BIBLIOTECA 2.0

Una risorsa per la ricerca e la conservazione del digitale: Internet Archive

ABSTRACT

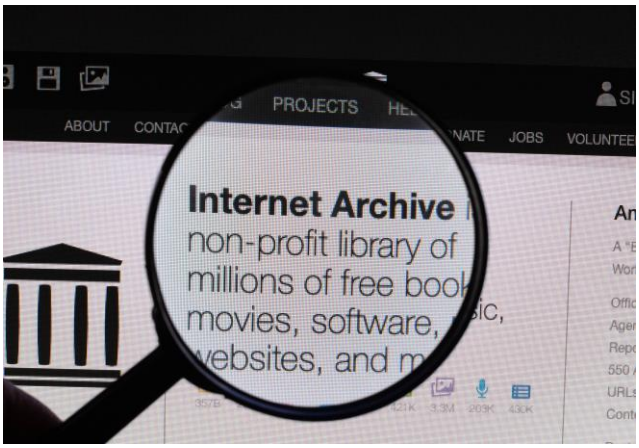
In occasione del venticinquennale dalla nascita di Internet Archive e del recente lancio del portale Internet Archive Scholar, ci ricolleghiamo al recente articolo in "MinervaWeb", n.s., 64 di agosto 2021 con un approfondimento sulla piattaforma digitale statunitense. Relativamente poco conosciuta e utilizzata in Italia, negli ultimi anni sta riscontrando una crescente popolarità grazie all'offerta di servizi unici nel suo genere. In questo articolo ne tracciamo brevemente la storia, per poi descrivere i vari progetti che la compongono e soffermarci infine su Internet Archive Scholar.

****

SOMMARIO

****

1. *La storia di Internet Archive*



Internet Archive è un'organizzazione no-profit nata nel 1996 nella Silicon Valley dal genio dell'imprenditore Brewster Kahle. Il progetto iniziale riguardava esclusivamente la realizzazione di un sistema automatizzato di archiviazione del World Wide Web, la cosiddetta Wayback machine; successivamente, grazie all'estensione dei servizi di archiviazione e autoarchiviazione al di là e al di fuori del web, è nata una vera e propria *digital library*, ossia sono stati avviati progetti di conservazione e diffusione di contenuti digitali come libri, registrazioni televisive, software, immagini, video e molto altro. La sede principale del progetto è localizzata a San Francisco, ma la crescente necessità di nuovi spazi di archiviazione e backup si è risolta

affidandosi a server sparsi per il mondo intero: basti pensare che ogni file digitale è conservato almeno in doppia copia.

Ad oggi l'intera mole di materiali digitali liberamente consultabili su Internet Archive supera i 70 *petabyte* .

Tutto il materiale digitale messo attualmente a disposizione da Internet Archive può essere oggetto di copie e consultazioni potenzialmente infinite, ma nasconde sfide di conservazione e fruizione nel tempo che ancora oggi sono al centro di un dibattito internazionale. Dietro alle potenzialità offerte da un sistema di archiviazione centralizzato, come quello proposto da Internet Archive, si celano problematiche relative al possibile deteriorarsi dei server di archiviazione, e al continuo sviluppo di nuove tecnologie che rendono obsoleti alcuni formati. Per fronteggiare questo problema la piattaforma statunitense si avvale di un continuo sistema di backup dei propri dati e di due sistemi di conservazione digitale, meglio conosciuti come 'migrazione' ed 'emulazione'. Per migrazione si intende la conversione di file digitali obsoleti in formati supportati dalle tecnologie esistenti; di contro, con l'emulazione si vuole ricreare l'ambiente software obsoleto al fine di leggere i materiali digitali non più supportati dalle moderne tecnologie.

****

2. La Wayback machine

Archive.org è l'accesso principale alle collezioni contenute in Internet Archive, il cui primo nucleo è rappresentato dai siti web archiviati mediante la Wayback machine. Questa è consultabile gratuitamente da tutti a partire dal 2001, anno di lancio ufficiale della piattaforma digitale realizzata da Alexa Internet. Il funzionamento è piuttosto semplice quanto efficace: una serie di *bot* (programmi automatizzati di raccolta dati), detti *crawler*, analizzano l'intero web in modo automatico raccogliendo informazioni e istantanee sui diversi siti in rete. Queste istantanee sono rese come pagine *html*, corredate da *hyperlink* che permettono la navigazione come nel sito originario, al fine di riprodurre fedelmente l'aspetto delle pagine nell'esatto momento del passaggio del *crawler*. Anche i materiali digitali 'catturati' in questo modo vengono conservati nei server di Internet Archive.

Questo strumento permette di studiare l'evoluzione dei siti web nel corso del tempo, memorizzandone le varie modifiche e rendendoli nuovamente disponibili nel caso in cui venissero cancellati dalla rete. Di contro, i siti così catturati non sono accessibili per intero: basti pensare ai database interni inaccessibili ai *crawler* o al famigerato *deep web*. Inoltre, la memorizzazione delle pagine non avviene giorno per giorno ma è scandita nel tempo in base alla popolarità del sito stesso, dunque alcune variazioni potrebbero non essere catturate.

I siti web sono inoltre oggetto di copyright: alcune realtà, difatti, potrebbero non voler essere tracciate dai *crawler*, ponendo la questione su un piano legale svantaggioso per l'organizzazione. Per ovviare a ciò, Internet Archive segue l'Oakland Archive Policy che prevede la possibilità degli amministratori dei siti web di tutto il mondo di impedire la cattura delle informazioni presenti sul proprio sito web, attraverso delle istruzioni da inserire nel file *robots.txt*. In alternativa i *webmaster* possono cancellare tutte le informazioni già presenti sulla Wayback machine, relative al proprio sito, facendone richiesta. (Segnaliamo a questo proposito l'interessante articolo di Ludovica Price, *Internet Archiving - The Wayback machine*)

Digitando il sito web che vogliamo analizzare all'interno del motore di ricerca della Wayback machine, sarà possibile visualizzare quante volte nel corso degli anni è stato catturato, mostrando nella *timeline* data e ora del passaggio del *crawler*. Di seguito un'immagine di come appariva il sito web del Senato della Repubblica il 15 febbraio 1997.

Una filiazione della Wayback machine è Archive-it.org, un servizio di archiviazione web in abbonamento nato nel 2006 allo scopo di conservare il materiale digitale proveniente dai siti web di biblioteche, musei, organizzazioni e istituti culturali presenti in tutto il mondo. Il materiale raccolto sarà conservato permanentemente nei server dell'organizzazione, garantendone la fruizione anche nel caso di cancellazione del sito web. A differenza della Wayback machine, Archive-it è una suite completa di gestione, controllo e raccolta delle collezioni digitale: questi strumenti permettono alle istituzioni di riorganizzare e personalizzare il sistema di raccolta di informazioni in modo completo, superando la 'staticità' dei *crawler* del servizio base.

In Italia, un esempio di utilizzo in ambito culturale di Archive-it proviene dalla Biblioteca nazionale centrale di Firenze, che all'interno del servizio 'Magazzini Digitali', raccoglie, conserva e rende accessibili in maniera permanente contenuti web di interesse per la cultura e la storia d'Italia. Per farlo,

dal 2018, essa ha avviato un programma di *Web archiving*, avvalendosi proprio della piattaforma Archive-it.

****

3. *Digital library e OpenLibrary*

Con l'estensione dei servizi di archiviazione e autoarchiviazione, Internet Archive ha creato una vera e propria *digital library*, estendendo a chiunque la possibilità di caricare contenuti digitali in piattaforma, pur con la necessaria attenzione che i materiali condivisi non siano soggetti a copyright. Per questa motivazione la maggioranza delle risorse condivise sono di pubblico dominio (come opere letterarie antiche) oppure materiali condivisibili perché distribuiti secondo licenze di utilizzo 'liberatorie' come la CC-BY delle Creative Commons (per altre informazioni sulle licenze Creative Commons si veda l'articolo sul n. 47, n.s., di "MinervaWeb").

Data l'eterogeneità delle risorse digitali in essa contenute la *digital library* di Archive.org consta di numerosi portali, ognuno dei quali dedicato a uno specifico progetto. Ad esempio, in Software Archive è possibile 'avventurarsi' fra migliaia di giochi e programmi software legalmente scaricabili, ma che il progresso tecnologico cancellerebbe per sempre dalla rete. Live Music Archive è invece dedicato alla conservazione di concerti live, o streaming, di artisti che condividono la propria musica con una politica non commerciale. Di notevole interesse per gli studiosi e per il mondo delle biblioteche sono i progetti dedicati al mondo del libro, come i portali dedicati ai libri digitali in prestito, ai microfilm e alle riviste. Per una panoramica completa dei diversi progetti si consiglia di visitare il sito Archive.org, punto di accesso alla biblioteca digitale.

Anche le biblioteche possono collaborare attivamente con la *digital library*, contribuendo alla disseminazione della conoscenza in rete. Difatti, fra i portali maggiormente consultati dagli utenti troviamo quello delle American Libraries che, lanciato nel lontano 2006, conta ad oggi un patrimonio di tre milioni di oggetti digitali caricati. Anche le biblioteche italiane sono sempre più attive su questo fronte; ricordiamo qui l'esempio della Biblioteca Nazionale Centrale di Firenze che dal 2014 si occupa di condividere in piattaforma digitalizzazioni di opere delle proprie collezioni. A tal proposito, informiamo i nostri lettori che presso la Biblioteca del Senato è in atto uno studio di fattibilità al fine di valutare un'estensione dei servizi digitali attraverso la *digital library* di Internet Archive.

Un altro servizio di Internet Archive dal sapore squisitamente bibliotecario è OpenLibrary, una biblioteca digitale lanciata nel 2006 grazie alle sovvenzioni della California State Library e della Kahle/Austin Foundation (fondata da Brewster Kahle e sua moglie). Scopo del progetto è la realizzazione di un vero e proprio 'catalogo' di tutti i libri esistenti, fornendo quando possibile l'accesso diretto alla copia digitale. Le risorse presenti possono essere in formato testuale o audiolibri, sia liberamente consultabili - quando esenti da copyright - che scaricabili secondo i modelli del *digital*

lending bibliotecario. Inoltre, grazie all'integrazione con altre piattaforme, come il catalogo WorldCat, è possibile ricercare l'esistenza di una versione cartacea nelle biblioteche della propria città. Le biblioteche di tutto il mondo possono partecipare attivamente al progetto OpenLibrary, effettuando donazioni di materiale librario o condividendo con Internet Archive le digitalizzazioni, contribuendo così alla disseminazione della conoscenza attraverso il web.

Durante l'emergenza pandemica, nel corso del 2020, OpenLibrary è stata al centro di un dibattito internazionale legato al diritto d'autore. Per contrastare le problematiche di accesso alla conoscenza dovute alla chiusura di molte biblioteche, l'intero patrimonio librario consultabile tramite il sistema del prestito digitale è stato reso liberamente fruibile senza limiti di consultazione. Inizialmente, il progetto della National Emergency Library si sarebbe dovuto concludere a fine giugno 2020, ma le accuse di violazione del copyright da parte degli editori hanno portato alla chiusura anticipata dell'iniziativa al 16 giugno.

****

4. Una nuova risorsa: Internet Archive Scholar

L'attività di Internet Archive si è recentemente ampliata nella direzione dell'archiviazione e conservazione dei prodotti scientifici, che è da sempre uno dei presupposti alla base dell'accesso aperto. Alcuni studi recenti hanno fatto emergere dati non confortanti riguardo il recupero delle informazioni a distanza di tempo, come nel caso delle 174 riviste *open access* pubblicate fra il 2000 e il 2019 non più disponibili in rete, indagate nel recente articolo di Mikael Laakso, Lisa Matthias e Najko Jahn, *Open is not forever: A study of vanished open access journals*. Questo problema sembra riguardare esclusivamente gli articoli contenuti nelle riviste ad accesso aperto - ovvero pubblicati secondo i principi della *gold road* - e non la documentazione conservata presso *repository* istituzionali - la cosiddetta *green road* - poiché non soggetta a problematiche di conservazione a lungo termine.

Il recente lancio di Internet Archive Scholar si pone dunque come un valido strumento per contrastare la problematica della conservazione digitale di questi articoli e si affianca al servizio simile e ben più noto, già da tempo lanciato da Google, Google Scholar.

La caratteristica del servizio Scholar di Internet Archive è fornire l'accesso alle diverse pubblicazioni *open access* 'catturate' dai servizi di Internet Archive e conservate presso i propri server. La pagina principale si presenta con un'interfaccia intuitiva attraverso cui è possibile lanciare ricerche semplici, ma anche complesse come vedremo più avanti, fra i *full text* di oltre 25 milioni di pubblicazioni

accademiche. Quando possibile il sistema fornirà all'utente la copia originale del documento ricercato, ma a volte può essere proposta una versione alternativa all'originale, come quella contenuta in una pagina *html* di un sito web. Ad ogni modo, i link di accesso alle risorse sono corredati da *tag* con informazioni aggiuntive sulla provenienza della copia; così facendo l'utente avrà sempre una panoramica chiara della documentazione ricercata, in special modo nel caso di versioni multiple della stessa copia. Le risorse sono conservate nelle collezioni di Internet Archive, come la Wayback machine, Archive-it, Archive.org, mentre i metadati provengono dal catalogo *open* Fatcat, garanzia di qualità in ambito di pubblicazioni accademiche. Il servizio - presentato alla conferenza del 2019 FORCE911 - è ancora nella fase beta del progetto, ed è dunque in continuo aggiornamento anche grazie ai feedback degli utenti.

Il sistema di *information retrieval* - ovvero di recupero delle informazioni ricercate - adottato dalla piattaforma non è dotato di una maschera di ricerca avanzata. Tuttavia, per interrogare la piattaforma è possibile sfruttare le potenzialità dei metadati, ossia le informazioni necessarie a identificare un documento (titolo, autore, data ecc.), e della sintassi di Lucene.

Lucene è una libreria Java ad accesso aperto, che permette di indicizzare in modo strutturato i *full text* delle pubblicazioni e di realizzare *query* di ricerca particolarmente potenti e complesse. Ad esempio, se si volessero recuperare documenti afferenti all'archeologia biblica a partire dal 2000, si digiterà quanto segue: *"biblical archaeology" access_type:ia_sim year:<2000*. Malgrado l'indubbia efficacia di questo strumento, l'esperienza utente risente della complessità nel formulare *query* così strutturate: probabilmente si tratta di una situazione temporanea dovuta alla fase sperimentale in cui la piattaforma si trova attualmente.

I metadati permettono invece di recuperare documentazione rispondente a determinate caratteristiche. È possibile anche restringere la ricerca ai casi in cui un campo esiste utilizzando un asterisco *author:*,* oppure negare un caso di ricerca utilizzando un punto esclamativo *!year:2005*. Se volessimo dunque ricercare tutti i lavori dell'autore Brian Fagan, escludendo eventuali monografie, basterà lanciare la ricerca in questo modo: *author:"Brian Fagan", !type:book* (un documento completo sull'utilizzo dei metadati è disponibile su Elasticsearch).

Infine, in via del tutto sperimentale è stata implementata anche una modalità di interrogazione tramite citazione bibliografica; in questo caso il sistema riconoscerà automaticamente che si tratta di una citazione e la confronterà con le bibliografie degli articoli. Per altre informazioni sul sistema di ricerca in Internet Archive Scholar, si consiglia di visionare la pagina web della Guida utente accademico.

Scholar si candida dunque a essere un servizio importante della costellazione di Internet Archive, che rende l'offerta della piattaforma sempre più completa e mirata, anche rispetto a esigenze informative più specialistiche e dettagliate.

****

5. Approfondimenti bibliografici

Considerata la relativa novità dell'argomento proposto rispetto al contesto bibliotecario, la specificità della risorsa descritta e la carenza di materiale monografico al riguardo, in questa particolare occasione – diversamente dalle altre uscite della rubrica "Biblioteca 2.0" – non viene proposto un percorso bibliografico nelle collezioni della biblioteca. Per approfondire le tematiche affrontate nell'articolo, si suggerisce in primo luogo la ricerca nelle banche dati consultabili dalle postazioni pubbliche della biblioteca, e in seconda battuta nel Catalogo del Polo bibliotecario parlamentare.