Biblioteca. La raccolta si presenta quindi a tutt'oggi molto parziale e frammentata rispetto alla produzione di interesse culturale presente sul web italiano. La scarsa sensibilità sul tema del web archiving e l'assenza di uno specifico quadro normativo rendono difficoltosa l'attuazione di strategie nazionali condivise, laddove un'azione sinergica sarebbe necessaria vista la rapida evoluzione del web e l'enorme quantità di risorse culturali che vi trovano sede e che rischiano di andare irrimediabilmente perdute a causa della rapida evoluzione del web e la sua 'fragilità' [18].

Il tema del social media archiving è stato preso in considerazione più tardi rispetto a quello del web archiving – nonostante che uno dei primi progetti 'pionieristici' di archiviazione di social media sia stato lanciato già nel 1994 (addirittura prima della nascita dei social media)[12] – e questo, ovviamente, per il fatto che discendono dal fatto che i social media sono comparsi molto più tardi rispetto al web. Ad ogni modo, negli ultimi anni sono stati avviati diversi progetti di archiviazione dei social media. In Francia, la Biblioteca Nazionale ha archiviato i

[12] Si trattava dell'Occasio project [31], condotto dall'International Institute of Social History di Amsterdam, che si prefiggeva lo scopo di conservare le conversazioni a tema sociale e politico postate, tra il 1988 e il 2002, nei gruppi di discussione online; ad oggi l'archivio non è più disponibile online ma se ne possono recuperare alcune pagine su Internet Archive.

post istituzionali su Facebook fin dal 2006, appena due anni dopo la nascita del social media, ma i cambiamenti tecnologici della piattaforma e le conseguenti difficoltà tecniche hanno costretto, nel 2010, la Biblioteca ad abbandonarne l'archiviazione sistematica. Nel Regno Unito le biblioteche titolari del deposito legale includono i social media nelle collezioni tematiche di web archiving, mentre i National Archives svolgono il servizio di social media archiving degli account social dei principali enti, istituti e organismi governativi. In Irlanda, il Digital Repository of Ireland in collaborazione con l'Insight Centre for Data Analytics News Lab ha portato avanti uno studio di fattibilità relativo all'archiviazione dei social media prendendo in considerazione contemporaneamente le esigenze della ricerca scientifica e quelle della ricerca storica, per arrivare alla costituzione del Social Repository of Ireland.[13] In Belgio, la Biblioteca Reale ha lanciato nel 2020 il progetto BESOCIAL, allo scopo di definire una strategia nazionale di conservazione dei social media; esso vede la compartecipazione di diversi enti e istituti di ricerca che si occupano, tra le altre cose, di Intelligenza Artificiale. La rete delle biblioteche e degli archivi afferenti agli Smithsonian Institutions archiviano, utilizzando Archive-it, i loro profili social e quelli di personaggi o istituti a loro correlati.[14] Negli Stati Uniti, al termine della presidenza Obama[15] e precedentemente all'insediamento di Donald Trump nel gennaio 2018, i National Archives and Records Administration (NARA) hanno archiviato tutta l'attività sui social di Barack Obama prima che gli account istituzionali venissero trasferiti al nuovo presidente.[16] Il principale strumento utilizzato è stato ArchiveSocial,[17] con il quale sono stati archiviati oltre 250mila tra foto, post, video e messaggi scritti da più di 100 profili ufficiali in qualche modo legati agli otto anni di presidenza Obama. Il 22 aprile 2019 sul China South Morning Post è apparsa la notizia che la National Library of

[13] Cfr. il sito del Digital Repository of Ireland, https://www.dri.ie.

[14] Cfr. https://www.archive-it.org/collections/3393.

[15] Obama è diventato presidente nel 2009, quando Facebook esisteva da cinque anni, Twitter da tre anni, mentre Instagram e Snapchat ancora dovevano essere inventati. Oltre a essere stato il primo presidente davvero digitale Obama è anche stato, per definizione della Casa Bianca, il «first social media president». Prima di lui nessuno aveva avuto un account Twitter "personale" e presidenziale (cioè quello @POTUS), nessuno aveva fatto una diretta su Facebook, un video in cui rispondeva su YouTube alle domande dei cittadini, una foto con un filtro Snapchat, delle playlist Spotify in cui consigliava ottima musica. Cfr. Come si archivia una presidenza social, Il Post, 6 gennaio 2017, https://www.ilpost.it/2017/01/06/barack-obama-social-archivi.

[16] La Casa Bianca aveva illustrato sul proprio sito web la strategia che avrebbe seguito per il passaggio degli account si sarebbe dovuta basare su tre regole principali, a prescindere dal sito, dall'app o dal tipo di profilo: 1. ogni cosa deve essere salvata dalla NARA (National Archives and Records Administration); 2. ogni cosa deve restare comunque pubblica, così che la si potesse consultare sul social su cui è stata pubblicata (anche se su un altro profilo); 3. ogni account deve passare alla nuova amministrazione (che deciderà poi se, come e quanto farne uso).

[17] ArchiveSocial è uno strumento software per archiviare e rendere disponibili contenuti di vario genere pubblicati su social media. Cfr. https://archivesocial.com. L'archivio della presidenza di Obama è raggiungibile all'indirizzo https://archivesocial.com/whitehouse.

China avrebbe conservato oltre 200 miliardi di post pubblici di Weibo, il più popolare sito di microblogging cinese, come parte di una iniziativa complessiva per conservare il patrimonio culturale digitale cinese.

La fragilità del web e dei social media

La necessità di preservare il web ha cominciato ad essere evidente fin dalla fine del secolo scorso. Sebbene per diversi anni il web sia stato considerato, con un ottimismo evidentemente eccessivo, un 'self preserving medium', ovvero capace di auto-conservarsi, con il trascorrere del tempo la realtà si è mostrata molto diversa. La rete è, infatti: [5]

> «effimera ed estremamente variabile: la vita media di un url è molto breve, la persistenza dell'informazione bassa e la natura complessa delle pagine, che contengono formati diversi e link, rende altrettanto complessa la loro conservazione.»

A questo proposito, ha suscitato una certa apprensione la pubblicazione dell'articolo di Mikael Laakso della Hanken School of Economics di Helsinki [15] dal quale è emerso che tra il 2000 e il 2019 oltre cento riviste scientifiche open-access sono svanite, rendendo le loro pubblicazioni sostanzialmente introvabili, e il fenomeno sembra possa essere molto più esteso e amplificarsi nel tempo. Il tema comincia ad essere sempre più evidente e allarmante: a differenza di quanto si pensava un tempo la memoria digitale affidata al web si è rivelata molto meno solida di quanto era logico aspettarsi, molto più fragile di quella affidata ad un supporto fisico – come un libro di carta o un documento cartaceo – assai più dipendente da variabili sulle quali non abbiamo grande controllo, come la questione dell'obsolescenza dei formati elettronici e dell'hardware utilizzato, l'architettura stessa della rete, basata su tecniche e protocolli che nel tempo diventano essi stessi obsoleti, la disponibilità di finanziamenti o anche semplicemente la riduzione della fornitura di energia elettrica, essenziale per la 'sopravvivenza' dei data center che ospitano i siti web e le piattaforme dei social media.

A questo proposito è illuminante rileggere un passo dell'articolo "Internet come fonte" che risale al 2007 e quindi fa capire come già all'epoca il problema fosse ben presente: [4]

> «Una serie di strumenti con i quali fino a ieri venivano scritti i documenti non esistono più. Solo dieci anni fa i browser più diffusi erano Netscape e Mosaic. Explorer era poco usato e Firefox non esisteva, le pagine html non erano ancora in grado di interagire con le applicazioni di scrittura e di calcolo, né, tanto meno, di mostrare immagini in movimento o far sentire suoni. Se superiamo la prima decade di Internet e andiamo ancora più indietro, i sistemi di videoscrittura usati nei primi anni Ottanta sono oggi forse reperibili, ma con grande difficoltà.»

Sono trascorsi 'solo' quindici anni dalla pubblicazione di quell'articolo e dobbiamo riconoscere che la situazione è completamente cambiata, a dimostrazione di quanto sia repentina l'evoluzione tecnologica e la conseguente obsolescenza: quei browser non esistono più, sostituiti da altri molto più performanti; le pagine web oggi sono molto più complesse ed articolate, e, soprattutto, hanno fatto irruzione in questo panorama i social media.

Non sono solo le pubblicazioni scientifiche a scomparire ma anche i siti web realizzati in occasione di progetti, anche internazionali, che a distanza, a volte, di pochi mesi o pochi anni dalla conclusione del progetto non sono più accessibili. A questo proposito si potrebbero portare numerosissimi esempi: si pensi al progetto di ricerca internazionale GAU:DI (Governance, Architecture and Urbanism: a Democratic Interaction), che aveva ricevuto dalla Comunità Europea un finanziamento triennale di 1,9 milioni di euro per il triennio 2002-2004 all'interno del programma Culture 2000, coinvolgendo diverse istituzioni europee con l'obiettivo di promuovere la collaborazione fra esse e favorire la conoscenza dell'architettura contemporanea: a distanza di pochi anni dalla conclusione del progetto il sito del progetto non risultava più raggiungibile, e con esso tutti i preziosi materiali ivi contenuti.[18] Oppure, si pensi in Italia, al sito della Fondazione Rinascimento Digitale che nel primo decennio del XXI secolo era stato un punto di riferimento sui temi legati alla conservazione del digitale ed il cui sito web,[19] anch'esso ricchissimo di materiali, è improvvisamente scomparso per un cambio di obiettivi da parte della Fondazione stessa. O, ancora, il sito del Centro di eccellenza italiano sulla conservazione digitale in Italia, che aveva anch'esso reso disponibili online tutta una serie di materiali estremamente utili e il cui sito web[20] risulta ormai perduto, sembra per un banale problema di rinnovo di dominio.

Analogamente sono scomparsi archivi digitali di prestigiosi quotidiani perché la tecnologia utilizzata a suo tempo è ormai diventata inutilizzabile;[21] sono scomparsi gli archivi musicali di MySpace perché a un certo punto della parabola discendente di uno dei primi social network qualcuno ha deciso che non era più economicamente conveniente dedicare risorse finanziarie alla manutenzione di quelle pagine che erano sempre meno consultate; è scomparso l'intero spaccato socioculturale contenuto nei sedici anni di (stravaganti) domande e risposte conservate su Yahoo Answer, anch'esso chiuso perché non più redditizio. Sono scomparsi 38 milioni di pagine utenti costruite su GeoCities, il servizio di web hosting fondato nel novembre 1994 ed acquisito da Yahoo! nel 1999, quando era il terzo sito web più visitato del World Wide Web;

[18] Il sito web del progetto era raggiungibile all'indirizzo www.gaudi-programme.eu, ma dopo pochi anni dalla sua conclusione il dominio era stato acquisito da un altro proprietario e non era più disponibile, portando nell'oblio tutti i preziosi materiali in esso contenuti, come le *Guidelines to managing architectural records.*

[19] Fino alla fine del 2012 il sito era raggiungibile all'indirizzo rinascimento-digitale.it; recentemente, dopo un 'buco' di quasi dieci anni, il dominio è stato riacquisito e il sito è stato parzialmente ricostruito, ma la maggior parte dei documenti che popolavano il 'vecchio' sito non è più disponibile.

[20] Il sito web era raggiungibile all'indirizzo conservazionedigitale.org. Attualmente il dominio risulta utilizzato da una fantomatica società cinese.

[21] Alla fine del 2020 aveva fatto scalpore, per la risonanza mediatica che ne era seguita, la notizia che l'archivio storico del quotidiano "La Stampa" sarebbe sparito dal web perché il sito era stato costruito in Flash, una tecnologia utilizzata per l'interfaccia divenuta ormai obsoleta e non più fruibile dai moderni browser. Si trattava di cinque milioni di articoli che nel corso di 150 anni avevano raccontato la storia di Torino, del Piemonte e dell'Italia. L'Archivio storico della Stampa è tornato nuovamente online il 15 febbraio, dopo due mesi di blocco.

nell'aprile 2009, circa dieci anni dopo che Yahoo! l'aveva acquisito, la società ne ha annunciato la chiusura [24].

Così scompaiono improvvisamente intere comunità online di persone perché la piattaforma in cui risiedevano ha cambiato proprietario; vengono cancellate le tracce del nostro archivio fotografico familiare perché lo spazio virtuale nel quale avevamo deciso di custodirle improvvisamente chiude; scompaiono ugualmente anche contenuti che la comunità stimerebbe probabilmente meno trascurabili delle foto dei nostri viaggi o delle chiacchiere virtuali con i nostri amici. Potrebbe sembrare, a prima vista, che le pagine andate perdute siano una trascurabile minoranza, ma non è così: uno studio condotto nel 2014 [33] mostrava come già allora il 75 per cento dei link contenuti sul sito della Harvard Law Review non fosse più funzionante. Si tratta, quindi, di un fenomeno che colpisce tutte, indistintamente, le risorse web, siano esse costituite da siti web più o meno complessi che piattaforme di social media più o meno diffuse.

Infine, c'è da considerare l'aspetto più generale della conservazione digitale a lungo termine. Non è sufficiente riuscire ad archiviare i siti web e le interazioni presenti sui social media ma occorre anche preoccuparsi della loro conservazione nel tempo, e questo significa mettere in campo tutta una serie di strategie volte ad assicurare che il materiale 'catturato' rimanga ancora fruibile a distanza di anni o decenni a partire da oggi.

L'importanza della conservazione del web e dei social media

Il dibattito sull'importanza della conservazione del web e dei social media è tuttora in corso, ma la consapevolezza che essi rappresentino una nuova e fondamentale fonte per la ricostruzione del periodo storico che stiamo vivendo si sta facendo strada. Già nel 2009 Stefano Vitali [30] indicava come prioritaria la necessità di curare la conservazione dei siti web con particolare riferimento a quelli di tipo istituzionale, ma non solo. In particolare, vista la loro popolarità ed il loro ampio utilizzo da ogni fascia della popolazione, i contenuti pubblicati sui social media rappresentano uno spaccato della società del nostro tempo ed è certo che costituiranno una fonte primaria per la comprensione e la ricostruzione della civiltà dei primi decenni del XXI secolo, non solo da parte delle future generazioni di storici, ma anche di sociologi, politologi, antropologi, umanisti, etc. [1]. Gli storici sono certamente i primi ad essere interessati alla possibilità di archiviare web e social media ed anche alla modalità di utilizzo di queste nuove fonti per la storia. A questo proposito Manetta [20] si domanda:

> «Come viene conservata l'eredità culturale e quali meccanismi sono adottati per lasciare ai posteri una traccia delle risorse native digitali e digitalizzate? Se si prende un preciso evento storico, come la recente campagna elettorale negli Stati Uniti d'America, come sarà possibile, per gli storici che tra cento anni lo vorranno, ricostruire il successo di Donald Trump usando come fonte primaria i suoi tweet?»

I social media sono spesso utilizzati come piattaforma per discutere di eventi importanti come le crisi politiche e le elezioni. Sono utilizzati anche da numerosi politici e altre figure pubbliche.

Ad esempio, gli storici del futuro potrebbero voler ricostruire la campagna elettorale del presidente americano Biden e la contrapposizione con Trump, e questo non sarebbe possibile senza avere a disposizione i rispettivi tweet pubblicati su Twitter. Oppure ricostruire la vicenda che ha portato recentemente alla caduta del governo in Italia e all'indizione di nuove elezioni, con la successiva campagna elettorale da parte delle diverse forze politiche coinvolte, la quale si è svolta in massima parte sui social media.[22] I social media sono utilizzati anche come spazio di discussione o di manifestazione di pensieri, opinioni, intenzioni relative a guerre, disastri naturali, eventi calamitosi in genere. Ad esempio, come già anticipato nell'Introduzione, l'attuale conflitto tra la Russia e l'Ucraina si sta combattendo non solo sul campo e a livello diplomatico, ma anche – e in misura certamente non meno rilevante – sui social media e sui mezzi di comunicazione, sia da una parte che dall'altra. Ad ogni azione di guerra corrisponde prontamente una reazione che si manifesta mediante un tweet o un post; ad ogni minaccia di un possibile attacco nucleare postata su Telegram dal presidente russo Putin corrisponde una reazione postata su Twitter dal presidente ucraino Zelensky o dal presidente statunitense Biden. Archiviare e conservare queste interazioni – che costituiscono delle fonti primarie – sarà assolutamente fondamentale per riuscire, un domani, a ricostruire correttamente la vicenda del conflitto e non sarà possibile farlo a partire, ad esempio, solo da quanto riportato dai giornali, che pure riprenderanno quei post e quei tweet, ma certamente con un filtro legato alla visione politica e alla linea editoriale del giornale stesso. In sostanza i social media e le altre forme di comunicazione online saranno utilizzate come fonti primarie dai futuri storici per comprendere i nostri tempi, come affermato dalla ricercatrice Katrin Weller, intervistata da Jason Steinhauer, in un articolo pubblicato online nel 2015: [26]

> «Social media data and other online communication data will surely be used by future historians to learn about our times. They won't be the only source material, as current traditional sources will still remain. But social media are already being used as a new type of data source by contemporary scholars in various disciplines: political science, sociology, linguistics, communication science, geography, physics, computer science and many more. It is logical to assume that future historians will also look at these sources.»

Certamente non si tratterà delle uniche fonti sulle quali gli storici e gli studiosi in genere si baseranno perché le tradizionali fonti rimarranno. Inoltre ci sarà da affrontare il problema della verifica dell'attendibilità delle fonti, dal momento che il fenomeno delle fake news è in continua espansione e, ad oggi, non sembra siano stati individuati dei 'rimedi' per arginarlo né per discriminare con esattezza tra notizie vere e notizie false; ma questo fenomeno riguarda non solo le comunicazioni sui social media ma anche quelle su fonti più tradizionali, come i quotidiani sia a stampa che online. Come ha ricordato Giovanni De Luna nella sua relazione al Convegno "Quale futuro senza la storia", del 5 ottobre 2019: [8]

[22] La presenza sui social media è talmente importante che ormai tutti i vari personaggi politici hanno alle loro dipendenze un team di social media manager che si occupano di tutte le questioni legate alla comunicazione e alla promozione del personaggio attraverso i vari canali, dai profili social ai siti web personali.

> «Si tratta quindi di lavorare anzitutto a una certificazione delle fonti online, suggerendo come criterio per valutarne la validità quello dell'esame dei contesti in cui affiora l'informazione in esse racchiusa. È evidente, infatti, come un documento reperito sul sito dei National Archives abbia un livello di attendibilità diverso da quello del materiale messo in rete da un amatore o da un sito – come quelli negazionisti – ideologicamente schierato. La riflessione sulla distinzione tra il "vero" e il "falso", quella sull'intenzionalità della fonte, sulla congruenza necessaria tra la stessa fonte e l'oggetto della ricerca, tutti i capisaldi, insomma, della concezione dinamica delle fonti ci tornano straordinariamente utili per affrontate il nodo del rapporto tra la storia e il web con la consapevolezza critica necessaria. Così come le narrazioni proposte dalla rete vanno valutate sulla base di criteri che appartengono integralmente alla nostra disciplina: l'ipotesi storiografica che ne è alla base, la congruenza delle fonti con tale ipotesi, la solidità e la coerenza dell'argomentazione e della interpretazione.»

Purtroppo l'archiviazione e la conservazione di social media e siti web si scontra con tutta una serie di difficoltà piuttosto complesse che devono essere affrontate con competenze specialistiche e che cercheremo di delineare nel seguito.

L'archiviabilità dei siti web

Prima di analizzare gli strumenti oggi disponibili per l'archiviazione di siti web è opportuno precisare che una buona riuscita della cattura non dipende solo dallo strumento utilizzato ma anche, ed in misura rilevante, da come è realizzato il sito web: vi sono siti facilmente catturabili ed altri la cui cattura è particolarmente complessa e non sempre possibile, almeno nella sua interezza. È per questo che durante la creazione di un sito web sarebbe davvero auspicabile che gli sviluppatori e i designer tenessero in considerazione, oltre ai criteri di accessibilità, di performance, di ottimizzazione ai fini dell'indicizzazione da parte dei motori di ricerca (Search Engine Optimization, SEO), di compatibilità con gli standard del W3C e di usabilità, anche quelli di archiviabilità dei siti web. Per archiviabilità si intende: [23]

> «l'insieme delle caratteristiche che i contenuti, la struttura, le funzionalità e le interfacce di un sito web devono possedere perché il sito stesso possa essere conservato e reso accessibile nel lungo periodo con gli attuali strumenti di web archiving.»

La redazione delle linee guida vede molto attivi gli istituti di area nordamericana e anglosassone, come la Library of Congress, che ha pubblicato una linea guida per la conservabilità dei siti web[24] e le Stanford Libraries, che hanno pubblicato una guida online sulla creazione di siti archiviabili.[25] In Europa va segnalato il contributo della Commissione Europea

[23] Cfr. la pagina della Biblioteca Nazionale Centrale di Firenze dedicata al tema dell'archiviabilità, https://www.bncf.firenze.sbn.it/biblioteca/archiviabilita-dei-siti-web.

[24] Cfr. la pagina "Creating Preservable Websites" sul sito della Library of Congress: https://www.loc.gov/programs/web-archiving/for-site-owners/creating-preservable-websites.

[25] Cfr. La pagina "Archiviability" sul sito delle Stanford Libraries: http://library.stanford.edu/projects/web-archiving/archivability.

che ha pubblicato le "Guidelines to make archivable websites"[26] e quello dei National Archives britannici.[27] In Italia il tema dell'archiviabilità dei siti web è stato affrontato dalla Biblioteca Nazionale Centrale di Firenze che ha pubblicato le "Linee guida per la realizzazione di siti web archiviabili",[28] che possono essere riassunte nelle seguenti nove regole:

1) strutturare il sito in conformità con i principali standard di accessibilità;
2) mantenere URLs stabili per contenuti importanti e reindirizzarli a nuovi URLs solo quando necessario;
3) dotare il sito di Protocollo Sitemap formato XML e/o RSS;
4) associare un link HTML/XHTML ad ogni contenuto del sito (pagine, immagini, video, documenti);
5) omettere l'esclusione robots.txt o limitarla alle aree non necessarie per l'archiviazione;
6) evitare l'utilizzo di formati proprietari per i contenuti importanti, specialmente nella homepage;
7) limitare l'utilizzo di contenuti inclusi in siti di terze parti;
8) utilizzare indirizzi web univoci che contengano informazioni sullo stato dei contenuti;
9) segnalare il tipo di supporto e di codifica dei caratteri.

Per valutare l'archiviabilità dei siti web sono stati elaborati alcuni metodi, come il metodo CLEAR+ (Credible Live Evaluation of Archive Readiness) [29]; inoltre è disponibile lo strumento online ArchiveReady[29] (cfr. Figura 4) nel quale è sufficiente inserire l'URL del sito da sottoporre a valutazione per ottenere come risposta un giudizio sulla sua archiviabilità.

[26] Cfr. la pagina "Guidelines to make archivable websites", https://op.europa.eu/en/web/web-tools/guidelines-to-make-archivable-websites.

[27] Cfr. la pagina "Archive a website or social media channel in the UK Government Web Archive" sul sito dei National Archives, https://www.nationalarchives.gov.uk/webarchive/archive-a-website. Si veda anche: The National Archives, Web Archiving Guidance, https://cdn.nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf.

[28] Cfr. la pagina della Biblioteca Nazionale Centrale di Firenze dedicata al tema dell'archiviabilità, *cit.*

[29] La piattaforma è disponibile all'indirizzo http://archiveready.com.

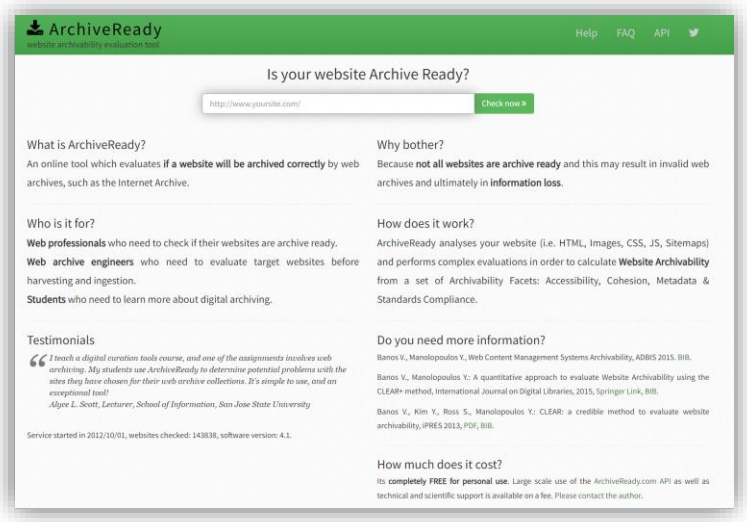

Figura 4. Lo strumento online per la verifica dell'archiviabilità dei siti web

L'archiviazione dei siti web

I siti web possono essere archiviati utilizzando strumenti software che ricadono in due grandi categorie [16]: da una parte quelli per un uso *personale*, con i quali è possibile archiviare piccoli siti oppure un numero limitato di pagine web; dall'altra, quelli per un uso di tipo *professionale/istituzionale* che consentono una maggiore flessibilità e completezza nella gestione dei processi di *harvesting*, ma richiedono competenze ed attrezzature informatiche decisamente più articolate. Questa distinzione non è mutualmente esclusiva in quanto esistono soluzioni in grado di coprire agevolmente entrambi i casi. È anche possibile distinguere gli approcci principali all'archiviazione del web sulla base della possibilità o meno di avere accesso al server sul quale è pubblicato il sito web. In tal caso si hanno generalmente tre possibilità [23]. La prima è l'*archiviazione lato server*: viene generata una versione funzionante del contenuto archiviato copiando i file contenuti nel server; questo approccio richiede la collaborazione attiva tra chi archivia e chi amministra il server su cui è ospitato il sito web. La seconda possibilità è l'*archiviazione lato client*: è l'approccio più utilizzato e si basa su programmi di web crawler che agiscono come dei client, cioè simulano il comportamento dei browser utilizzati dalle persone per la navigazione e utilizzano il protocollo HTTP per raccogliere le risposte con il contenuto fornito direttamente dal server del sito che si vuole archiviare; solitamente il crawler comincia da un URL di partenza e segue via via tutti i link fino ad una profondità prestabilita catturando le copie di tutti gli oggetti digitali disponibili. Esiste poi una terza possibilità, l'*archiviazione transazionale* ovvero un metodo che si concentra sugli eventi registrando il contenuto delle transazioni che si svolgono tra un server web ed i browser web quando l'utente naviga il sito;

questo tipo di archiviazione viene solitamente utilizzato quando si vuole raccogliere il contenuto di un sito insieme alla prova del funzionamento delle sue pagine; purtroppo le pagine che non vengono visitate non potranno essere archiviate.

Senza avere la pretesa di analizzare tutte le soluzioni presenti sul mercato – che sono numerose e in costante aumento – presentiamo di seguito alcuni dei più diffusi strumenti e servizi per il web archiving sia di tipo commerciale che open source.

Uno tra gli strumenti più conosciuti è certamente HTTrack, un'applicazione *open source* per il mirroring di siti web e la loro navigazione offline.[30] Ne esiste una versione con istruzioni testuali dalla riga di comando dei principali sistemi operativi come Linux, Windows e Mac, ed una versione dotata di interfaccia grafica predisposta per Windows (WinHTTrack) e per Linux (WebHTTrack) che ne rende sicuramente più semplice l'utilizzo. Il programma permette di scaricare dal server web al proprio computer locale un intero sito web ricostruendone l'intera struttura: vengono memorizzati il codice, le immagini ed ogni altro tipo di file. In questo modo è possibile navigare il sito esattamente come se si fosse online. Il software permette inoltre di configurare numerose opzioni per limitare o estendere la raccolta di base e per controllare il tipo e le caratteristiche dei file da scaricare sul proprio computer.[31]

Il secondo software che vogliamo segnalare è WebCopy,[32] uno strumento gratuito ma non *open source* sviluppato dall'azienda Cyotek e che consente di scaricare automaticamente il contenuto di un sito web sul proprio dispositivo locale. Come HTTrack, anche WebCopy esegue la scansione del sito web specificato e ne scarica il contenuto: i collegamenti a risorse come fogli di stile, immagini e altre pagine del sito verranno automaticamente 'rimappati' in modo da corrispondere al proprio percorso locale. Utilizzando il pannello di configurazione si potranno definire quali parti di un sito web verranno copiate, permettendo, ad esempio, di scaricare solo le immagini anziché l'intero contenuto. L'interfaccia d'uso è molto completa e sicuramente più facile da utilizzare rispetto HTTrack. Tra le opzioni principali è presente il comando che consente di effettuare una scansione completa del sito prima di iniziare a scaricarlo e creare così una mappa del sito stesso, utile per individuarne la struttura al fine di selezionare zone di particolare interesse. WebCopy riesce anche a scaricare il contenuto di aree protette da password avendo cura di inserire le credenziali richieste in fase di avvio della raccolta. Come HTTrack, anche WebCopy non consente di salvare il lavoro di copia nel formato WARC.

[30] Cfr. il sito web del produttore: https://www.httrack.com.

[31] Per ora non è prevista la possibilità di salvare il sito in formato WARC anche se in rete esiste un tool di conversione, httrack2warc, dai risultati ancora non del tutto affidabili.

[32] Il software è disponibile sul sito del produttore ma esiste solo la versione per il sistema operativo Windows (dalla versione 7 in poi). Sullo stesso sito è inoltre presente una corposa documentazione di supporto sia per l'installazione che per il suo utilizzo con diversi esempi pratici. L'ultima versione stabile è stata rilasciata alla fine di marzo 2021 ma la presenza di altre versioni in fase di test fa pensare che la casa produttrice sia intenzionata a seguire e migliorare il suo prodotto anche in futuro. Cfr. il sito web del produttore: https://www.cyotek.com/cyotek-webcopy.

Molto simile a Webcopy è WebCopier, uno strumento sviluppato dall'azienda MaximumSoft Corp., che consente di scaricare file, nonché interi siti Web o singole parti su Internet per poterli visualizzare offline. È disponibile per i sistemi operativi Microsoft Windows, Mac OS X e Linux

I software appena visti rientrano nella categoria dell'*archiviazione lato client* e dimostrano come le pratiche di web archiving possano essere implementate anche nell'ambito di progetti d'archiviazione di portata limitata. Infatti, sono efficaci pur rimanendo semplici da utilizzare, e, quindi, costituiscono un buon punto di partenza per un progetto di salvaguardia dei siti web, anche di tipo personale, e possono essere utilizzati anche da coloro che non hanno particolari competenze informatiche.

Tra gli strumenti che rientrano nella categoria dell'*archiviazione transazionale*, va citato Webrecorder, un servizio gratuito che permette di registrare e conservare le pagine visitate durante la navigazione di un sito web. È molto utile per conservare anche siti interattivi e contestuali, compresi i social media e altri contenuti dinamici, come ad esempio un video contenuto nella pagina e javascript, che non sarebbero facilmente catturabili con gli strumenti visti in precedenza. A differenza di altri sistemi, Webrecorder archivia il contenuto web attraverso la navigazione interattiva, catturando l'esatta sequenza di navigazione attraverso una serie di pagine web o oggetti digitali e preservando il percorso del singolo utente nella specifica interazione.

Lo strumento utilizza lo stesso software sia per acquisire che per riprodurre il sito.[33] Il formato di archiviazione è WARC e sono disponibili campi personalizzabili per ulteriori metadati in formato JSON.

Per progetti di più ampio respiro, come quelli condotti dalle biblioteche e dagli archivi nazionali di diversi Paesi, soprattutto del mondo anglo-sassone, questi strumenti non sono più sufficienti ed è necessario mettere in campo risorse – non solo tecniche ed economiche, ma anche umane – di dimensioni molto più ampie. In tali casi occorre appoggiarsi a strumenti professionali, come il già citato Archive-It,[34] (cfr. Figura 5) il servizio di archiviazione web commerciale offerto da Internet Archive. Si tratta di uno strumento di acquisizione che utilizza il web crawler open source Heritrix.[35] Il sistema consente di conservare i siti web archiviati e di

[33] Questo approccio è denominato archiviazione web simmetrica, e si contrappone all'archiviazione web asimmetrica dove sono utilizzati due distinti strumenti software, uno per catturare il sito e l'altro per riprodurlo dopo la cattura.

[34] Il servizio è raggiungibile all'indirizzo https://archive-it.org.

Il servizio permette di estendere le funzionalità di gestione con strumenti di analisi che partono da dati in formato WARC e Web Archive Transformation. Partendo da questi dati è possibile fare sia analisi di tipo big data sulle milioni di relazioni conservate nei siti web con report di tipo Longitudinal Graph Analysis sia estrazione di named entities dal testo dei siti web archiviati.

[35] Heritrix è il più diffuso web crawler di tipo professionale (enterprise) ed è stato sviluppato da Internet Archive già da molti anni. È distribuito con una licenza open source ed è scritto in

scaricare i contenuti memorizzati nel repository digitale di Internet Archive in formato WARC. Tuttavia non permette di eseguire la visualizzazione dei siti web archiviati, operazione per la quale è necessario un apposito strumento, la Wayback Machine, attraverso il quale è possibile consultare e navigare i propri siti web archiviati. È possibile anche utilizzare metadati specifici sia per la ricerca che per la gestione delle collezioni.

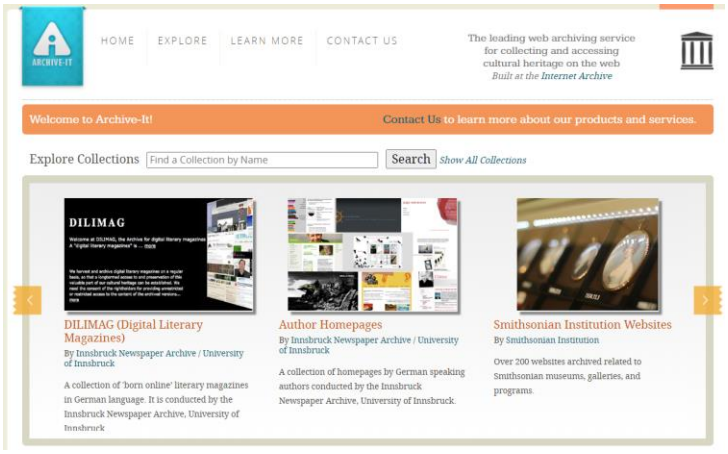

Figura 5. La home page del servizio Archive-it

Questa breve rassegna è necessariamente incompleta, dal momento che gli strumenti oggi disponibili sono numerosi e, come si diceva all'inizio, in continua crescita. Per una lista molto esaustiva degli strumenti disponibili si rimanda alla pagina specifica sul sito dell'International Internet Preservation Consortium.[36]

L'archiviazione dei social media

L'archiviazione dei social media rappresenta una vera e propria sfida per le istituzioni della memoria, che si trovano a dover affrontare tutta una serie di criticità per poter riuscire nell'intento. Una delle principali è senza dubbio costituita dall'immensa quantità di materiale

linguaggio Java. L'interfaccia principale è accessibile tramite un browser web ma è possibile automatizzare i processi di web crawling periodici con appositi comandi. Il risultato dei processi di crawling viene memorizzato in file WARC ai quali il programma aggiunge propri metadati specifici. Il tool richiede una buona conoscenza sistemistica per l'installazione e soprattutto per l'utilizzo in contesti complessi. È dotato di numerosi file di log che permettono di avere una fotografia precisa del processo e sono utili anche per la verifica di eventuali oggetti web mancanti e per poter ripetere scansioni in caso di errore.

[36] Si veda la pagina "Tools & software," https://netpreserve.org/web-archiving/tools-and-software.

che è presente nelle piattaforme dei social media e che giorno dopo giorno aumenta sempre più. I social media sono "luoghi" estremamente attivi con un numero di utenti che cresce a ritmi esponenziali, fino a coinvolgere quasi l'intera popolazione mondiale. Pertanto, la dimensione dei dati da archiviare è enorme. È evidente che l'archiviazione indistinta di tutto quanto viene prodotto sulle piattaforme social non è sostenibile. Occorre individuare criteri utili a discriminare che cosa sia rilevante conservare e cosa, invece, possa essere trascurato e sulla base di questi selezionare i contenuti che si desidera conservare e scartare quanto privo di interesse.[37]

Oltre alla criticità appena vista, che costituisce una sfida anche per il web archiving, se ne aggiungono di nuove, specifiche delle piattaforme social. Una forte criticità è costituita dal fatto che la possibilità di condividere testi, immagini, audio, video e altri oggetti digitali – caratteristica fondamentale di tutti i social media – fa si che i contenuti presenti siano estremamente variegati, complessi, interattivi e non standardizzati. Ad esempio, in un *post* su Facebook è possibile condividere un'immagine tratta da Flickr, un video caricato su YouTube o un documento in formato PDF presente su un sito web. Una soluzione di archiviazione deve essere in grado di includere il *post* originale, ma deve anche essere capace di seguire i collegamenti ipertestuali (*link*) e "catturare" l'immagine su Flickr, il video su YouTube o il documento PDF su un sito web.[38] A rendere le cose ancora più complesse è il fatto che i social media sono composti da oggetti estremamente "linkati" tra di loro; se ad esempio, in Twitter si clicca su un *tweet*, un *hashtag* o uno *username* si avvia uno *script* che riorganizza i contenuti fornendone una vista completamente nuova e comunque degna di essere "catturata" dal sistema di archiviazione. Riuscire a "catturare" e conservare questo ed altri tipi di interazione tra l'utente e la piattaforma non è certamente semplice; la complessità, è evidente, non è più quella di una semplice pagina web[39] ma vi sono sfide aggiuntive che rendono l'archiviazione e la conservazione dei contenuti dei social media un compito molto più impegnativo.

Vi sono, poi, le questioni da affrontare sotto il profilo tecnologico (ad esempio: quali sono gli strumenti e le tecniche da utilizzare per l'archiviazione? Quanto frequentemente va fatta la campagna di raccolta?) ma anche e soprattutto sotto il profilo legale e contrattuale (ad esempio: come capire se il contenuto è di carattere pubblico o privato? Quali sono i termini d'uso[40] e il

[37] A questo proposito occorre osservare che l'archiviazione indiscriminata dei *social media* porta ad una forte duplicazione dei contenuti, dovuta al fatto che lo stesso oggetto digitale può essere condiviso da più utenti. Si pensi, ad esempio, ad una foto condivisa su Facebook da più persone: si tratta della stessa risorsa presente nel "diario" di tutti gli utenti che l'hanno condivisa. Archiviare più volte uno stesso oggetto digitale è inutile e dispendioso.

[38] In caso contrario, si avrebbe una situazione analoga a quella che si verifica se si volesse archiviare una e-mail senza salvare i suoi allegati.

[39] Le criticità connesse all'archiviazione del web sono documentate esaurientemente dalla norma ISO/TR 14873:2013 - *Information and documentation -- Statistics and quality issues for web archiving.*

[40] Di solito i dati provengono da piattaforme di social media, gestite da grandi aziende – come Facebook o Twitter – che hanno ciascuna i propri termini di servizio. Molte delle interazioni sui

trattamento dei dati sottoscritti dagli utenti al momento dell'iscrizione ad una piattaforma social? Quali sono i termini d'uso consentiti dalle piattaforme alle terzi parti che vogliano raccogliere i dati? Come stabilire di chi è la proprietà intellettuale dei contenuti pubblicati sui social media e magari ripresi più volte dalle varie piattaforme? Come conciliare l'operazione di archiviazione con le norme sulla privacy e il diritto all'oblio?). È evidente che le strategie di conservazione devono avvenire all'interno di un quadro che garantisca la tutela degli utenti dei social media – che sono coloro che hanno effettivamente creato i contenuti all'interno di una piattaforma social – della loro privacy e dei loro interessi.

Va poi preso in considerazione il fatto che i social media sono piattaforme di "conversazione" nelle quali interagiscono tipicamente più soggetti: quali sono i contenuti che debbono essere archiviati dalle istituzioni della memoria? Consideriamo, ad esempio, il profilo pubblico di una organizzazione (un ente, una istituzione, etc.); solitamente ai post da questa pubblicati fanno seguito i commenti, le risposte, le 'reazioni', le condivisioni dei cittadini e di tutti coloro che seguono quel profilo. In casi del genere, cos'è che è importante archiviare, anche per finalità di una futura ricerca? Solamente il post dell'organizzazione oppure anche i commenti, le risposte, le condivisioni, etc.? E come trattare i contenuti condivisi (es: le immagini, i contenuti sonori ed audiovisivi), sia quelli incorporati che quelli 'linkati'?

Infine, occorre comprendere come riuscire a preservare, oltre al contenuto singolo, anche il contesto in cui avviene quella conversazione e l'esperienza di fruizione di ciascuna piattaforma social. Gran parte delle informazioni di contesto si perdono rapidamente, ma risulteranno molto importanti quando nel futuro si tratterà di interpretare i post o i tweet. Purtroppo, l'interfaccia di una piattaforma social cambia continuamente nel tempo a seguito dei vari aggiornamenti, a volte anche in maniera radicale, ed è molto difficile risalire all'aspetto che aveva una specifica piattaforma anche solo un paio di anni fa, perché di questo genere di informazione non rimane traccia. Ma il *look and feel* influenza fortemente il modo in cui le persone utilizzano i dati dei social media e interagiscono tra loro.

Infine, occorre domandarsi quale debba essere il soggetto che deve assumersi il compito di archiviare e conservare nel tempo i social media. Le aziende che hanno creato le piattaforme, come Twitter e Facebook? Le istituzioni archivistiche o le biblioteche a livello nazionale? Quelle a livello regionale e locale? Internet Archive o altre organizzazioni senza scopo di lucro? Gli studiosi interessati ad una determinata tematica? Le singole persone, ciascuno per i contenuti che lo riguardano? Come si può facilmente intuire, si tratta di domande alle quali non è facile fornire una risposta e che richiedono competenze estremamente diversificate, non solo tecnologiche o informatiche, ma almeno anche archivistiche, biblioteconomiche, giuridiche. Inoltre, le differenze tra le varie piattaforme social sono tali che ogni caso è diverso dall'altro e richiede una valutazione specifica, rendendo di fatto molto difficile fornire delle regole di carattere generale.

social media non vengono rese disponibili nella loro forma completa e l'accesso dipende spesso da accordi con le rispettive aziende, che possono avere o meno interesse a condividere l'accesso ai loro dati o a discutere le strategie di archiviazione.

Appare, quindi, evidente che non può esistere una risposta univoca, ma quello che è certo è che tale compito non può essere demandato ad un'unica organizzazione – come Internet Archive – che si occupi di conservare da sola il patrimonio mondiale dei social media e che certamente non può essere in grado di riuscire in un'impresa del genere. [41]

Per quanto riguarda i metodi di archiviazione dei social media, questi dipendono in larga misura dal modo in cui un'organizzazione li utilizza e da quali sono le piattaforme impiegate. Inoltre, l'effettiva attuazione di un piano efficace per la loro archiviazione richiede il supporto di tutto il personale coinvolto nell'uso dei social media dell'organizzazione. Per gli account di cui un'organizzazione è proprietaria (e titolare delle informazioni di accesso), può essere impiegata la funzione di download dell'archivio fornita da alcune piattaforme. Ad esempio, Facebook e Twitter, consentono ai proprietari degli account di scaricare i propri contenuti pubblicati sul social media. Tuttavia, per raccolte più ampie, il metodo di acquisizione migliore e più efficace è quello che si basa sull'utilizzo delle API (Application Programming Interface)[42] messe a disposizione dai social media stessi. Esistono, poi, servizi di archiviazione forniti da società commerciali che si 'appoggiano' sulle API messe a disposizione dalle varie piattaforme *social*. Tra le tante (*SocialSafe*, *PageFreezer*, *Edora*, etc.) merita di essere segnalata *ArchiveSocial*[43] i cui servizi sono probabilmente i più utilizzati anche dalle istituzioni della memoria.

I servizi di archiviazione forniti da aziende o enti istituzionalmente preposti a questo scopo sono sempre più utilizzati dalle organizzazioni che preferiscono evitare di realizzare e poi mantenere nel tempo una propria infrastruttura tecnologica di archiviazione e conservazione e preferiscono demandare il compito a società terze che possono fornire non solo la tecnologia, ma anche le competenze, molto specialistiche, e l'assistenza necessarie. La domanda è in forte crescita e questo ha favorito la nascita, negli ultimi quattro-cinque anni, di diverse aziende che offrono questo genere di servizi.[44]

---

[41] Sebbene il dibattito sul modello organizzativo sia ancora in corso, è presumibile che il compito di archiviare il web e i social media non possa essere demandato ad una sola organizzazione; semmai si tratterebbe di affidarlo ad un insieme di istituzioni della memoria (ad esempio, le biblioteche o gli archivi nazionali) ma probabilmente sarebbe opportuno che anche i singoli enti, le aziende e forse anche i privati cittadini provino a dotarsi di sistemi autonomi di web e social media archiving che possano venire in soccorso laddove le istituzioni della memoria non siano in grado – sia per la vastità del materiale da raccogliere che per questioni di vario genere, da quelle tecnologiche a quelle finanziarie, fino alla carenza di risorse umane da destinare allo scopo – di assolvere al loro compito.

[42] Una Application Programming Interface (API) è una procedura che viene resa disponibile ai programmatori per la creazione di applicazioni software. Sia Facebook che Twitter rendono disponibili le proprie API per consentire lo sviluppo di applicazioni che interagiscono con le loro piattaforme.

[43] Cfr. nota 17.

[44] Per una disamina delle varie soluzioni ad oggi disponibili si rimanda al *White Paper on Best Practices for the Capture of Social Media Records* [22].

Conclusioni

Come si è visto, l'archiviazione e la conservazione delle risorse presenti sul web e sui social media sono minacciate da tutta una serie di criticità – che riguardano vari aspetti: archivistico-biblioteconomico, tecnologico, giuridico, organizzativo, economico, etc. – e che fanno si che molti contenuti scompaiano: [21]

> «dentro un meccanismo quasi biologico di cancellazione e perdita di memoria digitale, altri per ragioni di sostituzione tecnologica dentro la quale l'innovazione vince ogni volta nei confronti della conservazione, ma la grande maggioranza dell'intelligenza del mondo si allontana oggi da noi per ragioni economiche.»

L'aspetto economico non è da sottovalutare: la sostenibilità di siti web e piattaforme social dipende in larga misura dalla remunerazione che ne traggono le aziende che li gestiscono o dalla disponibilità di finanziamenti: se non sufficientemente remunerative, o se vengono a mancare i finanziamenti, o, ancora, se i costi di gestione aumentano eccessivamente,[45] esse vengono dismesse e condannate all'oblio senza particolari remore [25].

Queste considerazioni imporrebbero di agire subito ed avviare senza indugio iniziative di archiviazione e conservazione[46], insieme ad iniziative di sensibilizzazione e di formazione delle competenze e delle professionalità necessarie per condurre progetti di archiviazione e conservazione del web e dei social media. Questo è un aspetto particolarmente importante. Infatti, se è vero che il tema dell'archiviazione e conservazione del web e dei social media sta acquisendo una rilevanza sempre maggiore, bisogna, purtroppo, riconoscere che ad oggi le figure professionali capaci di condurre progetti in questo ambito sono poche se non quasi del tutto assenti [16], salvo casi eccezionali e certamente meritevoli di segnalazione (come quello della Biblioteca Nazionale di Firenze). Ciò dipende non solamente dall'insufficiente interesse che fino ad oggi è stato riservato a questi temi ma anche dalla mancanza di percorsi formativi che sarebbero invece estremamente importanti anche in considerazione delle difficoltà – non solo di tipo tecnico ma anche economico ed organizzativo – che si incontrano per portare a termine progetti di questa natura. Purtroppo, su questo punto, la situazione in Italia appare molto in ritardo rispetto agli altri paesi europei ed enormemente in ritardo rispetto ai paesi dell'area anglosassone. Tuttavia recentemente sono state avviate alcune iniziative meritevoli di essere citate. Una di queste è l'attivazione nel 2021 della Summer school in "Web and social media archiving and preservation" presso l'Università degli Studi di Bologna e nata con

[45] Si, pensi, ad esempio ai costi di gestione dei data center che ospitano i siti web o le piattaforme social, che sono aumentati a dismisura a seguito della crisi energetica dell'ultimo anno e del conseguente incremento delle spese per la fornitura di energia elettrica.

[46] Purtroppo, per quanto auspicabile, l'individuazione di una strategia condivisa a livello nazionale riguardo la conservazione e la salvaguardia – anche parziale – dei contenuti del web e di quelli presenti sui social media, è un traguardo ancora molto lontano da raggiungere, anche in considerazione delle notevoli difficoltà che, come si è visto, occorre superare.

l'intento di «offrire una formazione di alto livello sui temi emergenti dell'archiviazione e conservazione dei siti web e dei social media, che rappresentano una nuova e diversificata tipologia di materiale la cui conservazione è imprescindibile per tutta una serie di ambiti scientifici (si pensi alla ricerca storica, sociologica, antropologica, etc.) ai fini della futura ricostruzione dell'attuale civiltà».[47] La Summer school intende «fornire le conoscenze e le competenze necessarie per favorire lo sviluppo di nuove professionalità ed avviare nuovi percorsi lavorativi da parte dei discenti interessati».[48] Con questa iniziativa, sorta proprio per venire incontro alle nuove esigenze di preservazione dei contenuti web e social, e con quelle che auspicabilmente verranno messe in campo, ci si propone di formare le professionalità che saranno in grado di conservare per il futuro almeno quella parte del web e dei social media che è fondamentale per la ricostruzione della nostra epoca e senza la quale la storia e le altre discipline interessate allo studio della civiltà ne risulterebbero irrimediabilmente menomate.

Riferimenti

[1] Allegrezza, S. (2015). "Archiviare e conservare i social media. Una sfida per gli archivisti". *AIDA Informazioni*, 2015, 1-2.

[2] Allegrezza, S. (2015). "Nuove prospettive per il Web archiving: gli standard ISO 28500 (formato WARC) e ISO/TR 14873 sulla qualità del Web archiving". *Digitalia*. Vol. 2015, pp. 49-61. http://digitalia.sbn.it/article/view/1473/981.

[3] BESOCIAL. (2020)."Towards a sustainable social media archiving strategy for Belgium, WP1 Report, An international review of Social Media Archiving initiatives", (M1-M6: December 2020) https://www.kbr.be/wp-content/uploads/2020/07/202012_BESOCIAL_Report_WP1_Review_of_existing_social_media_archiving_projects.pdf.

[4] Betta, E. Romanelli, R. (2007). "Internet come fonte?". *Dimensioni e problemi della ricerca storica*, Carocci, 2/2007, luglio-dicembre. ISSN: 1125-517X. DOI: 10.7376/72419.

[47] La prima edizione della Summer school, che si è svolta dal 6 al 10 settembre 2021, ha visto la partecipazione di quasi 40 discenti tra professionisti dei beni culturali (archivisti, bibliotecari, operatori museali), informatici, funzionari di enti pubblici ed aziende private, studenti/dottorandi in Library and information science e in Digital humanities oltre che persone interessate a vario titolo alle questioni legate all'archiviazione e alla conservazione dei siti web, dei blog e dei social media. La stessa partecipazione è stata riscontrata nella seconda edizione, che si è svolta dal 5 al 9 settembre 2022, a conferma dell'interesse su questi argomenti. Maggiori informazioni sulla Summer School sono disponibili sul suo sito web: https://site.unibo.it/web-and-social-media-archiving-and-preservation/it.

[48] Cfr. il sito della Summer school già citato.

[5] Bracciotti, L. (2019). "Il Web Archiving. Conservazione e uso di una nuova fonte". *Officina della Storia*. 10 gennaio 2019. https://www.officinadellastoria.eu/it/2019/01/10/ il-web-archiving-conservazione-e-uso-di-una-nuova-fonte.

[6] Bracciotti, L., (2020). "Pandemia e web archiving. Conservare le fonti online #igiornidellapandemia". *Il mondo degli archivi*. 2 maggio 2020, http://www.ilmondodegliarchivi.org/rubriche/archivi-digitali/815-pandemia-e-web-archiving-conservare-le-fonti-online-igiornidellapandemia.

[7] Costa, M., Gomes, D. & Silva, M.J. (2017). "The evolution of web archiving". *International Journal of Digital Libraries*, 18, 191–205 (2017). https://doi.org/10.1007/s00799-016-0171-9.

[8] De Luna, G. (2020). "Il web e la sfida insidiosa alla storia". *Gilda Professione Docente*, n. 1 - Gennaio 2020. https://gildaprofessionedocente.it/news/dettaglio.php?id=790.

[9] Dickson E., Ilya Kreymer. (2021). "Announcing WACZ Format 1.0". https://webrecorder.net/2021/01/18/wacz-format-1-0.html.

[10] Digital Preservation Coalition. (2013). "DPC Technology Watch Report: Web-Archiving". https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file.

[11] Giungato, L. (2022). "Memorie dal sottosuolo digitale: frontiere e prospettive del social web archiving". 28 Lug 2022. https://www.agendadigitale.eu/cultura-digitale/memorie-dal-sottosuolo-digitale-frontiere-e-prospettive-del-social-web-archiving.

[12] ISO 28500:2017 — Information and documentation — WARC file format, https://www.iso.org/obp/ui/#iso:std:iso:28500:ed-2:v1:en.

[13] ISO/TR 14873:2013 - Information and documentation — Statistics and quality issues for web archiving, https://www.iso.org/obp/ui/#iso:std:iso:tr:14873:ed-1:v1:en

[14] J. Masanés. (2006). "Web Archiving". Berlin. Springer. 2006. In particolare p. 7.

[15] Laakso, M; Matthias, L, & Jahn, N. (2017). "Open is not forever: a study of vanished open access journals". ArXiv, DOI: 10.1002/asi.24460.

[16] Landino, C. (2018). "Strumenti per il web archiving". 27 agosto 2018, https://www.webarchiving.it/2018/08/27/strumenti-per-il-web-archiving.

[17] Landino, C., Marzotti, L. (2018). "Memorie dinamiche". Roma. Edizioni ANAI. 2018.

[18] Landino, C.; Marzotti, L. (2019). "Perché dovremmo pensare al web archiving". Cantieri PA. 20 marzo 2019. https://www.forumpa.it/pa-digitale/gestione-documentale/perche-dovremmo-pensare-al-web-archiving.

[19] Landino, C. (2018). "Strumenti per il Web Archiving: alcune soluzioni". in Il mondo degli archivi, 6 luglio 2018.

[20] Manetta, L. (2017). "La storia di fronte al digital turn". In: *Fare storia nella società dell'informazione. Teorie, modelli, pratiche*. (tesi di laurea magistrale, relatore prof. Maurizio Vivarelli, Università degli Studi di Torino, a.a. 2015-2016) https://www.academia.edu/40678188/La_storia_di_fronte_al_digital_turn

[21] Mantellini, M. (2021) "Internet, in effetti, si è rotta". *Il post.* 11 ottobre 2021. https://www.ilpost.it/massimomantellini/2021/10/11/internet-in-effetti-si-e-rotta.

[22] National Archives and Records Administration (NARA). (2013). "White Paper on Best Practices for the Capture of Social Media Records". https://www.archives.gov/files/records-mgmt/resources/socialmediacapture.pdf.

[23] Rulent, M. (2017). "L'archiviazione web agli archivi storici dell'Unione Europea". In: Becherucci, A., Capetta, F. (a cura di). *The Net. La rete come fonte e strumento di accesso alle fonti*. Roma. Edizioni di Storia e Letteratura. 2017.

[24] Shechmeister, M. (2009). "Ghost Pages: A Wired.com Farewell to GeoCities". *Wired.com*. 3 novembre 2009. https://www.wired.com/rawfile/2009/11/geocities.

[25] Signorelli, A. D. (2021). "Internet si è rotta: coi siti cancellati sparisce la nostra memoria collettiva". *Domani*. 10 ottobre 2021. https://www.editorialedomani.it/tecnologia/internet-si-e-rotta-coi-siti-cancellati-sparisce-la-nostra-memoria-collettiva-rcifvprt.

[26] Steinhauer, J. (2015). "Preserving Social Media for Future Historians". 24 luglio 2015. https://blogs.loc.gov/kluge/2015/07/preserving-social-media-for-future-historians.

[27] Storti, C. (2019). "Web archiving, 'sfida culturale': il servizio della Biblioteca Nazionale Centrale di Firenze", Cantieri PA. 12 giugno 2019. https://www.forumpa.it/pa-digitale/gestione-documentale/web-archiving-sfida-culturale-il-servizio-della-biblioteca-nazionale-centrale-di-firenze.

[28] UNESCO. (2003). "Charter on the Preservation of Digital Heritage". http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION =201.html.

[29] V. Banos, Y. Manolopoulos. (2015). "A quantitative approach to evaluate Website Archivability using the CLEAR+ method", *International Journal on Digital Libraries* (Springer). 12 Marzo 2015.

[30] Vitali, S. (2010). "La conservazione a lungo termine degli archivi digitali dello stato". In S. Pigliapcoco (a cura di), *Conservare il digitale*. EUM Edizioni Università di Macerata. 2010.

[31] Vlassenroot, E., Chambers, S., Lieber, S. et al. (2021). "Web-archiving and social media: an exploratory analysis". *International Journal of Digital Humanities.* 2, 107–128 (2021). https://doi.org/10.1007/s42803-021-00036-1.

[32] Wikipedia. "List of Web archiving initiatives". https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives#Archived_data.

[33] Zittrain, J., Kendra Albert & Lawrence Lessig. (2014). "Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations". *Legal Information Management.* 14(02):88-99. DOI: 10.1017/S1472669614000255.