UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

La conservazione dei siti Web del LabCD: un servizio per l'Ateneo

Candidato: *Francesca Bertellotti*

Relatori: *Vittore Casarosa*

Enrica Salvatori

Correlatore: *Dario Besseghini*

Anno Accademico 2018-2019

Sommario

1. Introduzione

L'obiettivo di questa tesi di laurea è quello di affrontare il tema del *Web archiving* e di presentare una soluzione per la conservazione dei siti Web del Laboratorio di Cultura Digitale dell'Università di Pisa.

Infatti, in questi anni Internet è diventato una fonte giornaliera e continua di informazioni, scambio di dati e comunicazioni ed è entrato con i suoi servizi sempre di più nella nostra vita quotidiana con l'online banking, le e-mail, lo shopping e il telefono mobile.

La tecnologia è così divenuta da strumento di pubblicazione a strumento di comunicazione, non solo personale ma anche politica e scientifica, e i suoi usi vanno ben al di là della semplice navigazione in Internet.

Per lungo tempo si è pensato al Web come a uno strumento in grado di auto preservarsi e non si è prestata molta attenzione alla raccolta dei contenuti e alla loro conservazione a lungo termine, facendo affidamento sulla rete e sulle sue connessioni estremamente variabili.

La natura fluida e irregolare di Internet porta al fatto che ogni sito Web può cambiare e i contenuti possono scomparire senza lasciare nessuna traccia. Tutto il materiale presente nel Web è di fatto transitorio, fragile ed effimero.

Generalmente soprattutto tra gli utenti e i non specialisti prevale la convinzione che qualsiasi contenuto sarà reperibile anche a distanza di tempo, avendo sempre la possibilità di accedere alle risorse, di ottenerle e di modificarle, ma di fatto stiamo vivendo quella che può essere considerata a tutti gli effetti una perdita di memoria e di storia.

Uno tra i numerosi esempi che è possibile riportare è il caso "*The Crossing*" di Kevin Vaughan. Si trattava di una serie Web in 34 parti, finalista del premio Pulitzer, che riportava la storia di venti bambini che avevano perso la vita in un incidente nel 1961. Nel 2009 la rivista per cui lavorava Vaughan chiuse, il sito Web fu abbandonato e l'intera serie scomparve dalla rete dimostrando che è possibile la perdita di qualsiasi materiale, pure di un elaborato finalista di un premio così prestigioso.

Già agli albori del Web è stata riconosciuta l'importanza di preservare l'informazione digitale ma è solo verso la fine degli anni '90 che si è iniziato a lavorare a iniziative di *Web archiving*

vero e proprio, al processo di raccolta e conservazione di porzioni del World Wide Web con lo scopo di tramandarlo agli scienziati, agli storici e ai posteri.

Il primo problema da affrontare è l'estensione sterminata del Web, che rende di fatto impensabile l'idea di conservarlo nella sua interezza. È quindi necessario procedere in modo selettivo, considerando l'importanza dei siti che vogliamo conservare e stabilire ogni quanto il sito debba essere salvato. Decidere cosa è destinato o meno all'oblio è un'operazione delicata che porta ad affrontare problemi culturali, di identità e alla precisazione di cosa intendiamo con il termine "storia", col fine di definire cosa merita di essere ricordato per le future generazioni e cosa può descrivere al meglio gli anni più recenti dell'umanità.

All'interno del Laboratorio di Cultura Digitale (LabCD) dell'Università di Pisa è nata la necessità di conservare i progetti che hanno portato alla realizzazione di siti Web, salvarne le modifiche e memorizzarli su supporti esterni, al fine di poter accedere ai contenuti anche a lunga distanza di tempo, per avere a disposizione le scoperte e gli studi effettuati negli anni.

Nella seconda parte dell'elaborato viene presentata una possibile soluzione a questa esigenza: verrà spiegato cos'è *Heritrix* (uno dei più diffusi *software* per l'archiviazione di siti Web) e come può essere utilizzato, e verrà presentato come viene fatto il salvataggio dei siti e la loro catalogazione. Verranno anche messe in luce le problematiche incontrate e le soluzioni adottate.

Si auspica che l'Università possa utilizzare il *Web archiving* per conservare e preservare i propri materiali, dalle nuove ricerche alle tesi di laurea, per inserirsi così all'interno di un panorama dove l'archiviazione Web in Italia è ancora acerba sotto tanti punti di vista. Molto si sta facendo, ma ancora di più va ancora fatto; in mancanza di una precisa legislazione, le biblioteche e gli archivi sul territorio nazionale si stanno impegnando per colmare il *gap* culturale e di responsabilità presente per quanto riguarda la salvaguardia della memoria digitale.

2. Dalle biblioteche al Web come deposito della conoscenza e delle tradizioni

Fin dall'antichità tutto il sapere e la conoscenza che venivano prodotti ed era necessario ricordare, o per fini culturali o utilitaristici, venivano trascritti. Il materiale scritto è diventato così una forma esterna di memoria, un deposito di conoscenza e di tradizioni, ed è andato a costituire una testimonianza fondamentale delle credenze e delle abitudini dei popoli.

Nei secoli e a seconda delle zone i supporti di scrittura variavano: tavolette d'argilla, papiro, cera su legno o incisioni su pietra, alcune giunte fino a noi. Grazie a questi reperti oggi possiamo avere un'idea del tipo di cultura, religione e organizzazione socioeconomica delle varie popolazioni e possiamo interrogarci sul passato dell'umanità.

Gli scritti spesso venivano raccolti in archivi e generalmente si trattava di testi di legge, atti amministrativi, sentenze giudiziarie o contabilità di magazzini, solamente in seguito si iniziarono a scrivere veri e propri libri. La biblioteca divenne il luogo dove conservare e consultare questi documenti scritti.

Durante il Medioevo e con la caduta dell'Impero Romano le biblioteche subirono un'importante decadenza, in questo periodo fu importante per la preservazione di documenti politici e letterari l'azione svolta dai monaci all'interno dei monasteri e nelle abbazie dove veniva operata la trascrizione manuale di libri antichi e di atti ufficiali.

Il Rinascimento è stata un'epoca fondamentale per la nascita delle biblioteche in senso moderno specialmente grazie all'interesse dei nobili signori che videro nella costruzione all'interno delle proprie residenze di zone dedicate ai libri un simbolo di prestigio e di affermazione di potere. Ne sono un esempio la Malatestiana di Cesena, l'Estense a Ferrara, la Laurenziana di Firenze e la Marciana di Venezia.

La caratteristica fondamentale che accomuna questi periodi dell'umanità è la presenza di un supporto fisico e materiale, che esso sia pietra, argilla, papiro o pergamena, che ha lo scopo di mantenere lo scritto rendendolo, in linea teorica, permanente e statico. Con l'introduzione di Internet e la pubblicazione in rete dei contenuti questi documenti hanno perso questa proprietà fondamentale. L'intangibilità ne permette sì, una miglior diffusione, modifica e condivisione, ma rende possibile anche la perdita totale del prodotto.

Negli ultimi venti-trenta anni gran parte della produzione dell'ingegno umano è stata pubblicata sul Web e spesso non ha una sua controparte fisica all'interno di archivi e biblioteche, rendendola di fatto del tutto effimera e provvisoria. Oggigiorno tesi di laurea e di dottorato, dati sensibili, musica, ricerche, scoperte scientifiche, articoli di giornale, inchieste, e-book, home banking, tutto viene pubblicato su questa enorme piattaforma, il Web, che (secondo le statistiche mantenute sul sito www.Internetworldstats.com) a Giugno 2019 contava circa quattro miliardi e mezzo di utenti in tutto il mondo[1].

Negli ultimi anni si è assistito anche a una crescita d'importanza dei social network, influenti per la vita sociale e politica e ultimamente fonte principale di oggetto di ricerca in ambito antropologico e culturale, che porta alla ridefinizione di distanza e vicinanza, di comunità e massa.

Alla fine del 2016 la Casa Bianca ha comunicato le modalità di conservazione dei profili social di Barack Obama e nel 2017 è stata approvata la *Covfefe Act (Communications Over Various Feeds Electronically for Engagement Act),* che prende il nome da una parola errata di un tweet di Donald Trump e che richiede agli archivi nazionali di conservare i post social del Presidente degli Stati Uniti in quanto sono atti a tutti gli effetti che documentano la vita politica e il consenso di una affermazione [2].
Internet è così divenuto un deposito di conoscenza e tradizioni: racchiude non solo nozioni, definizioni e spiegazioni ma anche le nostre attività quotidiane, abitudini, credenze, notizie e discorsi politici.

Due esempi che dimostrano l'importanza del materiale digitale sono riportati da Ralph Schroeder e Niels Brügger nel loro capitolo *The Web as a History* dell'omonimo libro.

Nel 2013 nel Regno Unito è stato scoperto che il partito conservatore aveva cancellato dai siti Web alcuni discorsi politici e ne aveva ridotto la possibilità di accesso tramite Google. Questa scoperta portò a ricercare questo materiale che era stato archiviato da una collezione speciale della *British Library*. Questo avvenimento, che aveva lo scopo di limitare e impedire la conoscenza di questi discorsi, mise in cattiva luce il partito conservatore attirando l'attenzione sulla pericolosità della memoria su Internet.

[1] *Internet World Stats*, https://www.Internetworldstats.com/stats.htm (visitato il 26/10/2019).

[2] *Officina della storia*, https://www.officinadellastoria.eu/it/2019/01/10/il-Web-archiving-conservazione-e-uso-di-una-nuova-fonte/ (visitato il 13/9/19).

Un altro caso è avvenuto nel 2014 durante la guerra tra Russia e Ucraina. Un russo ha comunicato tramite un post su un social di aver abbattuto un aereo ucraino, post immediatamente eliminato ma ritrovato tramite *Internet Archive* rendendo possibile la determinazione del responsabile[3].

Il punto fondamentale è che i documenti contano e questo vale sia per quelli cartacei che per quelli digitali e per molte altre fonti di prova sul passato.

[3] Brügger, Niels, e Ralph Schroeder, a c. di. *The Web as History*: *Using Web Archives to Understand the Past and the Present,* UCL Press. 2017.

3. Perché archiviare: l'importanza degli archivi Web come memoria collettiva

3.1 *Web as a history*: gli archivi Web e la conservazione della cultura

A partire dagli anni ‘90 il Web è divenuto parte integrante dell’infrastruttura comunicativa dell’età moderna. Si è sviluppato inizialmente come media a sé stante (come la radio, la televisione e i giornali) per poi essere intimamente coinvolto nella vita sociale, culturale e politica. Inoltre, nella vita di tutti i giorni, la cultura giovanile è la protagonista di questa dimensione e si manifesta tramite piattaforme come YouTube, Facebook e Instagram.

Una caratteristica importante da sottolineare per quanto riguarda la natura del Web, che lo differenzia da altre forme di comunicazione, è la sua capacità di non sottostare a forme di coercizione di natura politica o statale. Eccettuati alcuni casi, come la Cina, Internet è libero e garantisce in linea di massima uno scambio democratico, fluido e illimitato di informazioni e di notizie, accrescendo così la sua importanza come specchio autentico di una società.

Pertanto, possiamo aspettarci che nei prossimi anni la rete costituirà un importante oggetto di ricerca per gli studiosi di comunicazione ma anche per gli storici in generale che vorranno ottenere una visione approfondita di questioni importanti del recente passato.

Va considerato che, oltre che per gli storici, gli archivi Web saranno una fonte di studio anche per sociologi, antropologi, giornalisti, linguisti e per le altre figure professionali che possono attingere informazioni da questo vasto bacino di dati.

Le conclusioni possono riguardare come si è evoluta un’istituzione, la storia di un’impresa, come e cosa il governo ha deciso di comunicare con il pubblico e in quale periodo, gli effetti del commercio online su quello reale, quali prodotti vengono comprati su Internet e perché, lo sviluppo e le modifiche del linguaggio e tracciare i movimenti delle persone e delle idee politiche[4].

Per quanto gli archivi Web possano essere frutto di studio e di ricerche manca ancora una vera e propria cultura digitale: spesso gli umanisti sono privi di una formazione informatica e tecnologica adeguata a padroneggiare questo strumento. È chiaro che i ricercatori di discipline

[4] Ibidem.

umanistiche debbano acquisire nuove competenze e sviluppare nuove metodologie per utilizzare gli archivi Web come fonti.

Quindi dobbiamo tener presente che il Web è la nostra storia ed è necessario preservarne il contenuto per poterlo studiare e tramandare ai posteri. Va comunque considerato l'effettivo problema dato dalla mancanza di complete competenze digitali da parte degli umanisti, dalla gestione di grossi dati e dell'accesso agli archivi, ancora sempre troppo limitato.

3.2 *History of the Web*: gli archivi Web come testimonianza dell'evoluzione tecnologica

Il risultato del processo di *Web archiving* è costituito da una serie di pagine Web navigabili che corrispondono alle modifiche del sito al trascorrere del tempo. Un sito significativo che si occupa del processo di archiviazione Web e permette di visitare i siti più conosciuti e di vederne i cambiamenti è la *WayBackMachine.*

Questa piattaforma è stata creata nel 2001 dall'*Internet Archive*, un'organizzazione no profit localizzata a San Francisco. La *WayBackMachine* ha iniziato a conservare siti nel 1996 per poi essere lanciata e presentata al pubblico cinque anni dopo. Lo scopo di questa organizzazione è fornire libero accesso alla conoscenza per le persone e il motto è "*Our mission is to provide Universal Access to All Knowledge*"[5].

Si tratta di un archivio dove è possibile cercare un determinato sito, selezionare un periodo o un anno e visitarlo così com'era. Ad oggi contiene tra l'altro 20 milioni di libri messi a disposizione da Università, biblioteche e vari enti, 4.5 milioni di registrazioni audio e più di 330 miliardi di pagine Web.

Comunque il caso della *WayBackMachine* è un esempio di come queste risorse possano essere oggetto di studio per informatici e tecnici per il loro valore in sé e perché permettono di ripercorrere la storia di Internet negli anni: possiamo così parlare di *History of the Web* ovvero notare come è cambiato il World Wide Web per ricostruirne la storia.

L'attività di *Web archiving* permette così di studiare il design dei siti, le possibilità offerte dalla tecnologia e dalla grafica con il passare degli anni, gli interessi e gli strumenti messi a disposizione e così via. In questo senso gli archivi Web diventano testimonianza

[5] Sito della *WayBackMachine*, https://archive.org/about/ (visitato il 26/10/19).

dell'evoluzione tecnologica e informatica e permettono di accedere e visitare piattaforme che altrimenti sarebbero andate perdute, permettendo ad anni di distanza di ricordarle.

3.3 Le forme della memoria nel Web: scrittura, linguaggi fotografici e video

Una pagina Web contiene di base testo, immagini ed eventualmente video. Queste forme di espressione sono linguaggi della memoria ed entrano a far parte di ciò che viene raccolto e conservato durante il processo di *Web archiving*.

Un aspetto ovvio per il quale diventa necessario preservare i testi digitali è quello di conservare ciò che è conosciuto e i progressi fatti in ambito scientifico, medico e nelle altre discipline: raccolta e conservazione che precedentemente veniva garantita grazie alle biblioteche. Di fatto, senza documenti, si potrebbe fare solo una ricostruzione ipotetica del passato, tentativo di dubbia utilità [6].

Bisogna anche considerare che la lingua è una forma di organizzazione dell'esperienza entro la quale un individuo è immerso e diventa il filtro attraverso il quale noi recepiamo e pensiamo gli avvenimenti. Quindi, le parole, si intersecano con qualsiasi aspetto della cultura, espletando numerose funzioni allo stesso tempo: esprimere pensieri, essere il principio di scambi, rituali e cerimonie, costituendo il fulcro della vita comunicativa all'interno della società.

Il lessico è un deposito fondamentale in cui si stratificano significati ed elementi culturali relativi a diversi argomenti, per questo può essere interessante studiare ciò che viene scritto e il Web costituisce un enorme bacino dal quale attingere informazioni e dati che, essendo già digitalizzati, permettono una più facile analisi tramite programmi appositi.

Per i linguisti può essere interessante studiare le nuove parole entrate nel vocabolario, quali necessità di espressione hanno i parlanti, i neologismi, i forestierismi, i termini più utilizzati per esprimere emozioni e così via.

Un altro aspetto da considerare è che la parola, come suggerisce Monod (2009, p. 626), è molto di più della sua concretizzazione, del suono o del segno grafico, perché "implica una vicinanza, reale o fittizia, tra i protagonisti dello scambio verbale". Le parole uniscono chi le pronuncia, chi le ascolta, chi le scrive e chi le legge.

[6] Radcliffe-Brown, Alfred Reginald, *The Andaman Islanders*. 1922.

La scrittura diviene così oltre che uno spazio di memoria uno spazio di condivisione, che porta in sé la testimonianza dell'agire sociale, di movimenti del pensiero e di abitudini. Per questo diventa necessario preservare questa forma di memoria, così diffusa nel Web, e gli archivi diventano custodi di documentazione generalmente di natura pratica (politica, giuridica, economica, amministrativa), connessa ad attività svolte dai vari soggetti al fine di soddisfare specifiche esigenze[7], ma anche di documenti che costituiscono forme artistiche e legami.

Le immagini sono sempre state largamente utilizzate, in qualsiasi epoca e ben prima della scrittura. Possono essere definite come "oggetti da raccogliere e allo stesso tempo (…) contenitori con cui raccogliere dati. In ogni cultura circolano densi flussi di immagini, sotto forma di rappresentazioni di vario tipo prodotte con diverse tecnologie." (Pennaccini, 2012). A partire dal 1816, con la produzione della prima fotografia, si è passati dal dover rappresentare il reale tramite le capacità artistiche e la sensibilità degli artisti, alla possibilità di rappresentarlo così come appare all'occhio umano.

Negli ultimi anni l'insieme delle interazioni della comunicazione visiva si sono fatte sempre più complesse e fitte, contribuendo alla globalizzazione culturale. Infatti, i limiti imposti dalla fisicità dei singoli media (fotografia, cinema, televisione) sono andati svanendo con la diffusione del Web, e i prodotti tipici di uno specifico mezzo di comunicazione sono confluiti all'interno di questa piattaforma, creando un prodotto che contiene tutti gli altri. Basti pensare alla possibilità di vedere programmi Tv, pellicole cinematografiche, gallerie fotografiche sul computer o addirittura sul nostro telefonino.

La comunicazione tramite immagini viene largamente utilizzata, oltre come forma di espressione artistica, anche come forma di testimonianza, di prova o come pubblicità, e sono uno degli elementi principali dei siti Web e dei social network. Le immagini una volta percepite entrano nell'immaginario e diventano prodotti culturali che vengono trasmessi attraverso i media in continua evoluzione tecnologica. Il prodotto visivo diventa così un simulacro della realtà e una rappresentazione di come le persone l'hanno percepita.

La diffusione dei video ha contribuito in maniera notevole alla creazione di un'identità nazionale e in generale alla identificazione di un soggetto all'interno di un gruppo. Aver

[7] Zanni Rosiello, Isabella. *Andare in archivio*. *Orientamenti*. Bologna: Il Mulino. 1996.

condiviso la stessa esperienza, aver visto lo stesso filmato, diventa un'occasione per parlare e scambiare idee e commenti. Ma la loro funzione non si limita a questo.

Generalmente, all'interno dei siti Web, i video sono posti per illustrare meglio quanto viene descritto nel testo (ricette di cucina, documentari, come costruire un oggetto, etc.) e per mostrare le azioni da svolgere passo dopo passo. I filmati possono anche essere di puro intrattenimento (reality show e competizioni), per appagare il nostro gusto estetico (paesaggi e arredamento) o per suscitare la nostra curiosità (popolazioni lontane e scoperte scientifiche). La loro funzione può andare ben al di là del semplice accompagnamento di quanto scritto nel sito e possono costituirne il fulcro dell'attività comunicativa.

La necessità di preservare queste forme di comunicazione sta nell'importanza svolta a livello culturale e dalla possibilità data dai video di ricostruire cosa è considerato interessante e rilevante all'interno di una società. I video, come le immagini, mostrano ciò che conosciamo, gli abiti, gli usi, i costumi, ciò che piace e ciò che non piace; permettono di ricostruire l'identità di una popolazione in un dato momento storico.

Nel processo di *Web archiving* è necessario considerare la natura, anche dal punto di vista antropologico e culturale, di questi elementi per comprenderne le caratteristiche e poterli conservare al meglio. Un problema da non sottovalutare è infatti l'accesso a lungo termine; la tecnologia evolve continuamente e in modo molto veloce rendendo le informazioni digitali obsolete in pochi anni. Quando un software o un hardware non sono più disponibili o utilizzabili le informazioni possono non essere più raggiungibili. Si parla in questo senso di obsolescenza digitale.
Una sfida importante per l'archiviazione Web è non solo catturare i siti e conservarli ma anche preservarne l'intero contenuto a lungo termine.

4. Iniziative di *Web archiving*

L'importanza di preservare l'informazione digitale è stata riconosciuta solamente a partire dagli anni '90 e vari istituti, generalmente biblioteche e archivi, hanno attivato iniziative di conservazione dei propri contenuti, allargando le loro responsabilità rispetto a quanto definito dal deposito legale. Recentemente la tematica del *Web archiving* ha ottenuto un'attenzione sempre maggiore portando alla creazione di nuove associazioni, organizzazione no profit e servizi commerciali che si occupano della preservazione di siti Web.

Nel 2003, l'Organizzazione delle Nazioni Unite per l'educazione, la scienza e la cultura (UNESCO) ha considerato i materiali digitali come un patrimonio culturale e ha sollevato la necessità di agire per preservare questo patrimonio[8]. Di seguito sono riportate le iniziative più importanti in tale ambito.

4.1 *Internet Archive*: "*Universal Access to All Knowledge*"

L'Internet Archive è una biblioteca digitale no-profit, fondata nel 1996 da Brewster Kahle, negli Stati Uniti, e fa parte della IIPC (*International Internet Preservation Consortium*). Permette di accedere a collezioni storiche e materiali digitali e inizialmente l'archivio aveva contenuti provenienti da *Alexa Internet*, società creata dallo stesso ideatore dell'*Internet Archive*. Oggi collabora con più di 625 biblioteche e altri partner localizzati in varie regioni del mondo.

Negli ultimi anni hanno esteso il numero e la tipologia di contenuti archiviati. A partire dal 2005 si occupa, oltre che dei siti Web e delle risorse digitali, della digitalizzazione dei libri. Il contenuto attuale è costituito da:

- 330 milioni di pagine Web
- 20 milioni di libri e testi
- 4.5 milioni di registrazioni audio (inclusi 180.000 concerti dal vivo)
- 4 milioni di video (inclusi 1.6 milioni di programmi TV)
- 3 milioni di immagini

[8] Toyoda, M., e M. Kitsuregawa. *The History of Web archiving.* 2012. https://doi.org/10.1109/JPROC.2012.2189920.

- 200.000 programmi software

Qualsiasi persona, dopo la creazione di un account gratuito, può caricare i propri contenuti che verranno conservati dall'*Internet Archive*.

Per motivi di copyright, la piattaforma permette di scaricare liberamente i contenuti precedenti al 1923, mentre gli altri contenuti possono essere consultati online. Ogni giorno più di mille libri vengono digitalizzati in 28 diverse parti del mondo per conto dell'*Internet Archive*. Dopo il 2001, con l'attacco alle Torri Gemelle, *Internet Archive* si occupa anche di conservare le notizie riportate dalla televisione e dal 2009 anche di salvare i programmi televisivi.

Il sito Web dell'*Internet Archive* è tra i 300 siti Web più visitati al mondo, l'intera collezione occupa circa 45 Petabytes di spazio su memorie esterne e ogni risorsa viene archiviata almeno due volte [9].

4.2 PANDORA *Austrialian Web Archive*

Pandora (*Preserving and Accessing Networked Documentary Resources of Australia*) è l'archivio Web dell'Australia, fondato nel 1996 dalla Biblioteca Nazionale dell'Australia e adesso costituito dalla collaborazione con altre nove biblioteche e organizzazioni di raccolte culturali.

Le sue origini risalgono al 1995 quando la Biblioteca Nazionale ha evidenziato il problema del crescente numero di informazioni pubblicate online, successivamente ha accettato la responsabilità di raccogliere e conservare le pubblicazioni australiane, indipendentemente dal formato.

Nel gennaio del 1996 venne creato un comitato (*Selection Committee for Australian Online Publications*) per sviluppare linee guida per la selezione del materiale e nell'aprile dello stesso anno fu istituita l'Australian *Electronic Unit* (ribattezzata *Digital Archiving Section* nel 2003) per selezionare le pubblicazioni online secondo le linee guida, per negoziare con gli editori il diritto di archiviarle e catalogarle nel database bibliografico nazionale. Il piano si muoveva così in due direzioni: stabilire cosa archiviare e archiviare effettivamente il materiale.

[9] Sito dell'*Internet Archive*, https://archive.org/about/ (visitato il 28/10/2019).

L’anno dopo tutte le infrastrutture furono sufficientemente sviluppate per invitare le biblioteche statali a diventare partecipanti e dopo la *State Library of Victoria* aderirono anche tutte le altre biblioteche. Inizialmente PANDORA utilizzava un software di dominio pubblico che però presto divenne non sufficiente per gestire la quantità di dati: venne così creato il programma PANDAS.

L'Archivio PANDORA ha raggiunto i seguenti obiettivi:

- un archivio a livello mondiale di pubblicazioni online australiane selezionate, come riviste elettroniche, pubblicazioni governative e siti Web di ricerca o di rilevanza culturale;
- la definizione di politiche, procedure e linee guida di selezione per la raccolta e la fornitura di accesso a lungo termine agli articoli nell'Archivio;
- un approccio nazionale collaborativo all'archiviazione e alla conservazione a lungo termine delle pubblicazioni online australiane, coinvolgendo la partecipazione di biblioteche statali e altre istituzioni culturali;
- un sistema di archiviazione digitale (PANDAS) per semplificare la raccolta e il caricamento di pubblicazioni nell'archivio, definire le informazioni su di esse e gestire l'accesso pubblico;
- uno schema per nominare in modo persistente tutti gli oggetti nell'archivio;
- sviluppo di politiche e ricerca sulla conservazione delle pubblicazioni online, inclusi metadati di conservazione, migrazione e analisi dei rischi[10].

4.3 IIPC*: International Internet Preservation Consortium*

L’International Internet Preservation Consortium nasce nel 2003 dall’iniziativa di 12 “istituzioni della memoria” (principalmente Biblioteche Nazionali) con lo scopo di coordinare gli sforzi per il mantenimento del Web nel futuro. I partecipanti si impegnano a finanziare e partecipare a progetti e lavori di gruppo per raggiungere questo obiettivo. Nel 2010 contava già 35 enti aderenti all’iniziativa, ed oggi conta più di 50 istituzioni, costituite essenzialmente da biblioteche, archivi, musei e istituti per i beni culturali.

L’IIPC si è posto degli obiettivi che riguardano l’acquisizione, la preservazione e l’accessibilità di risorse provenienti da Internet per le generazioni future. In particolare, il gruppo di ricerca

[10] Sito di *PANDORA*, http://www.pandora.nla.gov.au/historyachievements.html (visitato il 28/10/2019).

si occupa di promuovere l'uso degli standard, incoraggiare le biblioteche e gli archivi ad adoperare strumenti e tecniche di archiviazione, e consentire la raccolta di un corpus di contenuti Internet da tutto il mondo in modo da poterli archiviare, proteggere e accedere nel tempo[11].

I tre principali progetti attualmente attivi sono:

- la costruzione di collezioni collaborative, ovvero la creazione di raccolte di archivi Web pubblici, basate su temi transnazionali o eventi di reciproco interesse. Gli argomenti sui quali si sta lavorando includono la crisi europea dei rifugiati, la cooperazione internazionale, le Olimpiadi e la commemorazione della Prima guerra mondiale;
- *Memento*, sono metadati utilizzati degli archivi IIPC;
- una mailing list elettronica aperta a chiunque sia interessato ai problemi associati alla raccolta di siti Web, all'archiviazione e ai problemi di manutenzione della qualità[12].

4.4 Il *Digital Preservation System* dell'Unione Europea

Il *Digital Preservation System (*DPS) è un progetto iniziato a settembre del 2016, promosso dagli Archivi Storici dell'Unione Europea (*Historical Archives of European Union* – HAEU), con lo scopo di raccogliere, archiviare, conservare e rendere accessibili i contenuti digitali e materiali derivanti da archivi privati di eminenti politici o movimenti e associazioni europee. I contenuti e i materiali in questione devono essere stati prodotti dalle Istituzioni dell'Unione Europea o devono essere rilevanti per la storia dell'UE.

La prospettiva della preservazione a lungo termine viene attuata impiegando due diverse politiche. La prima è che gli strumenti e le tecnologie si appoggino sui criteri e le raccomandazioni più rilevanti a livello mondiale tra le quali OAIS, PREMIS e ISO 16363 poiché definiscono i ruoli e l'architettura del sistema e per allinearsi con le direttive internazionali.

La seconda politica riguarda la sicurezza e la privacy. Il DPS pone particolare attenzione a questa tematica adottando sistemi di firewall, sicurezza perimetrale, password, autenticazione

[11] Sito dell'*IIPC*, http://netpreserve.org/ (visitato il 28/10/19).

[12] Wikipedia, voce *International Internet Preservation Consortium*, https://en.wikipedia.org/wiki/International_Internet_Preservation_Consortium (visitato il 26/10/19).

e autorizzazione dell'utente, backup, copia in remoto e sistemi di recupero. In questo senso è una delle poche organizzazioni in ambito della preservazione digitale che esplica in modo trasparente l'attenzione posta verso la sicurezza e la tutela di chi effettua l'accesso.

Il progetto DPS è realizzato dallo sforzo congiunto dell'HAEU e del gruppo italiano di tecnologia dell'informazione, *Dedagroup*. L'approccio metodologico si basa su due versioni principali che vanno di pari passo con le fasi del progetto. Questo approccio assicura un'adeguata gestione della trasformazione dei processi dell'HAEU, compresi i temi dell'istruzione e della formazione, e limita i rischi associati all'introduzione di nuove tecnologie dell'informazione[13].

4.5 Il Regno Unito e la legge sul deposito legale di siti Web

Nel Regno Unito, a partire dal 6 aprile 2013 è in vigore la *Non-print Legal Deposit Regulation*, una legge che ha esteso l'obbligo di deposito anche alle pubblicazioni non a stampa, e dunque includendo il digitale e l'online. In realtà, la costruzione di un deposito di siti, veniva già effettuato da più di un decennio, tramite accordi con i proprietari, portando alla costruzione di un *open archive* di 72.000 *snapshots*. La legge ha permesso di superare le difficoltà connesse allo stringere accordi preventivi con i responsabili dei siti.

La conservazione è garantita grazie a una collaborazione tra istituzioni governative e accademiche, e viene effettuata in quattro copie: presso la *British Library* nelle sue due sedi di Londra e Boston, la *National Library of Scotland* e la *National Library of Wales*.

È possibile inoltre effettuare la consultazione presso tre sedi universitarie: la *Bodleian Library* di Oxford, la *University Library* di Cambridge e la biblioteca del *Trinity College* di Dublino. Il Regno Unito è uno dei pochi casi di *Web archiving* dove si adottano tre diversi tipi di approcci:

- *Whole of domain,* si archiviano tutti i siti che ricadono sotto uno specifico dominio nazionale e quindi si basa su un criterio geografico;

[13] Sito dell'EUI, https://www.eui.eu/Research/HistoricalArchivesOfEU/FindingAidsAndResearch/Digital-Preservation-System (visitato il 28/10/19)

- *Archiviazione selettiva*, si sceglie cosa archiviare e cosa non archiviare, generalmente questa scelta viene fatta da organi appositi;
- *Archiviazione tematica*, si concentra sulla raccolta di siti che riguardano uno specifico argomento. La scelta viene fatta sulla base di quello che viene considerato socio-culturalmente rilevante.

Nel Regno Unito, quindi, la raccolta di siti Web riguarda tutto quello che ricade sotto il dominio .uk (circa 10.000.000 siti); inoltre viene raccolto il materiale che non ricade nel dominio .uk, ma è pubblicato e reso disponibile da persone o enti inglesi nei propri siti, e tutto ciò che viene pubblicato su domini stranieri che riguarda eventi importanti per la Nazione (personaggi influenti, informazioni da agenzie di stampa, scelte economiche estere, eventi culturali e così via). Il materiale raccolto è poi organizzato su base tematica o temporale.

Per i siti con dominio .uk la scansione viene effettuata in modo automatico mentre per gli altri si procede verificando la rilevanza; nel 2014 sono stati scaricati 2.500.000 siti non .uk.

A partire dall'approvazione della legge il numero di siti raccolti è andato crescendo, nel 2015 sono stati archiviati 57 TB, di cui 2,5 TB provenienti da siti non .uk. Per quanto riguarda la selezione e conservazione dei social media ci sono alcuni problemi particolari, dovuti all'enorme numero di pagine da salvare alla tutela della privacy e dei diritti.

Come in Francia, l'accesso ai contenuti archiviati è limitato ed è possibile solo nelle sale di consultazione delle sette biblioteche titolari del *Legal Deposit*. L'accesso è ristretto per motivi legati ai diritti degli editori e anche ai diritti dei cittadini a essere dimenticati, il così detto diritto all'oblio[14].

4.6 Il caso Wikipedia

Wikipedia è un'enciclopedia libera che si basa sulla collaborazione e cooperazione degli utenti che possono creare una pagina, modificarla o usufruire delle informazioni e del contenuto. È il settimo sito a livello mondiale per numero di accessi e costituisce un caso particolare e unico

[14] Bergamin Giovanni, Augusto Cherchi e M. Alessandra Panzanelli Fratoni. «Archiviare la rete: strumenti e servizi - PDF. https://docplayer.it/42899979-Archiviare-la-rete-strumenti-e-servizi.html (visitato il 9/11/2019).

di archiviazione Web. La sua organizzazione infatti prevede un "auto-archiviazione" delle pagine create, ciascuna caratterizzata da un identificatore denominato *Oldid.* Utilizzando questo identificatore è possibile accedere a tutte le versioni precedenti di una specifica pagina, mentre il *permanent link* porta sempre all'ultima versione della pagina.
Per esempio al link https://it.wikipedia.org/w/index.phptitle=Informatica_umanistica&oldid=9863804 possiamo vedere la prima pagina Wikipedia alla voce 'Informatica umanistica', creata il 14 luglio 2007; mentre al link permanente https://it.wikipedia.org/wiki/Informatica_umanistica si vede sempre la versione corrente che può cambiare nel tempo.

4.7 Limiti e assenza dell'Italia nel panorama della conservazione Web

Quelle illustrate in precedenza sono solo alcune delle più rilevanti iniziative a livello mondiale di *Web archiving*, alcune basate sul salvataggio delle pagine che corrispondono a un dominio nazionale e altre concentrate sulla rilevanza delle informazioni presenti nei siti Web e al rispetto del deposito legale.

In Italia, negli ultimi anni, grazie all'attenzione posta in merito all'argomento, è stata emanata la legge 106/2004 sul deposito legale esteso ai documenti diffusi tramite rete informatica. Le due novità più rilevanti introdotte per rendere possibili il raggiungimento di questi obiettivi sono: l'aumento delle tipologie di documenti oggetto di deposito legale e il deposito di questi documenti a più biblioteche su base regionale.

D'altro canto, il deposito legale dei documenti diffusi tramite rete informatica non è obbligatorio perché lo stesso D.P.R. 252/2006 all'art. 37 prevedeva la redazione di un regolamento tecnico che, dopo 13 anni, non è ancora stato emanato. Inoltre, nella legge è presente anche la "clausola costo zero" ovvero la possibilità di evitare la raccolta se comporta delle spese eccessive a carico dello Stato.

Una maggiore attenzione viene data alla conservazione di documenti e risorse prodotte dalla Pubblica Amministrazione, come stabilito dal Codice dell'Amministrazione Digitale. A tal proposito l'11 marzo 2019 è stato presentato il Piano Triennale per l'informatica nella Pubblica Amministrazione, con lo scopo di creare una strategia condivisa per una trasformazione digitale del Paese. Nel piano vengono definite le linee guida riguardo alcune tematiche come la sicurezza in Internet, creazione di piattaforme, governare la trasformazione digitale, occuparsi della spesa e del bilancio, etc.

Scarsa attenzione viene posta, invece, ai documenti digitali di interesse culturale. L'assenza di un quadro normativo specifico porta a una forte limitazione per quanto riguarda lo stanziamento di risorse, umane ed economiche, per far fronte alle necessità di gestione e tali da garantire uno sforzo adeguato verso una conversione al digitale.

Tuttavia, a partire dal 2006, la Biblioteca Nazionale Centrale di Firenze, in collaborazione con la Biblioteca Nazionale Centrale di Roma e la Biblioteca Nazionale Marciana di Venezia, ha attivato un progetto per la realizzazione di un'infrastruttura condivisa chiamata "Magazzini Digitali". In dieci anni sono stati salvate 96.000 tesi di dottorato (depositate presso le due Biblioteche Nazionali Centrali di Roma e Firenze), 40.000 articoli di riviste ad accesso aperto e circa 500 e-book. Inoltre, nel 2006, è stato fatto un esperimento di *Web archiving* sull'intero dominio ".it", grazie al quale sono stati raccolti 6 TB di dati[15].

In Italia la Biblioteca Nazionale di Firenze sta inoltre partecipando al progetto IIPC (vedi paragrafo 4.3) e dal 2018 ha attivato un nuovo progetto di *Web archiving*. Qualsiasi ente o associazione, che si occupa di siti Web di natura culturale, può partecipare e collaborare. I criteri di selezione tengono conto di quanto stabilito nel D.P.R 252/2006 e tengono anche conto della privacy e delle limitazioni imposte dai responsabili dei siti.

Infatti, dopo aver selezionato quello che viene ritenuto rilevante, si attua la fase di richiesta per poter poi effettivamente procedere alla raccolta: in mancanza di una normativa di riferimento è necessario chiedere un esplicito permesso ai gestori dei siti, rallentando e sovraccaricando il lavoro gestionale. Inoltre, non potendo richiedere ai programmatori delle specificità riguardo la pubblicazione dei siti, si hanno delle forti limitazioni per quanto riguarda l'usabilità e la qualità della preservazione e dell'accesso.

In confronto ad altri paesi europei in Italia il *Web archiving* è più una sfida culturale che tecnologica (Storti, 2018). Manca consapevolezza da parte degli utenti della rete ma c'è anche un vuoto a livello legislativo per quanto riguarda la gestione dei diritti e delle responsabilità. In assenza di una copertura legislativa delle reali necessità non possono essere finanziati progetti

[15] FPA. *Web archiving "sfida culturale": il servizio della Biblioteca Nazionale Centrale di Firenze*, giugno 2019. https://www.forumpa.it/pa-digitale/gestione-documentale/Web-archiving-sfida-culturale-il-servizio-della-biblioteca-nazionale-centrale-di-firenze/ (visitato il 9/9/19)

e fondi, e non esistono corsi che formino delle figure professionali in grado di assolvere alle competenze richieste in questo ambito.

5. Problematiche del *Web archiving*: dalle dimensioni del Web ai diritti dell'individuo

I problemi da affrontare per quanto riguarda il salvataggio del Web sono di quattro tipi principali: culturale, tecnico, economico e legale (Lyman, 2002).

Livello culturale. Il ritmo del cambiamento tecnologico rende difficile la conservazione dei media digitali. Tutti i documenti seguono un ciclo di vita da prezioso a obsoleto ma poi alcuni diventano storicamente importanti. Il lavoro degli archivisti consiste nel decidere quanto salvare, cosa e come.

Livello tecnico. Ogni nuova tecnologia impiega alcune generazioni per diventare stabile e per affermarsi o scomparire. Dal momento che questo avviene molto velocemente non pensiamo di conservare hardware e software necessari per mostrare una risorsa nata in quel preciso contesto. Gli archivi devono risolvere questo problema raccogliendo continuamente quanto serve e mantenendo le connessioni e i rapporti tra risorse.

Livello economico. È ancora da definire chi abbia la responsabilità di preservare il contenuto di Internet. Un archivio Web richiede un iniziale investimento di grandi dimensioni per la tecnologia, la ricerca, lo sviluppo e la formazione e, se necessario, deve essere realizzato su larga scala per consentire il salvataggio di grosse porzioni del Web.

Livello legale. Sebbene il Web sia popolarmente considerato una risorsa di dominio pubblico, in realtà è coperto da copyright. Recentemente le leggi sulla proprietà intellettuale sono state in parte estese anche ai documenti digitali ed è pertanto necessario stabilire accordi per usufruire dei materiali e delle informazioni.

5.1 Le dimensioni elevate del Web: cosa salvare e perché

Il Web è il documento più grande che sia mai stato scritto: si trova distribuito negli hard disk di 1.724.395.750 di siti Web[16] che contengono testo, immagini, video e grafici. Questi dati non considerano le dimensioni del *deep Web*, ovvero le risorse informative non indicizzate dai

[16] Statistiche prese da https://www.Internetlivestats.com/ (visitato il 28/10/2019).

normali motori di ricerca, che non sono misurabili ma sono stimate intorno ai 550 miliardi di documenti.

Il Web è scritto in 220 lingue (il 78% dei contenuti sono scritti in inglese) da autori provenienti da tutto il mondo. Va notato che il WWW ha solo 20 anni, e le modifiche che sta esercitando a livello economico, sociale, intellettuale sono solo all'inizio. Le sue dimensioni crescono molto velocemente, 7 milioni di nuovi siti Web al giorno, e allo stesso tempo molti altri spariscono. Infatti, la vita media di una pagina Web è solo 44 giorni, e gli utenti si rendono conto di questo solo quando cercando una pagina attraverso l'URL vedono scritto "404–Site Not Found".

La possibilità, mai avuta in nessuna epoca storica, di avere così tante informazioni insieme e, teoricamente, di poter comunicare con chiunque in qualsiasi parte del mondo, sta avendo un impatto sui nostri valori e sulle nostre vite i cui effetti saranno visibili solo a lungo termine. Il concetto stesso di memoria ha avuto una modifica radicale: da limitata, umana, da capacità da esercitare, a veloce, illimitata, a nostra completa disposizione.

L'eccesso di notizie, di conoscenza, ha però provocato una perdita del valore di quanto viene comunicato e inoltre il passaggio da analogico a digitale ha portato con sé il problema della perdita. Se da una parte su un e-book io posso avere anche 10.000 libri, dall'altro basta un click per eliminarli del tutto[17].

Nel processo di *Web archiving*, dove non è fattibile archiviare tutti i contenuti presenti sul Web, stabilire cosa consegnare all'oblio e cosa no, diventa un'operazione delicata e culturalmente rilevante. Possono essere utilizzati criteri dei tipo politico-economico (si salvano i giornali, gli articoli di finanza, etc.), culturale (riviste scientifiche, siti di musei, gallerie virtuali) od ogni Nazione si occupa di preservare quello che ritiene più importante che sta sotto il suo dominio (.it, .uk, e così via).

5.2 L'obsolescenza tecnologica e l'importanza degli standard

La Treccani definisce l'obsolescenza tecnologica come "*la condizione che rende non leggibile e non intelligibile e, quindi, inutilizzabile una risorsa digitale, a seguito dell'indisponibilità dei*

[17] *Costantino Landino - Web archivingWeb archivingWeb archiving,* https://www.youtube.com/watch?v=csBMR-y3b3Y (visitato il 28/9/19).

supporti o degli strumenti di lettura e di trattamento dei dati."[18]. Questo fenomeno può riguardare tre componenti diverse: i supporti fisici per la memorizzazione dei bit, gli hardware che assicurano la lettura del supporto fisico e i software che permettono la fruizione della risorsa digitale.

L'obsolescenza tecnologica costituisce un grosso limite alla preservazione a lungo termine, basta pensare al passaggio dalle schede perforate ai floppy disk, poi dai CD ROM alla pennina USB ed infine ai servizi di *Cloud* e alla quantità di dati andata persa. Gli spazi di archiviazione in rete, oggi largamente utilizzati, non danno la certezza che i documenti che vi sono depositati al loro interno siano accessibili tra 20 anni.

Le informazioni ottenute tramite la regolare raccolta di siti Web possono fornire alcune indicazioni sui file che stanno diventando non più accessibili. Quando il numero di file di un determinato formato inizia a diminuire, è il segno che il formato non è più in uso e sarà presto obsoleto. A quel punto è necessario compiere uno sforzo per garantire l'ulteriore disponibilità alla risorsa, traslandola su un altro supporto.

Un esempio di buona gestione dell'obsolescenza tecnologica è quella che riguarda l'Archivio Storico Multimediale del Mediterraneo. Il progetto, realizzato tra il 2006 e il 2009, consisteva in una banca dati di una vasta mole di documenti e collezioni cartografiche, conservati in diversi Archivi di Stato italiani. Il materiale, che copriva tutto il XV secolo, comprendeva atti notarili, pergamene, documenti degli Stati preunitari contenenti informazioni riguardo ai rapporti tra i Paesi del Mediterraneo. La fruizione da parte del pubblico di questo corpus è stata possibile tramite un portale che nel 2015 è stato chiuso a causa dell'obsolescenza tecnica del sistema. L'Istituto Centrale per gli archivi ha dunque traslato il contenuto dentro una piattaforma open source per la gestione dei beni culturali MetaFAD[19].

L'applicazione di uno standard costituisce lo strumento fondamentale per abbattere il problema dell'obsolescenza, unendo anche informazioni relative alla *policy*, che definiscano

[18] In "Enciclopedia Italiana" voce *obsolescenza digitale*, http://www.treccani.it//enciclopedia/obsolescenza-digitale_(Enciclopedia-Italiana) (visitato il 28/10/19).

[19] Istituto Centrale Per gli Archivi (ICAR). *Archivio Storico Multimediale del Mediterraneo: un progetto per il recupero e il rilancio*, http://www.icar.beniculturali.it/index.php?id=384 (visitato il 28/10/19).

esplicitamente responsabilità, ruoli e indicazioni sulle metodologie da adottare per garantire la sopravvivenza delle memorie, in modo da abbattere i costi di manutenzione.

5.3 Aspetti legali e limitazione all'accesso degli archivi Web

5.3.1 Copyright e diritti d'autore

Un aspetto fondamentale per quanto riguarda l'archiviazione Web è quello dei diritti. Il diritto alla privacy, il copyright e il diritto all'oblio limitano fortemente l'azione dei *crawler* e riducono la possibilità di accedere alle risorse archiviate. A meno che non venga stabilito a livello nazionale un protocollo di archiviazione che estenda la legge del deposito legale anche ai contenuti digitali, i responsabili devono confrontarsi direttamente con gli autori della risorsa, procedura inattuabile quando si tratta di collezioni estese.

Bisogna tener presente che la raccolta tramite *crawler* costituisce una copia e in quanto tale è vietata dalla normativa sul diritto d'autore. Solo l'autore ha infatti «il diritto esclusivo di autorizzare o vietare la riproduzione diretta o indiretta, temporanea o permanente, in qualunque modo o forma, in tutto o in parte» (art. 2 Direttiva 2001/29 CE del 22 maggio 2001)[20]. La normativa comunitaria considera come eccezioni quei casi in cui non ci sia un guadagno economico, se il processo è temporaneo e quando l'azione costituisce una parte essenziale e imprescindibile della trasmissione nella rete, sempre nel rispetto di un uso legittimo. Questo è il caso, per esempio, dei motori di ricerca.

Per i proprietari di un sito Web è possibile definire all'interno di un file apposito, il robot.txt, le regole per l'accesso, vietando la raccolta per alcune o tutte le pagine. Queste informazioni possono anche essere inserite in opportuni campi HTML contenenti meta-informazioni sulla pagina stessa (campi META). I *crawler* sono tenuti a seguire le regole definite nei file; quello *dell'Internet Archive*, per esempio, rispetta quando definito all'interno del sito. Nonostante questa procedura possa essere considerata un buon rimedio per il rispetto dei diritti d'autore, ciò non toglie che il detentore dei diritti non è tenuto a inserire un file robot.txt nei propri siti e la sua assenza non può essere interpretata come un segno di assenso alla raccolta.

[20] Bergamin, Giovanni, Augusto Cherchi e M. Alessandra Panzanelli Fratoni. *Archiviare la rete: strumenti e servizi - PDF*, https://docplayer.it/42899979-Archiviare-la-rete-strumenti-e-servizi.html (visitato il 2/11/19).

Un altro aspetto da considerare è quello della limitazione all'accesso. A seconda delle regole vigenti in un determinato Stato, sempre per motivi legati ai diritti, la consultazione del materiale archiviato può essere limitata. Alcune organizzazioni, come l'*Internet Archive* e la *Library of Congress* americana, non presentano restrizioni di alcun genere; mentre l'archivio Web della Norvegia segue delle regole molto severe. In alcuni casi viene consentito solamente l'accesso in loco, in altri è possibile consultare solo determinati contenuti, in altri ancora la lettura è concessa solo ai ricercatori e agli studiosi.

La questione del copyright e dei diritti diventa poi particolarmente delicata quando si tratta di social network e piattaforme online, per i contenuti personali che li caratterizzano. Bisogna considerare il diritto alla privacy per quello che viene pubblicato, il diritto d'autore e la proprietà intellettuale sui contenuti. Gestire migliaia e migliaia di contributi diventa una procedura estremamente complessa e praticamente impraticabile.

5.3.2 Il diritto all'oblio

Il diritto all'oblio è il diritto che riguarda la non diffusioni di informazioni che possono ledere l'onore di una persona, in particolare se si tratta di precedenti giudiziari. Infatti, un individuo accusato, se il nome è presente in rete, rimane marchiato anche se poi verrà assolto o scagionato; inoltre bisogna ricordare che una persona è innocente fino alla condanna definitiva. In Italia, inoltre, la legge impedisce di fare il nome di chi subisce violenza sessuale, dei collaboratori di giustizia e dei congiunti di persone coinvolte in fatti di cronaca.

La corte europea ha riconosciuto il diritto di un singolo di eliminare le tracce della propria esistenza nel Web. Questo diritto, nonostante tutto, è in conflitto con altri due: "il diritto di essere ricordato" e "il diritto all'informazione". Infatti, un cittadino di una società democratica deve poter accedere nel modo più ampio e libero possibile alla conoscenza e al sapere, sfruttando tutti i mezzi messi a disposizione.

Il diritto all'oblio non è riconosciuto in tutti i Paesi, nel Regno Unito vige la *Data Protection Act*, che sposta la responsabilità sul richiedente che deve dimostrare l'onere della prova. Questo significa che chi vuol far valere in giudizio un diritto deve dimostrare i fatti costitutivi, quelli che ne hanno determinato l'origine, ovvero chi vuole appellarsi al diritto all'oblio deve dimostrare che è vittima di sofferenza o di danni conseguenti alla diffusione di notizie che lo riguardano. In Italia ci si può appellare a questo diritto se viene a mancare uno dei principi della

corretta informazione: verità, continenza (linguaggio appropriato) e pubblico interesse (utilità di quanto viene comunicato).

Nel 2018 risultava che in tre anni e mezzo di diritto all'oblio quasi 400mila cittadini europei si sono rivolti a Google per pretendere la rimozione dal motore di ricerca di quasi 2,4 milioni di pagine che riguardavano loro stessi, la loro azienda o persone decedute. Sotto esame articoli di testate giornalistiche, documentazioni varie e, soprattutto, pagine di social network. Google ha analizzato ogni richiesta e ha provveduto, se la domanda avesse fondamento, alla *delisting* dal motore di ricerca. La *delisting* non è una cancellazione dal Web ma la perdita di visibilità su un motore di ricerca[21].

La sentenza pioniera per quanto riguarda il diritto all'oblio e il Web è quella C-131/12 del 2014. Al signor Mario Costeja González, spagnolo, era stata pignorata la casa e alcuni quotidiani avevano pubblicato la notizia. Dopo 15 anni, dato che digitando sui motori di ricerca il proprio nominativo apparivano ancora tali notizie, il signor Costeja Gonzales si è appellato al garante della privacy, affinché venisse rimosso dal quotidiano e da Google il suo nome. La decisione spettò alla Corte di Giustizia dell'Unione Europea che dovette considerare se effettivamente poteva dirsi di Google España la responsabilità della rimozione del link. Gonzalès vinse la causa e venne deindicizzato, mentre l'articolo rimarrà nel server sul quale originariamente è stato caricato[22].

5.3.3 Tre diversi approcci per la gestione dei diritti

Senza il permesso di archiviare non può esserci una corretta gestione del ciclo di vita del materiale. I diritti devono riguardare non solo l'acquisizione del materiale ma anche la possibilità di immagazzinare il materiale, fornire l'accesso ed eseguire gli interventi di conservazione al fine di garantire la preservazione a lungo termine. Esistono tre diversi tipi di

[21] IctBusiness. *Diritto All'oblio Rivendicato Su 2,4 Milioni Di Pagine Web e Social*, febbraio 2018, http://www.ictbusiness.it/cont/news/diritto-all-oblio-rivendicato-su-2-4-milioni-di-pagine-Web-e-social/41023/1.html#.XYH1FZMzbBI (visitato il 2/11/19).

[22] Il Post. *La sentenza sul diritto all'oblio e Google* in "Il post". 2014. http://www.ilpost.it/2014/05/13/sentenza-corte-europea-diritto-oblio-google/ (visitato il 4/11/19).

approcci per la gestione dei diritti dell'archiviazione Web, presentati nel *Digital Curation Manual* (Thompson, 2008):

- *Rights secured*, si basa sulla richiesta ai proprietari dei diritti. Questo è stato l'approccio adottato dall'archivio Pandora della *National Library of Australia* e dall'archivio UKWAC (*UK Web archiving Consortium*). Sicuramente permette di evitare problemi legali ma presenta diversi inconvenienti come lunghe tempistiche per la richiesta e la minore quantità di materiale archiviato.
- *Rights assumed*, si archivia il materiale senza richiedere permessi. L'organizzazione salva i siti Web e consente ai proprietari di richiedere la rimozione dei contenuti. Questo è il modello adottato dall'*Internet Archive*. Un archivio, nel caso in cui non rimuova quanto richiesto, può trovarsi di fronte a minacce legali. Tuttavia, se si vuole creare un archivio completo, contattare ogni detentore è un ostacolo notevole.
- *Mandated* o *legislated role*, il diritto di acquisire e rendere possibile l'accesso agli utenti è dato agli archivi e alle biblioteche come estensione del deposito legale. I diritti non vengono intaccati, ma ne subentrano di nuovi riguardo al salvataggio di siti e alla loro preservazione come previsto dalla legge.

6. Il formato standard ISO 28500:2009 (WARC) per la preservazione a lungo termine

6.1 Nascita e funzionalità del formato WARC

Circa un decennio fa i software utilizzati per "catturare" parti del Web (detti *crawler*, *spider* o *robot*) salvavano le informazioni utilizzando formati proprietari, codificandole in modo diverso. Ben presto è nata la necessità di definire uno standard che permettesse la cattura, la gestione e l'interscambio dei dati tra organizzazioni diverse. Il formato Web ARChive (WARC) soddisfa questo bisogno specificando un metodo che combina le risorse digitali e le relative informazioni all'interno di un unico file "contenitore".

In realtà questo formato costituisce una modifica del formato ARC creato nel 1996 da Brewster Kahle e Mike Burner, all'interno dell'*Internet Archive*, che serviva per archiviare il risultato delle operazioni di *crawling,* e poi successivamente adottato da istituzioni culturali. La ragione della modifica del formato ARC è che l'IIPC trovava sempre maggiori difficoltà nella gestione di dati a mano a mano crescenti e quindi avviò un progetto denominato "WARC file format", del quale si occupò un gruppo di lavoro dell'ISO.

Il progetto, iniziato nel 2005, si è sviluppato molto velocemente fino ad arrivare alla pubblicazione delle specifiche del formato. Nel maggio 2009 con la pubblicazione della norma ISO 28500:20094 "*Information and documentation- WARC file format*", viene ufficialmente riconosciuto come standard ISO. Il formato raccoglie i dati a prescindere dalla tipologia di risorsa e dal metodo di raccolta utilizzato, che può essere stata effettuata tramite il protocollo *Hypertext Transfer Protoco*l (HTTP), il *Domain Name System* (DNS) o il *File Transfer Protocol* (FTP).

Rispetto al formato ARC, WARC è più flessibile, aggiunge nuove funzionalità, memorizza le richieste HTTP, permette di inserire metadati arbitrari che uniscono una risorsa ad altre risorse archiviate, gestisce i duplicati, assegna un identificativo per ogni oggetto e permette la segmentazione dei contenuti raccolti su più record. I file WARCapplication/warc. hanno un loro identificatore di tipo MIME: application/warc.

6.2 Struttura di un file WARC

Essendo, il formato WARC, di tipo "contenitore" non ha bisogno di conoscere il tipo di oggetto che contiene. Un file WARC è costituito da più record di cui il primo è una descrizione degli altri. Ogni record è composto da un **text header** e un **content block**.

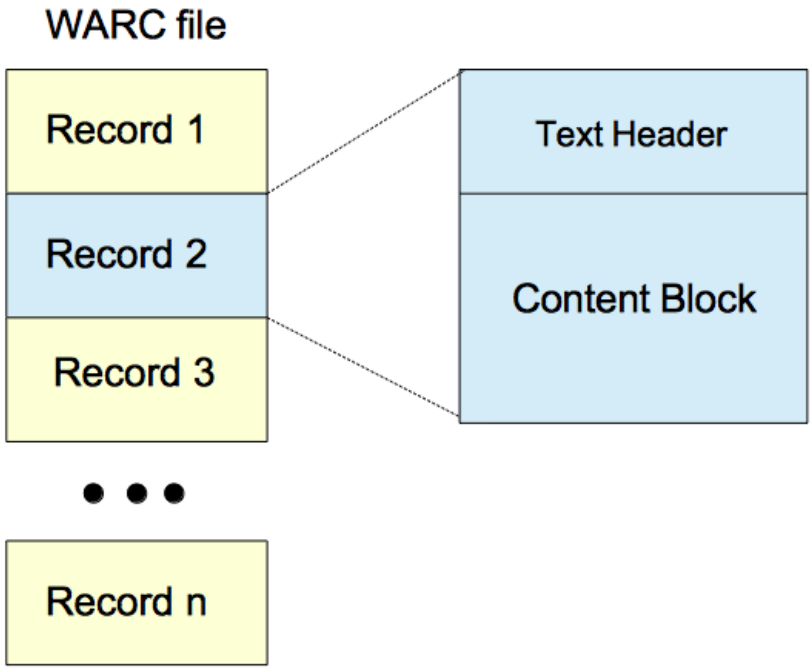


Figura 3 La struttura di un file WARC

La prima linea della text header dichiara la conformità a una versione (come "WARC/1.0), mentre le altre linee hanno la forma nome:valore ed indicano, per esempio, la data di cattura, l'URI della risorsa, il tipo di contenuto, etc. Una riga vuota separa la text header dal content block.

Il content block ha al proprio interno il contenuto vero e proprio ma può anche avere altri dati che forniscono informazioni ulteriori sugli oggetti archiviati (metadati). A seconda della tipologia di risorsa i record possono far parte di una delle seguenti categorie specificate nel campo WARC-Type:

- ***warcinfo***, descrive i record che lo seguono;
- ***response***, contiene una risposta a una richiesta, comprende il protocollo di rete;
- ***resourcer***, contiene una risorsa senza specificarne le informazioni relative al protocollo. Un esempio può essere un file recuperato direttamente da un repository accessibile localmente;
- ***request***, contiene i dettagli di una richiesta specifica. Il contenuto esatto di un record di "richiesta" è determinato non solo dal tipo di record ma anche dallo schema URI;
- ***metadata***, riguarda contenuti creati per descrivere, spiegare o accompagnare ulteriormente la risorsa raccolta per spiegarne aspetti non coperti da altri tipi di record. Un record di questo tipo farà quasi sempre riferimento a un record di altro tipo come quello che contiene l'originale o il contenuto trasformato;
- ***revisit***, si riferisce alla rivisitazione di un contenuto già archiviato. Il record "*revisit*" viene utilizzato al posto di un record "*response*" o "*resource*" per indicare che il contenuto visitato è un duplicato e non è l'originale;
- ***conversion***, deve contenere una versione alternativa del contenuto di un altro record creato come risultato di un processo di archiviazione. In genere, viene utilizzato per contenere le trasformazioni di un contenuto. Ogni trasformazione deve infatti tradursi in un record indipendente, completo, senza dipendenza dal record originale;
- ***continuation***, sono aggiunti ad altri file WARC per costruire il record originale; infatti la dimensione massima consigliata è 1 GB, se si supera questo limite il file viene spezzettato. Questo tipo di record contiene al suo interno il *Segment-Origin-ID* (identificatore della risorsa originale), il *WARC-Segment-Number* (numero che identifica i segmenti) e *WARC-Segment-Total-Length* (la lunghezza totale)[23].

[23] Landino, Costantino. *Strumenti per il Web: alcune soluzioni*. http://www.ilmondodegliarchivi.org/rubriche/archivi-digitali/650-strumenti-per-il-Web-archiving-alcune-soluzioni (visitato il 4/11/19).

```
WARC/1.1
WARC-Type: resource
WARC-Target-URI: file://var/www/htdoc/images/logoc.jpg
WARC-Date: 2016-09-30T16:40:32Z
WARC-Record-ID: <urn:uuid:23200706-de3e-3c61-a131-g65d7fd80cc1>
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:DBXHDRBXLF4OMUZ5DN4JJ2KFUAOB6VK8
WARC-Block-Digest: sha1:DBXHDRBXLF4OMUZ5DN4JJ2KFUAOB6VK8
Content-Length: 1662
 [image/jpeg binary data here]
```

Figura 3 Un esempio di un Text Header di una risorsa di tipo "resource"

6.3 Potenzialità e l'importanza di uno standard

Il formato WARC ha alcune caratteristiche che lo rendono adatto alla conservazione digitale. Infatti, non è proprietario, è aperto, è uno standard ed è trasparente. Nel particolare:

1. È stato sviluppato dall'*International Internet Preservation Consortium* (IIPC) e il gruppo di lavoro ISO si occupa del suo mantenimento. Non è sottoposto a limitazioni o brevetti;
2. Le sue specifiche sono disponibili e il formato è documentato;
3. È stato riconosciuto standard ISO ed è ampiamente utilizzato;
4. Può contenere di fatto qualsiasi tipo di formato e di oggetti digitali.

Oltre a queste caratteristiche se ne aggiunge un'altra molto importante, infatti il formato è auto-documentato (ogni risorsa è accompagnata dalle sue informazioni)[24].

Ad oggi quasi tutti i sistemi di *Web archiving* utilizzano questo formato per salvare le porzioni di Web scansionate. Ad esempio, *Heritrix*, il *crawler* utilizzato da *Internet Archive* e da numerose istituzioni culturali, adotta il WARC come formato predefinito per le sue "catture" (vedi capitolo 10).

Il formato WARC nonostante le sue caratteristiche e i suoi pregi non può essere considerato una soluzione definitiva al problema dell'obsolescenza digitale. WARC comunque è indubbiamente un notevole passo avanti verso la soluzione del problema dell'archiviazione,

[24] Dal sito della *Bibliothèque nationale de France*. *Information and documentation — WARC file format*, http://bibnum.bnf.fr/warc/WARC_ISO_28500_version1-1_latestdraft.pdf (visitato il 4/11/19).

consentendo la raccolta, la gestione e l'interscambio delle risorse "catturate" nel Web. Il fatto di poter ricondurre tutte le tipologie di formati a uno solo è un grande vantaggio.

Il suo riconoscimento come formato standard permette alle istituzioni che si occupano di *Web archiving* di svolgere il proprio dovere con efficacia e di scambiare tra di loro informazioni. Nel complesso il formato WARC alimenta la speranza di riuscire a conservare per le generazioni future le informazioni presenti sul Web.

7. Il *Web archiving* e il Laboratorio di Cultura Digitale dell'Università di Pisa

7.1 Necessità e problematiche del LabCD

Il Laboratorio di Cultura Digitale dell'Università di Pisa (LabCD) ha come scopo la promozione del patrimonio culturale e della sua trasmissione tramite i nuovi mezzi di comunicazione.

All'interno del laboratorio collaborano studiosi, docenti e alunni, che creano progetti digitali, in particolare nell'ambito della grafica virtuale, codifica di testi, creazione di siti Web *user friendly*, visualizzazione di dati, biblioteche e archivi digitali, piattaforme e-learning e così via. Molti di questi progetti includono anche lo sviluppo di siti Web, ospitati dal Laboratorio di Cultura Digitale.

Negli ultimi anni, come abbiamo visto (vedi cap. 3), l'attenzione verso l'archiviazione digitale è andata crescendo, e si è diffusa la consapevolezza della necessità di preservare quanto viene prodotto sul Web.

Per evitare di perdere definitivamente il materiale (le modifiche in corso, siti obsoleti, etc.), all'interno del Laboratorio di Cultura Digitale si è pensato di attuare le metodologie adottate da vari istituti e centri di ricerca per la conservazione del proprio lavoro e di quanto è stato fatto negli anni, tramite una soluzione di "*Web archiving"*. La soluzione completa dovrà prevedere sia l'aspetto di *crawling* e archiviazione di siti Web, sia la navigazione tra i siti archiviati e visualizzazione dei siti di interesse.

L'analisi delle esigenze del laboratorio è stata ottenuta conducendo alcuni colloqui con i responsabili e analizzando lo stato dell'arte riguardante l'acquisizione e il salvataggio di pagine Web.

Nel particolare, un progetto di questo genere deve:

- essere flessibile. Le pagine archiviate hanno caratteristiche e tipi di file al proprio interno molto differenti tra di loro (xml, pdf, html, css, js, etc);
- rispettare la *politeness*. Rispettare quanto stabilito dal file robot.txt per i diritti ed evitare di sovraccaricare il server;
- essere *user-friendly*. È fondamentale creare un'interfaccia quanto più possibile usabile e intuibile per evitare che l'utente abbia difficoltà nell'utilizzo degli strumenti;

- scalabile. Il sistema deve essere pronto ad adattare le proprie capacità alle richieste dell'utente;
- efficiente. La capacità di svolgere il lavoro e il rendimento devono essere alti.

7.2 Presentazione di una possibile soluzione

Presa coscienza delle due necessità, archiviare i contenuti online creati dal Laboratorio di Cultura Digitale e navigare e cercare tra i contenuti archiviati, è stato deciso di tenere separate le due funzionalità.

Così sono state create due piattaforme distinte: una di tipo server dedicata all'archiviazione vera e propria dei siti che contengono i progetti del LabCD, la seconda realizzata in WordPress per l'accesso ai contenuti archiviati (http://wayback.labcd.unipi.it/).

In questa prima fase del progetto la piattaforma di archiviazione è accessibile solamente agli amministratori. Sulla macchina sono installati un *crawler* (*Heritrix* versione 1.14.4), java versione 1.6.0_45, un visualizzatore di file WARC (il *Webrecorder Player* del https://Webrecorder.io messo a disposizione su *Github*) e una cartella-archivio contenente i file.

La seconda piattaforma, la Wayback Machine, permette agli utenti di accedere all'archivio dei siti Web salvati. Una volta individuato un progetto salvato che ci interessa visualizzare, è possibile, tramite il lettore e visualizzatore di file WARC, ripristinare le pagine salvate e navigarle come se fossero l'originale.

7.3 Scansione, salvataggio e accesso ai siti

Come già detto, possiamo distinguere due fasi fondamentali del progetto per il LabCD: la scansione e salvataggio dei siti e l'accesso ai siti salvati effettuato dall'utente.
La scansione del sito viene effettuata impiegando un *Web crawler*, *Heritrix*, che percorre i link interni alla pagina, salvandola con formato standard. I siti archiviati vengono automaticamente posti dentro una cartella del server, denominata *WARC*.

Per quanto riguarda il rispetto del copyright (vedi cap. 6), nel nostro caso, essendo i siti Web di proprietà del Laboratorio di Cultura Digitale, non abbiamo dovuto contattare i responsabili. Comunque, nelle impostazioni di *Heritrix*, abbiamo definito il rispetto del file robot.txt, che specifica quali pagine sono disponibili per i robot di *crawling*.

La parte dell'accesso è garantito dal sito Web *Wayback Machine*. Quest'ultimo, creato con WordPress, è composto da quattro sezioni principali: home, archivio, salva il tuo sito e contatti.

Nella home page è presente una descrizione sintetica di cosa sia il *Web archiving,* lo scopo del sito e il motivo per il quale l'archiviazione è importante. Vengono inoltre gli ultimi sei siti salvati.

Nella sezione *archivio* vengono mostrati sinteticamente tutte le pagine salvate, in ordine alfabetico. Per ogni sito salvato viene presentato uno *screenshot* della sua homepage per avere una visuale d'insieme, viene fornita una breve descrizione e il link per accedere al sito salvato. Inoltre, ad ogni sito è assegnata una categoria ed è contraddistinto da alcuni tag che permettono di fare una ricerca mirata. Nella presenta implementazione le categorie definite sono: tesi di laurea, progetti conclusi, progetti in corso, riviste e conferenze concluse.
La colonna di destra permette una navigazione tra i siti salvati. Dopo la lista degli ultimi sei siti salvati, è presente una casella di ricerca (alla Google) che cerca tra i nomi e le descrizioni di tutti i siti salvati; è presente anche una lista temporale (mese e anno) che permette di accedere ai siti salvati in quel mese; è presente infine una lista delle categorie e dei tag, che permette di accedere a tutti i siti in una data categoria o che hanno quello specifico tag.

La sezione *salva il tuo sito* è un *contact form* nel quale sono richieste alcune informazioni nel caso in cui l'utente sia interessato ad archiviare una pagina Web. La richiesta viene inoltrata al responsabile del progetto che provvederà a effettuare il processo di *crawling*.
All'interno del *contact form* sono indicati quei campi fondamentali al fine del salvataggio del sito:

- il nome del responsabile;
- l'indirizzo e-mail del richiedente;
- una breve descrizione;
- l'URL del sito da archiviare;
- il titolo del sito da archiviare;
- la descrizione del sito da archiviare;
- la categoria a cui assegnare quel sito;
- lo *screenshot* della pagina da salvare (facoltativo).

L'ultima voce del menu *contatti* raccoglie i nomi di coloro che hanno contribuito alla realizzazione del progetto e i diritti.

8. *Heritrix*: funzioni, caratteristiche e limiti

8.1 Software per l'acquisizione di siti Internet

Il metodo più comune di archiviazione utilizza *Web crawlers* che, lavorando in automatico, accedono ai siti a un livello di profondità determinato dal programmatore. Attualmente esistono numerosi strumenti che permettono di effettuare un processo di *harvesting*, alcuni altamente specialistici e altri adatti a un'archiviazione di tipo personale. I principali sono:

- *Heritrix*, (il sistema che è stato scelto) è il più diffuso a livello professionale. Il suo funzionamento verrà approfondito in seguito;
- *HTTRack*, applicazione *open source* che permette di lavorare sia da riga di comando che da interfaccia utente. Supporta HTTP e FTP ma non HTTPS e serve per riprodurre un sito in locale. In pratica una copia viene scaricata sul dispositivo per renderla navigabile; i link sono riorganizzati in modo da consentire l'accesso ai contenuti anche *offline*;
- *Archive-It*, è un servizio messo a disposizione da Internet Archive (vedi par. 4.1) e si basa sull'utilizzo di *Heritrix* e del formato WARC. I sottoscrittori del servizio possono utilizzare un'applicazione *user friendly* che permette anche ai non specialisti di archiviare le proprie pagine Web e di effettuare ricerche *full text* nei contenuti archiviati;
- *NetarchiveSuite*, progetto nato nel 2004, basato su *Heritrix* e impiegato per la raccolta del Web danese. Spazia in un ampio *range* di attività, dalla raccolta tematica limitata a porzioni ristrette di Web fino a quella che cattura un intero dominio (*whole of domain)*. I file sono salvati in formato WARC e vengono accompagnati da metadati estratti automaticamente;
- *SiteStory*, è un tipo di archiviazione transazionale, ovvero registra lo scambio di dati tra un Web server e un Web browser. Un protocollo che impiega questo software è Memento;
- *Social Feed Manager*, è un software open source che permette di raccogliere dati dai social network come Twitter, Tumblr, Flickr. Le immagini e le pagine Web sono catturate usando *Heritrix*; il risultato è memorizzato in un file WARC;

- *Webrecorder*, è un software gratuito che permette di archiviare i siti Web, mantenendone la dinamicità e l'interattività. Scaricando un'ulteriore applicazione è anche possibile navigare offline i siti. Anche in questo caso, il formato di archiviazione è il WARC[25].

8.2 La scelta del *crawler*

In questo paragrafo verrà confrontato il *crawler Heritrix* con quelli più utilizzati a livello internazionale, in particolare con *HTTRack*, *Archive.it* e *WebRecorder.io*.

Prima di tutto bisogna distinguere la staticizzazione, la copia statica in HTML di un sito, dalla procedura di archiviazione. La prima è utile nel caso in cui sia necessario effettuare il *backup,* per evitare problemi nel caso in cui ci siano inconvenienti, o per quelle pagine i cui proprietari non hanno intenzione di continuare a modificarle. La procedura di archiviazione, soprattutto volendo rispettare gli standard come nel nostro progetto, ha obiettivi diversi. Infatti, si vuole preservare il contenuto garantendone l'accesso a distanza di tempo e, per quanto riguarda le risorse digitali, può essere fatto utilizzando WARC.

Per questo motivo è stato scartato *HTTRack*, software che permette di effettuare una copia statica, ma non di salvare i siti utilizzando il formato standard.

Archive.it mette a disposizione una Web app a pagamento ed è basato su *Heritrix*. Non è possibile definire le proprie necessità e la raccolta deve rimanere sul server dell'Internet Archive, senza possibilità di scaricare le copie in locale.

WebRecorder.io, implementato in python, è un servizio disponibile online che permette di "registrare" la pagina Web, con un'ottima resa dei siti dinamici. In questo caso è possibile utilizzare il software online per archiviare ed è possibile navigare il sito sia online sia offline, grazie a un'applicazione desktop[26]. Quest'ultima viene impiegata nel progetto per visualizzare i file salvati con *Heritrix*. Registrandosi con WebRecorder.io, all'utente vengono concessi 5GB di spazio di archiviazione. Questo spazio di memoria non risulta sufficiente per gran parte dei siti Web del Laboratorio di Cultura Digitale (come il sito stesso del Laboratorio: http://www.labcd.unipi.it/).

---

[25] Landino, Costantino. *Strumenti per il Web archivingWeb archivingWeb archiving: alcune soluzioni,* http://www.ilmondodegliarchivi.org/rubriche/archivi-digitali/650-strumenti-per-il-Web-archiving-alcune-soluzioni (visitato il 4/11/19).

[26]Sito *Webrecorder.io,* https://Webrecorder.io/ (visitato il 4/11/19).

Alla luce di queste considerazioni, è stato deciso di usare *Heritrix* (il cui nome significa "erede"), che è un *Web crawler* open-source, estensibile e scritto in Java[27]. È attualmente il più diffuso e viene utilizzato in numerosi progetti nazionali. È un *crawler* di tipo *client-side* ovvero si comporta come un *client*, usando il protocollo HTTP per richiedere i contenuti direttamente dal server. Questa tecnica, nata per l'indicizzazione da parte dei motori di ricerca, è stata poi successivamente applicata ai bot. Si può interagire con il programma sia tramite l'interfaccia utente o da linea di comando. La memorizzazione dei file viene fatta utilizzando il formato WARC e non è possibile visualizzare direttamente il risultato, per il quale servono dei *tool* appositi.

8.3 Una panoramica sul funzionamento del tool

Heritrix gestisce il lavoro tramite i *job*, sui quali è possibile specificare i moduli, sotto-moduli impostazioni generali, sostituzioni e perfezionamenti. Può essere eseguito un solo job per volta. Inizialmente si può creare un lavoro sulla base di uno precedente o si può lavorare utilizzando un profilo, creato dall'utente o di default. All'interno della casella sottostante va inserito l'URI della risorsa da archiviare.

I **moduli**, una volta definiti, possono essere collegati tra di loro e si distinguono in tre grandi categorie: ambito (crawl scope), frontiera e processori.

- Un ambito di ricerca per indicizzazione (crawl scope) è un oggetto che decide per ciascun URI rilevato se rientra nell'ambito della ricerca per indicizzazione corrente, quindi definisce quali URI presenti nel sito saranno archiviati e quali no. Quest'impostazione serve per limitare o allargare la profondità di ricerca. Tra le diverse opzioni quelle più interessanti sono PathScope e BroadScope. La prima limita gli URI rilevati a una sezione di percorsi su host definiti dal primo che abbiamo inserito, per esempio se la radice che immettiamo è *'archive.org/example/'* tutti gli URI sotto il percorso *'example'* saranno scannerizzati (come *'archive.org/examples/hello.html').* Il secondo non impone alcun limite agli host, domini o percorsi URI sottoposti a scansione, rischiando di andare fuori dalla capacità di memoria.

[27]Sulla piattaforma *Github* alla voce *Heritrix3*, https://github.com/Internetarchive/heritrix3, visitato il 29/10/19.

- La frontiera mantiene lo stato interno della scansione. È il processo che tiene conto degli URI scoperti, quelli che si stanno processando, quelli processati e l'ordine con cui devono essere processati quelli in coda. Quella predefinita è BdbFrontier che controlla, rallenta o velocizza la processione di alcuni URI.
- Quando un URI viene sottoposto a scansione, viene in effetti passato attraverso una serie di processori. Questa serie è suddivisa per comodità in cinque fasi e l'utente può aggiungere, rimuovere e riordinare i processori su ciascuna di queste fasi. Nel caso in cui ci sia un errore o il responsabile decida di interrompere per qualche motivo il processo, l'elaborazione passa direttamente alla fine, alla fase finale di pulizia. Le cinque fasi sono:

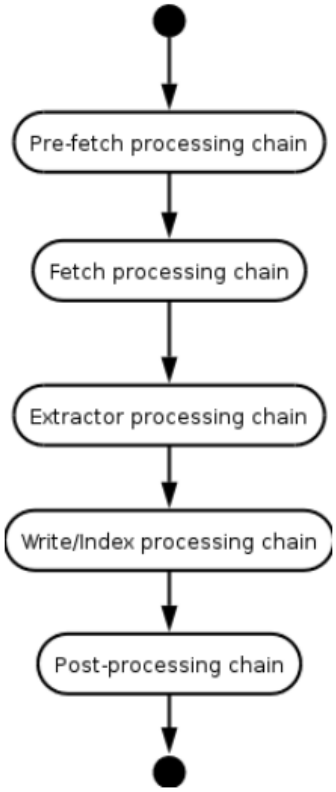
  Pre-fetch. Compie un controllo iniziale di verifica sugli URL da scansionare, controllando tutte le condizioni impostate e il robots.txt.

  Fetch. Vengono scaricati i dati dal server remoto.

  Extractor. Vengono estratti i link dei documenti presenti sul sito. È presente un processore per ogni tipo di documento scansionato: html, pdf, doc, swf, etc. Nelle impostazioni dei moduli possono essere aggiunti o tolti estrattori per specifiche tipologie di file.

  Write/Index. Si occupa della memorizzazione dei file scansionati, di default il formato utilizzato è il ARC.

  Post-processing. Vengono fatti controlli sui link estratti.



Nella scheda **sotto-moduli** è possibile andare a modificare dettagli di quanto si è specificato nella sezione precedente. Per esempio, possono essere impiegati per imporre dei filtri sugli URL scoperti (rifiutare se troppo lontano dalla radice, rifiutare di default i link trovati, etc.).

Un altro caso che può essere gestito grazie ai sotto-moduli è quello della trasformazione degli indirizzi. *Heritrix* mantiene un elenco di URL già visti e, prima di recuperarli, cerca nell'elenco di quelli trovati per vedere se l'URL è già stato sottoposto a scansione. Spesso un URL può essere scritto in più modi, ma la pagina recuperata è la stessa in ogni caso. Quindi il software attua una serie di passaggi di regolarizzazione per testare se un link è stato precedentemente trovato o meno e per capire se quella che sta archiviando è una copia della stessa pagina o dello stesso documento. L'utente può modificare queste trasformazioni.

Le altre impostazioni della scheda sotto-moduli sono relative alle cinque fasi viste precedentemente.

La scheda delle **impostazioni** presenta ulteriori alternative che è possibile selezionare. Qui sotto verranno presentate le più importanti per questioni di sinteticità e di semplicità.

Due sono i campi obbligatori da impostare sotto l'etichetta *http-headers*: user-agent e from. Nell'*user-agent* va inserito l'URL dell'organizzazione che si occupa del *crawling*, il campo ha la forma: Mozilla/5.0(compatible; /1.14.4 +http://www.labcd.unipi.it/).

Nel campo from va inserito l'indirizzo e-mail del responsabile.
Nella sezione *crawl-order* è possibile imporre dei limiti sulla durante e l'estensione del *crawling*, o imponendo il numero massimo di byte (max-byte-download) o interrompendo dopo aver scaricato un numero fisso di documenti (max-document-download) oppure dopo un certo numero di secondi trascorsi (max-tempo-sec).

Nella sezione *robots-honoring-policy* si può decidere se rispettare il robot.txt ed eventualmente fare alcune modifiche.

Le impostazioni della *frontier* riguardano l'imposizione di tempi di attesa tra la fine dell'elaborazione di un URI fino all'inizio di quello successivo.

Nella scheda delle **sostituzioni** si offre la possibilità di sostituire le singole impostazioni in base al dominio. Non sono state utilizzate per questo progetto.

L'ultima pagina è quella dei **perfezionamenti**. I perfezionamenti sono simili alle sostituzioni in quanto consentono all'utente di modificare le impostazioni in determinate circostanze. Vi sono tuttavia due differenze principali: i perfezionamenti vengono applicati in base a criteri arbitrari anziché al dominio URI rilevato ed è possibile impostare criteri in base all'ora del giorno, un'espressione regolare corrispondente all'URI e al numero di porta dell'URI. Anche in questo caso non se ne è fatto uso per il progetto[28].

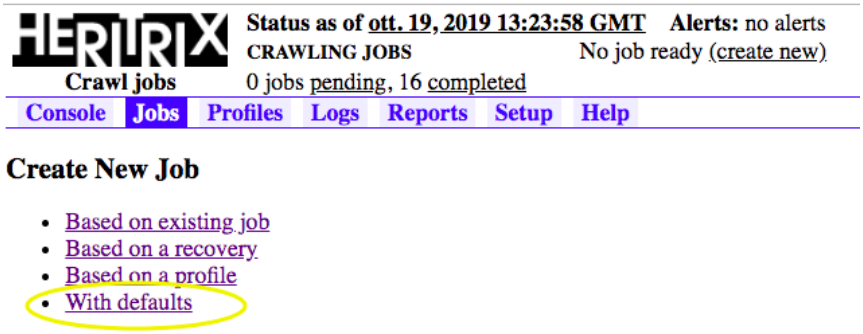
8.4 Un esempio di funzionamento del programma su un sito-progetto del LabCD

Per dimostrare l'utilizzo del programma e come è stato impiegato per fare *crawling* dei siti Web dei progetti del Laboratorio di Cultura Digitale sarà riportato di seguito un esempio.

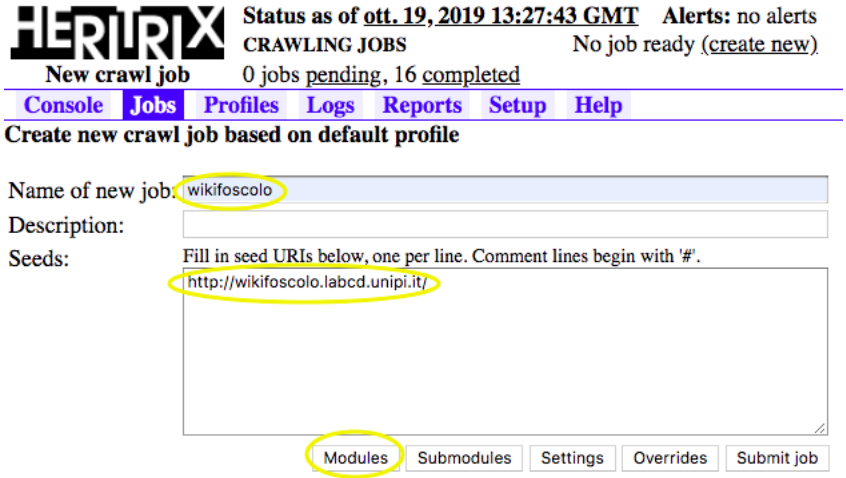
Per prima cosa consideriamo una pagina come http://wikifoscolo.labcd.unipi.it/, sito che ha come scopo promuovere progetti e ricerche sull'autore.

Creiamo un *job* con il profilo di default.

[28] *Heritrix User Manual*, http://www.crawler.archive.org/articles/user_manual/index.html (visitato il 4/11/19).

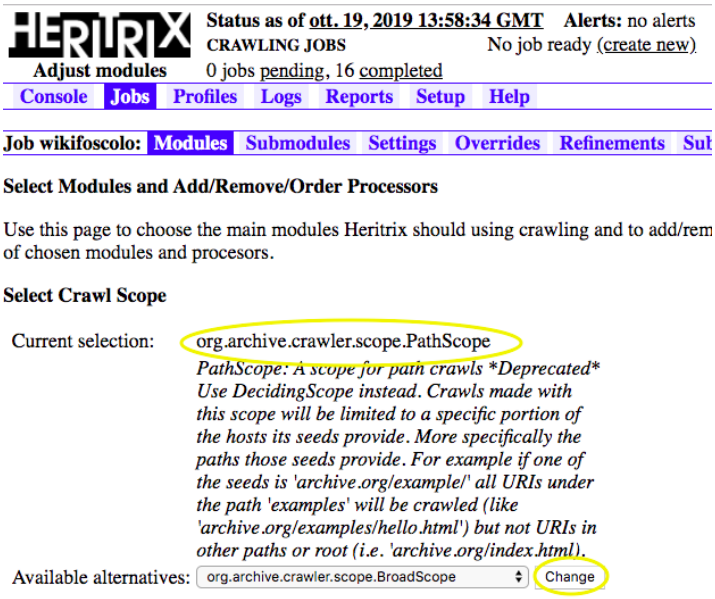


Inseriamo il nome del *job,* l'URI nella casella apposita e andiamo nella sezione *modules*.



A questo punto abbiamo davanti la schermata dove è possibile andare a modificare i moduli. Per prima cosa ci interessa modificare la profondità di ricerca del crawl.

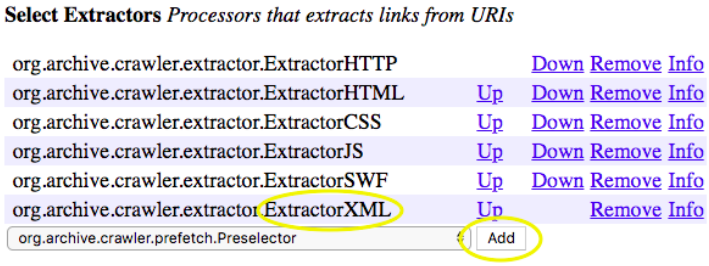
Nel nostro caso vogliamo che a partire dal *seed* http://wikifoscolo.labcd.unipi.it/ vengano scaricate tutte le pagine figlie come http://wikifoscolo.labcd.unipi.it/index.php/biografia/, quindi selezioniamo l'opzione che permette di fare questo, *Pathscope,* e clicchiamo su *change*.



Successivamente quello che ci interessa modificare rispetto alle opzioni di default sono gli estrattori di link. All'interno di una pagina Web possono infatti essere presenti file di formato diverso ed è necessario indicare estrattori specifici per ogni tipo.

Nel nostro caso guardando i file contenuti all'interno del sito possiamo osservare alcuni xml, dobbiamo quindi selezionare l'estrattore in grado di catturare questi file.

Scendiamo quindi fino a cercare tra i moduli la voce *Select extractors*. Nel menu a tendina selezioniamo quello che ci interessa e clicchiamo su *add*. Una volta aggiunto sarà nella lista sopra.



Ovviamente a seconda del sito che stiamo scansionando e a seconda di specifiche esigenze è possibile aggiungere più di un estrattore.

A questo punto ci possiamo spostare sulle impostazioni generali (**settings**) tramite il menu in alto o quello in basso. Andiamo ad inserire nei campi user-agent e from le informazioni obbligatorie.

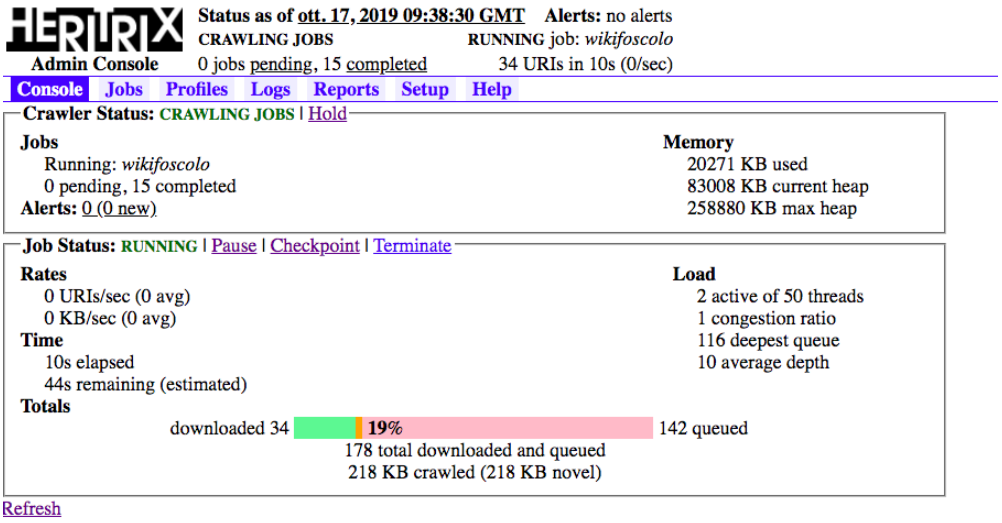
Nel primo campo andiamo a inserire l'indirizzo Web dell'ente responsabile del *crawling*, questo permette al Webmaster di avere informazioni sull'ente responsabile.

Nel secondo campo digitiamo l'indirizzo e-mail di chi si è occupato dell'archiviazione del sito.

A questo punto, definite le regole dell'operazione di crawl possiamo, dal menu in grassetto in alto o in basso, andare su ***Submit Job*** e avviarlo.

8.5 Analisi dei risultati e riflessioni sui limiti

Dalla pagina iniziale di *Heritrix*, su **Console**, si può monitorare l'andamento tramite una barra che mostra la percentuale dei link estratti e convertiti in WARC.



Heritrix fornisce un ulteriore strumento per controllare l'andamento del lavoro e la generazione di errori: i log, informazioni di monitoraggio che il programma scrive mentre procede. È possibile visualizzarle dal menu in alto, alla voce **Logs** ed inoltre vengono automaticamente inserite all'interno della cartella finale che contiene i file WARC.

I log sono di cinque tipi:

1. crawl.log. Per ogni URI estratto una riga riporta il successo o il fallimento del processo. Per ogni riga sono indicati in questo ordine: il *timestamp* in formato ISO8601, il codice di stato

(200, 404, etc.), la grandezza del file in bytes, l'URI del documento scaricato, il percorso del download, il tipo di documento, l'id di chi l'ha scaricato, la data, lo SHA1 del contenuto (funzione crittografica di *hash*), il source tag e infine le annotazioni.

2. local-errors.log. Sono gli errori che occorrono quando si prova a processare un URI, dovuti solitamente a problemi di rete. Possono essere ignoranti senza gravi danni.
3. progress-statistics.log. Sono scritti dal cosiddetto *StatisticsTracker,* che riporta le informazioni riguardo una URI processata. In particolare: la data, numero degli URI trovati, gli URI che stanno aspettando di essere processati, quelli scaricati, numero di documenti scaricati al secondo, kilobytes scaricati al secondo, numero di URI che *Heritrix* fallisce nel processare, numero di *thread* occupati nell'elaborazione e la quantità di memoria assegnata alla Java Virtual Machine.
4. runtime-errors.log. Rileva le eccezioni ed errori imprevisti che si verificano durante la raccolta. Alcuni possono essere dovuti a limitazioni hardware ma la maggior parte sono dovuti a un bug nel software.
5. uri-errors.log. Si trovano quando ci sono errori nelle URI[29].

Durante il processo di archiviazione è possibile che non tutte le risorse vengano completamente scaricate. Non è raro che navigando attraverso un sito processato alcune pagine restituiscano l'error 404 o alcune immagini non vengano visualizzate.

[29] Ibidem.

9. L’accesso e la gestione dei progetti a lungo termine

La costruzione del servizio costituito dalle due piattaforme non si limita al suo uso temporaneo e provvisorio, ma la volontà è quella di attivare uno strumento che sia utilizzato e garantisca la conservazione del sapere interno dell’Università.

Un interrogativo al quale dobbiamo provare a rispondere è la gestione del *tool* a lungo termine. Infatti, per quanto sia importante la nascita e l’ideazione di un progetto, la sua sopravvivenza dipende dal fatto che venga impiegato o meno e dalle modifiche che verranno apportare per supportare le necessità dei fruitori.

Nei seguenti paragrafi verranno affrontate alcune tematiche e possibili soluzioni.

9.1 Estensione del servizio di *Wayback Machine* ai dipartimenti dell’Ateneo

All’interno della *Wayback Machine* sono raccolti circa una ventina di progetti, tutti appartenenti al Laboratorio di Cultura Digitale, per il quale lo strumento è stato creato.

La volontà, però, è quello di estendere il servizio all’intero Ateneo pisano. In quest’ottica sarà necessario affrontare alcune questioni:

- uno spazio di memoria più grande;
- gestione di richieste multiple al server (scalabilità);
- suddivisione dei siti tramite categorie e tag appropriati;
- dotare il servizio delle opportune risorse.

Inoltre, bisognerà effettuare alcune modifiche sul sito della *Wayback Machine* per renderlo più adatto a consentire l’accesso a un numero maggiore di utenti.

9.2 Sviluppi futuri dell'archivio Web

Attualmente l’utente che è interessato ad archiviare un sito può farne richiesta tramite un apposito *contact form* presente all’interno della *Wayback Machine* del LabCD.

Un possibile sviluppo futuro è quello di creare un *self-preserving system*, ovvero dare la possibilità agli utenti di archiviare da soli un sito per poi accedervi successivamente.

Attualmente l'interfaccia della piattaforma di archiviazione, la *Webarchive*, è quella predefinita che mette a disposizione la GUI di *Heritrix*. L'idea è quella di costruirne una *user friendly*, più facile da impiegare e più intuitiva. Infatti, il software è molto complicato e potente e non sono necessarie tutte le funzionalità per estrarre un normale sito Web. La cosa migliore è quella di semplificare l'archiviazione, impostando di default tutti gli estrattori, rispettando la policy del *robot.txt* e richiedendo all'utente di inserire alcune informazioni. Tra queste sono fondamentali:

- l'URL della risorsa;
- un nome identificativo per il job;
- una breve descrizione del contenuto per la creazione dell'articolo su *Wordpress*;
- il nome del responsabile

Per quest'ultimo punto la modalità più efficiente di gestire l'accesso alla piattaforma è quella di permettere di utilizzare il servizio solo se l'utente è registrato e appartiene all'Università di Pisa. Il sistema di registrazione richiederà l'e-mail, il ruolo (docente, studente, etc.) e una password.

Inoltre, per quanto riguarda il copyright, si specifica che l'Università si solleva dalla responsabilità nel caso in cui vengano archiviate pagine Web coperte da diritti d'autore o i cui contenuti siano illegali. Inoltre, nel caso in cui il detentore dei diritti ne faccia richiesta, ci si riserva di eliminare totalmente il contenuto archiviato.

10. Problematiche incontrate e soluzioni adottate

Durante la realizzazione del progetto si sono presentati alcuni ostacoli i quali verranno descritti brevemente di seguito. Per una questione di chiarezza verranno distinti quelli affrontati per la piattaforma di archiviazione (Webarchive.labcd.unipi.it) da quelli della piattaforma di visualizzazione (wayback.labcd.unipi.it).

Per quanto riguarda la piattaforma di archiviazione uno dei punti che è stato affrontato è la mancanza di un lettore di file WARC messo a disposizione da *Heritrix*. Infatti, il *crawler* salva le pagine Web utilizzando il formato standard ma non prevede uno strumento per visualizzare quanto è stato archiviato. Per ovviare al problema si è utilizzato il lettore di file WARC di *Webrecorder* (vedi par. 9.2).

Bisogna considerare che *Heritrix* è un software specialistico e in quanto tale richiede una conoscenza approfondita dello stato dell'arte e delle sue impostazioni che possono risultare non banali per l'utente comune. Inoltre, un grosso limite del *crawler* è l'incapacità di processare più *job* alla volta e di scaricare automaticamente le eventuali modifiche dei contenuti archiviati. Infatti, nel caso in cui un sito venga cambiato in alcune sue parti, *Heritrix* lo ri-archivia nella sua interezza, senza duplicare il materiale che già possiede e processare solamente quello nuovo.

Per quanto riguarda la piattaforma di visualizzazione, la creazione di articoli da pubblicare sulla pagina Archivio per la consultazione delle risorse è stata fatta manualmente, compromesso accettabile dato che il materiale del Laboratorio di Cultura Digitale era poco numeroso. Nell'ottica di estendere il servizio agli altri dipartimenti dell'Ateneo questo procedimento andrà automatizzato, utilizzando come base le informazioni inserite dall'utente al momento della richiesta di *crawling* come specificato nel paragrafo precedente (par. 7.2)

11. Conclusioni

Questo elaborato ha cercato di chiarire cosa sia il *Web archiving* e la sua importanza per la preservazione dei contenuti digitali. Il servizio messo a disposizione dal Laboratorio di Cultura Digitale è un esempio concreto di come possano essere impiegati gli strumenti presenti online per la creazione di un archivio digitale, nell'ottica di salvaguardare la cultura presente sul Web. La soluzione implementata prevede l'utilizzo del *crawler open source Heritrix,* di un visualizzatore di file WARC e di un sito per l'accesso ai contenuti archiviati.

La scelta di tenere separati i *tool* per l'archiviazione e quello per la visualizzazione delle pagine Web, ha permesso di semplificare notevolmente lo strumento, permettendo la creazione di un'interfaccia utente intuitiva e di facile utilizzo.

Considerando le risorse messe a disposizione, il progetto risulta efficiente e adatto alle necessità del Laboratorio. Una volta garantiti una migliore scalabilità dello strumento ed effettuate delle modifiche per una navigazione ottimale anche nel caso di grosse quantità di pagine archiviate, il servizio sarà estendibile all'intero Ateneo.

Con il tempo, grazie alla *Wayback Machine*, sarà possibile fare una ricostruzione virtuale della storia del LabCD, dello sviluppo dei progetti, delle ricerche e il mutamento degli interessi nella comunità delle *digital humanities*.

Proprio la denominazione "Laboratorio di Cultura Digitale" porta dentro di sé la missione di raccolta della cultura digitalizzata, che ha come risultato pagine Web che accolgono gli studi in ambito culturale.
Citando una parte di presentazione del Laboratorio presente sul sito, descrivendo la propria missione, "*La nozione di Cultura Digitale comprende l'impegno verso la conservazione e trasmissione del patrimonio culturale tramite i nuovi media e lo studio sull'evoluzione dei contenuti, determinata dai nuovi mezzi di comunicazione.*"[30].

Il progetto di laurea si è concentrato proprio sulla conservazione dei contenuti e la piattaforma, con il trascorrere del tempo, permetterà di mettere in luce l'evoluzione del dialogo tra l'ambito umanistico e l'informatica.

[30] Sito del Laboratorio di Cultura Digitale, http://www.labcd.unipi.it/laboratorio/ (visitato il 4/11/19).

12. Bibliografia

Allegrezza, Stefano. «Requisiti e standard dei formati elettronici per la produzione di documenti informatici». 2010. http://www.comune.castelleone.cr.it/public/upload/file/UNIONE_GERUNDO/Formazione/dematerializzazione/Dispensa%20sui%20formati%20elettronici.pdf.

Bergamin, Giovanni, Augusto Cherchi e M. Alessandra Panzanelli Fratoni. «Archiviare la rete: strumenti e servizi - PDF». Consultato il 9 novembre 2019. https://docplayer.it/42899979-Archiviare-la-rete-strumenti-e-servizi.html.

Bnf. *Information and documentation — WARC file format*. Consultato il 4 novembre 2019. http://bibnum.bnf.fr/warc/WARC_ISO_28500_version1-1_latestdraft.pdf.

Brügger N. e Ralph Schroeder, a c. di. *The Web as History: Using Web Archives to Understand the Past and the Present*. UCL Press. 2017. https://doi.org/10.2307/j.ctt1mtz55k.

Clausen, Lars R.. «Handling File Formats». The State and University Library, Århus, Denmark The Royal Library, Copenhagen, Denmark. 2004. http://netarkivet.dk/wp-content/uploads/FileFormats-2004.pdf.

Friedlander, Amy C. . 2002. «Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving».

Lavoie. Brian F..«Il modello di riferimento per un Sistema informativo aperto per l'archiviazione».

Michetti, Giovanni. «Il modello OAIS». 2008. http://www.interpares.org/display_file.cfm?doc=ip3_italy_dissemination_jar_michetti_digitalia_3_2008.pdf.

Ockenden, Will. «HANDLING FILE FORMATS». The State And University Library, Universitetsparken.

Radcliffe-Brown, Alfred Reginald. *The Andaman Islanders*. Cambridge: Cambridge Univ. Press. 2013.

Sigurðsson, Kristinn. «Adaptive Revisiting with *Heritrix».* 2005.

Testoni, Laura. «Digital curation e content curation: due risposte alla complessità; dell'infosfera digitale che ci circonda, due sfide per i bibliotecari». 2013. https://www.aib.it/aib/sezioni/emr/bibtime/num-xvi-1/testoni.htm#nota10.

The consultative Commitee for an for Space Data System. «Reference Model for an Open Archival Information System (OAIS)». 2012.

Thompson, Dave. *DCC Digital Curation Manual: Instalment on Archiving Web Resources*. 2010. HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils. 2008. http://hdl.handle.net/1842/3375.

Toyoda, M., e M. Kitsuregawa. «The History of Web archiving». 2012. https://doi.org/10.1109/JPROC.2012.2189920.

Zanni Rosiello, Isabella. *Andare in archivio*. Orientamenti. Bologna: Il Mulino. 1996.

13. Sitografia

Bracciotti, Lorenzana. «Il Web archiving. Conservazione e uso di una nuova fonte – Officina Della Storia». https://www.officinadellastoria.eu/it/2019/01/10/il-Web-archiving-conservazione-e-uso-di-una-nuova-fonte/. Consultato il 13 settembre 2019.

Conservazione Digitale. «Il modello OAIS». http://www.conservazionedigitale.org/wp/materiale-didattico/il-modello-oais/. Consultato il 14 settembre 2019.

DCC. «What is digital curation? | Digital Curation Centre». http://www.dcc.ac.uk/digital-curation/what-digital-curation. Consultato il 10 ottobre 2019.

Enciclopedia Italiana, voce *obsolescenza digitale*, http://www.treccani.it//enciclopedia/obsolescenza-digitale_(Enciclopedia-Italiana). Consultato il 28 ottobre 2019.

European University Institute. «Digital Preservation System (DPS)». https://www.eui.eu/Research/HistoricalArchivesOfEU/FindingAidsAndResearch/Digital-Preservation-System.aspx. Consultato il 26 agosto 2019.

FPA. «Web archiving: il servizio della Biblioteca Nazionale Centrale di Firenze». *FPA* (blog). 12 giugno 2019. https://www.forumpa.it/pa-digitale/gestione-documentale/Web-archiving-sfida-culturale-il-servizio-della-biblioteca-nazionale-centrale-di-firenze/. Consultato il 9 ottobre 2019.

Github*,* voce *Heritrix3*, https://github.com/internetarchive/Heritrix3. Consultato il 29 ottobre 2019.

Github. «The WARC Format». https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/. Consultato il 16 settembre 2019.

«*Heritrix* User Manual». http://www.crawler.archive.org/articles/user_manual/index.html. Consultato il 4 novembre 2019.

IctBusiness. «Diritto All'oblio Rivendicato Su 2,4 Milioni Di Pagine Web e Social». http://www.ictbusiness.it/cont/news/diritto-all-oblio-rivendicato-su-2-4-milioni-di-pagine-Web-e-social/41023/1.html#.XYH1FZMzbBI. Consultato il 2 novembre 2019.

IIPC. «IIPC-Projects». http://netpreserve.org/projects/. Consultato il 26 agosto 2019.

Il Post. «La sentenza sul diritto all'oblio e Google». 2014. http://www.ilpost.it/2014/05/13/sentenza-corte-europea-diritto-oblio-google/. Consultato il 4 novembre 2019.

Internet Archive, https://archive.org/about/. Consultato il 28 ottobre 2019.

Internet World Stats, https://www.internetworldstats.com/stats.htm. Consultato il 26 ottobre 2019.

Istituto Centrale Per gli Archivi (ICAR). «Archivio Storico Multimediale del Mediterraneo: un progetto per il recupero e il rilancio». http://www.icar.beniculturali.it/index.php?id=384. Consultato il 28 settembre 2019.

Laboratorio di Cultura Digitale. http://www.labcd.unipi.it/laboratorio/. Consultato il 4 novembre 2019.

Landino, Costantino. *Costantino Landino - Web archiving*. https://www.youtube.com/watch?v=csBMR-y3b3Y. Consultato il 28 settembre 2019.

Landino, Costantino. (Istituto Centrale per gli Archivi). «Strumenti per il Web archiving: alcune soluzioni». http://www.ilmondodegliarchivi.org/rubriche/archivi-digitali/650-strumenti-per-il-Web-archiving-alcune-soluzioni. Consultato il 4 novembre 2019.

«lineeguidaperlaconservazionedeisiti_0.pdf». https://ict.sns.it/sites/default/files/lineeguidaperlaconservazionedeisiti_0.pdf. Consultato il 22 ottobre 2019.

«L'obsolescenza tecnologica in ambito sanitario: criticità e metodi di prevenzione». *ICT Security Magazine* (blog).
7 dicembre 2016. https://www.ictsecuritymagazine.com/articoli/lobsolescenza-tecnologica-ambito-sanitario-criticita-metodi-prevenzione/. Consultato il 4 novembre 2019.

PANDORA, http://www.pandora.nla.gov.au/historyachievements.html. Consultato il 28 ottobre 2019.

SiteStory. https://coptr.digipres.org/SiteStory. Consultato il 20 ottobre 2019.

Social Feed Manager. https://gwu-libraries.github.io/sfm-ui/. Consultato il 3 novembre 2019.

Treccani, alla voce obsolescenza digitale. http://www.treccani.it//enciclopedia/obsolescenza-digitale_(Enciclopedia-Italiana). Consultato il 28 settembre 2019.

WayBackMachine, https://archive.org/about/. Consultato il 26 ottobre 2019.

Webrecorder.io, https://Webrecorder.io/ . Consultato il 4 novembre 2019.

Wikipedia, voce *International Internet Preservation Consortium*, https://en.wikipedia.org/wiki/International_Internet_Preservation_Consortium. Consultato il 26 ottobre 2019.

WikiZero, voce *Memoria digitale*. https://www.wikizero.com/it/Memoria_digitale. Consultato 4 novembre 2019.