L'organizzazione e la descrizione di un fondo nativo digitale: PAD e l'archivio Franco Buffoni

Paul Gabriele Weston **Primo Baldini** **Laura Pusterla**
Università degli studi di Pavia
{name.surname}@unipv.it

Abstract

English: Within PAD-Pavia Archivi Digitali, a project aimed at the medium and long-term preservation of born digital archives belonging to Italian writers and humanists, several procedures based on Franco Buffoni's files have been tested in order to provide a better and more sustainable description of an archive of such nature. Through IT capabilities, PAD has sought to provide the end user with the greatest possible number of access points to the resource. PAD has also experimented implementing by computer-assisted treatment, based on comparison algorithms, the description carried out manually. This test took place on textual material from the author's website which had been previously saved.

Italiano: PAD-Pavia Archivi Digitali, progetto volto alla conservazione a medio e lungo termine di archivi d'autore, ha sperimentato le procedure occorrenti alla descrizione di un archivio nativo digitale sul Fondo Franco Buffoni (scrittore e poeta, traduttore, anglista). Attraverso le potenzialità informatiche, PAD ha cercato di fornire all'utente finale la maggior quantità possibile di punti di accesso alla risorsa. Dopo averne effettuato la descrizione in modo tradizionale, PAD ha sperimentato una forma di descrizione assistita dal computer, basata su un algoritmo di comparazione. Questa prova è avvenuta sul materiale salvato dal sito web dello stesso autore.

1 Introduzione

Il progetto PAD-Pavia Archivi Digitali dell'Università di Pavia è nato nel 2009 con lo scopo di preservare dalla scomparsa gli archivi delle memorie digitali di autori contemporanei. L'Università, che attraverso il Centro per la tradizione manoscritta di autori moderni e contemporanei dal 1969 salvaguarda i documenti cartacei di scrittori e giornalisti italiani, ha voluto estendere questa esperienza ai documenti nativi digitali. Fu il giornalista e scrittore Beppe Severgnini, ex alunno dell'Università, che, partendo dalla constatazione che una parte maggioritaria della produzione culturale letteraria si basa ormai sull'utilizzo di supporti informatici, sollecitò questo ampliamento di prospettive. Per rendere tali documenti ricercabili e leggibili è necessario servirsi di infrastrutture hardware e software in continua evoluzione, un ostacolo che rende le procedure della conservazione progressivamente più impegnative con il passare degli anni. La volontà di arginare questa perdita di testimonianze della nostra storia culturale è stata alla base della creazione del progetto PAD.
PAD conserva diverse tipologie di materiali digitali, garantisce la tutela a lungo termine dei fondi ed eventualmente può essere accessibile agli studiosi, nel pieno rispetto, come è ovvio, delle disposizioni ricevute dagli autori.

Fino ad ora il progetto si è focalizzato soltanto sulla conservazione a lungo termine degli archivi digitali in locale, ospitati cioè sui dispositivi di scrittura correntemente utilizzati dagli scrittori o su apparecchiature non più utilizzate, ma da essi conservate, nonché sui supporti di archiviazione utilizzati dagli stessi nel corso degli anni (nastri magnetici, floppy di diverse dimensioni e densità di archiviazione, cd, dvd, unità compatte di archiviazione massiva). La crescente tendenza ad avvalersi della rete per comunicare ed archiviare dati ha reso necessario mettere a punto una strategia e dei dispositivi finalizzati alla salvaguardia di risorse digitali, siti web e contenuti sui social media. A questo modulo del sistema è stato dato nome PAD Web Archiving. L'intento di PAD non è, ovviamente, quello di competere con analoghi progetti internazionali di ben altro respiro, ma mantenersi come un progetto sostenibile, sia tecnologicamente, sia finanziariamente, che garantisca, ad onta delle sue dimensioni contenute, dei risultati di qualità. Spetta agli autori stessi o alle istituzioni culturali alle quali fanno capo i siti interessati richiedere espressamente che anche questa componente venga inserita nel piano complessivo di preservazione dell'archivio. L'accordo è indispensabile al fine di interagire direttamente con il committente per stabilire tempi e metodi per il salvataggio e la consultazione. Tutto il materiale resta ovviamente di proprietà dell'autore, che può in ogni momento decidere di rimuoverli dall'archivio e di rinunciare al prosieguo del progetto.

2 Il fondo Franco Buffoni

Franco Buffoni, anglista, poeta, prosatore e traduttore, il cui archivio cartaceo si trova già in deposito presso il Centro Manoscritti della stessa Università, avendolo lui conferito in anni precedenti, è uno degli autori che, nel corso degli anni, hanno conferito a PAD i propri archivi digitali. Nel 2016, nel rispetto dell'iter messo a punto allo scopo da PAD, una copia del suo ampio archivio è stata riversata in PAD. L'iter a cui si fa qui riferimento prevede che, dopo la firma di un contratto legale, un operatore di PAD si rechi presso la residenza dell'autore per prelevare una copia dei file che lui stesso ha selezionato per la conservazione. Il riversamento dell'archivio Buffoni ha riguardato 1065 elementi, per complessivi 758 MB, comprendenti tipologie di file di diversa natura: documenti di testo, immagini, video, audio, link. Nel 2019, con l'intenzione di sperimentare lo strumento messo a punto per la salvaguardia a lungo termine delle risorse web, PAD ha concordato con l'autore di utilizzare il suo sito personale (www.francobuffoni.it), ritenendolo particolarmente idoneo allo scopo a motivo della ricchezza di contenuti e della varietà di formati e tipologie.

Per prima cosa, su richiesta esplicita del proprietario del sito, PAD ne ha prelevato una copia. Dato che i siti web possono essere modificati o aggiornati anche molto di frequente, si è concordato con l'autore di procedere con l'effettuazione di salvataggi a cadenza prestabilita. In questo modo si possono conservare le varie versioni del sito, che possono essere messe a disposizione dell'utenza secondo la volontà del proprietario. Attraverso un software per il web scraping, il sito dell'autore è stato riprodotto in locale, in modo da garantirne il browsing offline. Così l'utente futuro potrà navigare liberamente nella copia dell'intero sito. Per progettare questa implementazione, si è dovuto tenere conto della struttura anche molto complessa che i siti possono talvolta presentare, comprendente riferimenti numerosi ad altre pagine, interne o esterne nel web. Per questo di ogni pagina che compone il sito web, PAD memorizza, oltre alla pagina stessa, i link anche alle pagine esterne, con un'immagine della pagina a cui il link conduce, nonché i documenti allegati. In questo modo si può tenere meglio traccia dei path che il creatore del sito ha voluto valorizzare. Se, ad esempio, un link a una pagina esterna non fosse più funzionante o se la pagina non risultasse più esistente, una parte di ciò che l'autore intendeva comunicare, una componente probabilmente significativa del suo pensiero, andrebbe perduta.

Quando il progetto ha avuto inizio è stata presa in considerazione l'idea di utilizzare principalmente il formato di archiviazione WARC (Web ARChive). Sebbene questo formato sia stato standardizzato nel 2009 (ISO 28500:2017) il suo utilizzo da parte delle grandi aziende informatiche (Microsoft, Apple, Google ecc.) non ha mai goduto negli anni della diffusione che sarebbe stata auspicabile. Un'accurata serie di verifiche ha permesso di accertare che i browser più diffusi non lo riconoscono. Allo stesso tempo il sistema di memorizzazione sembra essere stato realizzato per essere installato e usato solamente da un sistemista esperto, ciò che rischia di creare notevoli problemi agli utenti comuni. Si è preferito quindi adottare un prodotto di più facile utilizzo per l'elaborazione del sito. Il formato WARC è, invece, stato mantenuto per la parte del progetto che si occupa di preservazione a lungo termine. Quindi in PAD per il Web Archiving vengono gestiti due sistemi diversi. Il primo utilizza il software Heritrix, sviluppato da Internet Archive, mentre il risultato dei processi di crawling viene memorizzato in file con formato WARC.

Per l'elaborazione delle pagine il software che PAD utilizza prevalentemente si chiama HTTrack, un Web crawler open-source. Consente di scaricare un sito web da internet in una directory locale, ottenendo HTML, immagini e altri file dal server al computer. HTTrack mantiene la struttura originale del sito, compresi i link, permettendo all'utente di navigare da una pagina all'altra, come se la stesse visualizzando online. Esso preleva anche tutte le altre tipologie di documenti e immagini che si trovano allegate alle pagine web. Pur non basandosi su uno standard, HTTrack presenta il vantaggio della semplicità nell'aprire le pagine web o nell'estrarre il testo per elaborarlo. Alcuni autori, come ad esempio Francesco Pecoraro, hanno richiesto di creare una copia offline del proprio sito, con l'intenzione di rimuoverlo successivamente dalla rete. Questa copia resta ovviamente a loro disposizione per l'accesso, qualora ne facciano richiesta, anche a distanza. Si è visto come la versione 'mirror' del sito è stata considerata come quella più semplice da inviare e da consultare da parte di utenti non particolarmente esperti. La possibilità di navigare all'interno del sito locale, senza l'obbligo di installare preventivamente specifici software, ha reso questo servizio estremamente semplice da gestire. Al contrario, per le questioni ricordate in precedenza, l'utilizzo di un archivio WARC avrebbe comportato la necessità di assistere l'utente nel corso della procedura di consultazione.

Tutto il materiale così raccolto è stato sottoposto alle procedure di conservazione sperimentate da PAD nel corso degli anni. La prima operazione è creare più copie dell'archivio, ubicate su diversi server. Oltre che sul server interno di PAD infatti, esso viene replicato sui server dell'Università di Pavia e su quello della sede

distaccata dell'Università a Cremona, città distante da Pavia circa 70 chilometri, in modo da salvaguardare la sicurezza delle informazioni in caso di disastro ambientale. Un'ulteriore copia viene memorizzata su supporto hardware esterno. Una volta assicurata la preservazione dell'originale, l'archivio passa in un'area di working.

Vengono estratti i metadati, fondamentali per poter poi svolgere l'operazione di normalizzazione, che comporta il salvataggio di ogni documento in diversi formati, a seconda della tipologia. Terminate le operazioni preliminari, si procede a descrivere i file.

3 La descrizione

Rispetto al trattamento di un archivio cartaceo, un archivio nativo digitale presenta peculiarità, come ad esempio il numero dei file che lo costituiscono, che talvolta possono assommare a molte migliaia, che rendono la descrizione effettuata seguendo consuetudini e procedure tradizionali inadeguata e persino non sostenibile. È stato, perciò, necessario individuare strategie che potessero consentire di sfruttare al massimo le potenzialità offerte dall'informatica.

Al contempo, si registrano problematiche simili, come la questione dell'accessibilità e della riservatezza. Trattandosi di documentazione prodotta molto di recente, si è dovuto tener conto del fatto che, probabilmente, una parte anche significativa dei file non possano essere resi disponibili all'utenza senza che ciò comporti la violazione di disposizioni legislative, come quelle sulla privacy o sulla proprietà intellettuale. Anche il fatto che i documenti conferiti o salvati dal web siano stati indicati espressamente dal conferente, non è sufficiente a garantirne la libera consultazione da parte degli studiosi. Si rende perciò necessario, durante le fasi preliminari della descrizione, sottoporre ogni singolo file ad una attenta disamina volta ad escludere che contenga dati sensibili o creazioni intellettuali la cui responsabilità non sia in capo al conferente, al soggetto produttore dell'archivio o al titolare del sito. Anche lo scrittore che ha conferito il proprio archivio potrebbe richiedere, per ragioni personali, che alcuni documenti siano secretati e di conseguenza esclusi dalla consultazione, anche per motivi di studio, per un determinato periodo di tempo, il cosiddetto embargo. Per tener conto di queste evenienze, l'operatore di PAD, nel vagliare ogni file dell'archivio digitale, gli assegna una categoria di rischio, in base alla quale esso viene automaticamente reso o meno consultabile da parte degli utenti.

Si passa poi alla fase del riordino. Come per gli archivi cartacei, viene creata una struttura ad albero rovesciato che comprende le diverse serie, alle quali vengono poi assegnati i singoli file. Già in questa procedura si manifestano quelle potenzialità dell'informatica, prima ricordate, che offrono un significativo contributo agli archivisti e nuove opportunità agli utenti. In primo luogo, il sistema offre la possibilità di assegnare un singolo file a più di una sezione dell'archivio. In secondo luogo, se nell'archivio cartaceo il riordino comporta la modifica della sistemazione pensata dallo scrittore, l'archivio digitale può consentire di mantenere ad un tempo l'aspetto originale e contemporaneamente collocare i documenti in un ordinamento che segua altri criteri. Questi criteri possono anche essere più di uno, quando le esigenze lo richiedano. Durante la descrizione, infatti, l'archivista assegna a ogni documento dei tag, che hanno lo scopo di aiutare l'utente a comprendere meglio la tipologia del materiale in questione (se, ad esempio, si tratta di un testo, di una recensione, di un'immagine, di un video e così via). In ogni archivio, ovviamente, non tutti i file sono prodotti da colui o colei che conferisce l'archivio stesso. Sono molto frequenti i casi di documenti frutto del lavoro intellettuale di terze persone. La possibilità di collegare a ogni file uno o più nomi di persona che abbiano in qualche modo contribuito alla sua produzione è funzionale anche a questo scopo. Al tempo stesso, il nome della persona o dell'ente viene associato ad una tipologia di responsabilità intellettuale, espressa attraverso un vocabolario controllato, implementabile a seconda delle esigenze mediante l'inserimento di nuovi termini in una tabella. Ricorre, poi, il caso di documenti - il termine viene qui utilizzato in senso generale, senza cioè fare riferimento a funzioni di natura amministrativa - che siano stati estratti o ricavati da pubblicazioni più ampie (ad esempio, un capitolo da un libro, una poesia da una raccolta, un brano da un'intervista o da una recensione e così via). Qualora il collegamento tra i due documenti sia riconosciuto dall'archivista, il sistema consente di esplicitare la relazione anche a beneficio dell'utente, dal momento che è possibile stipulare un collegamento tra l'oggetto e il titolo della risorsa che lo contiene o del quale è parte. L'inserimento del codice ISBN o DOI nel caso di un libro o dell'ISSN per una rivista è funzionale a rendere più riconoscibile e in modo inequivoco la fonte e, di conseguenza, a consentirne più facilmente il recupero.

La presenza di tutte queste informazioni inserite dall'archivista (tag, nomi, responsabilità, identificativi univoci), unitamente ai metadati tecnici estratti direttamente dalla macchina, consente a PAD di mettere a disposizione dell'utente percorsi di ricerca alternativi, o, per meglio dire, complementari e trasversali, rispetto a quelli consueti. Infatti, vi è la possibilità di estrarre i dati dei documenti secondo l'ordinamento per serie

archivistiche, oppure secondo la disposizione originale dello scrittore, o ancora ordinati per soggetto produttore o per provenienza. In questo modo PAD cerca di venire incontro alle variegate necessità dell'utente che si trova a consultare l'archivio.

4 La descrizione del sito web

Analizzando i conferimenti effettuati dai diversi autori ci si è resi conto che spesso l'autore ha conservato i testi che poi vengono riversati sul web sotto forma di file allegati alla pagina o come link. In altri casi, i testi della pagina web si ispirano, prendono spunto oppure sono almeno in parte i medesimi di quelli presenti nell'archivio. Operando sui metadati estratti, avendo selezionato quelli più importanti per identificare in modo puntuale la risorsa web, tali correlazioni tra nativo digitale e web possono venire ricercate ed individuate.
Le procedure occorrenti non si discostano molto da quelle normalmente messe in atto in PAD al momento dei conferimenti per analizzare la struttura e la consistenza dell'archivio. Come primo passo, si procede a creare una mappa del sito; in secondo luogo si estraggono i metadati, il cui numero viene ridotto in seguito alla scrematura di quelli di scarsa utilità; infine, i documenti vengono sottoposti ad operazioni di normalizzazione. Conclusa questa fase, i documenti del web possono essere descritti.

Il sistema di descrizione è in parte automatizzato, dato che la grande quantità di materiale disponibile sul web richiederebbe tempi eccessivamente lunghi e un'attività dispendiosa, se ad occuparsene fosse un operatore. È evidente, infatti, che, una volta ricevuto il comando, il computer possa elaborare per diverse ore il materiale, fino ad arrivare alla conclusione. Per ottenere un risultato molto simile, un operatore dovrebbe lavorare per giorni. Il software PAD Web Analyzer confronta ogni pagina web - prendendo in considerazione sia il testo vero e proprio della pagina, sia eventuali file allegati - con i documenti presenti nell'archivio. Procede, quindi, a mostrare, affiancati, i testi del web e i corrispondenti testi del nativo digitale, sulla base di indicatori di similitudine. Per ciascuna corrispondenza viene stabilito un indice di similitudine, che indica in percentuale quanta parte del testo sul web sia uguale a quella di un documento presente nell'archivio nativo digitale. Un risultato molto basso, indicativamente inferiore al 50 %, non viene tenuto in considerazione dall'operatore. Al contrario, come hanno permesso di accertare le prove effettuate, la certezza di aver individuato il medesimo documento si ha quando l'indice presenta un valore intorno al 98%. In presenza di valori intermedi sarà compito dell'operatore effettuare i riscontri necessari, ma anche in questo caso il sistema è in grado di offrire assistenza.

Se il valore è relativamente elevato, il computer, assumendo di aver individuato due file "probabilmente" corrispondenti propone l'assegnazione ad una delle serie inserite con la descrizione manuale dell'archivio. Se, viceversa, il valore è relativamente baso, la casella delle assegnazioni resta vuota e deve essere quindi l'operatore a riempirla sulla base di una ricognizione autoptica. Per effettuare queste valutazioni si è tenuto conto del fatto che talvolta i documenti in archivio possono essere in formati differenti rispetto a quelli pubblicati o allegati sul web (ad esempio un file Word solitamente viene convertito in PDF per essere messo sul sito) e questo comporta un ragionevole abbassamento dell'indice. Poiché la macchina non ha le stesse capacità di discernimento di un operatore, è opportuno che il controllo finale sia effettuato esaminando in parallelo i due testi. Una specifica funzione del software di PAD mostra i due testi affiancati in modo che la ricognizione possa procedere speditamente. È ovviamente è possibile che il testo sia stato prodotto direttamente sulla rete, oppure che il file originale sia stato eliminato o non conferito: in tal caso non si evidenziano corrispondenze.

A questo scopo PAD ha implementato all'interno del software l'algoritmo Levenshtein distance, adattandolo alle proprie esigenze. La procedura appena descritta risulta attualmente vantaggiosa solo se la parte dell'archivio digitale nativo sia stata già trattata e viene quindi utilizzata in questo momento unicamente per la descrizione dei siti web o in caso di secondi (o comunque successivi) conferimenti.

Le informazioni ottenute attraverso questo procedimento serviranno all'archivista per assegnare il documento a una particolare serie inventariale, riguardante una specifica opera dell'autore o una specifica tipologia documentale. Integrando i documenti provenienti dai siti web con quelli nativi digitali si ottiene un archivio il più completo possibile. Se all'interno di una serie viene inserito del materiale proveniente dal web, viene creata una sottoserie apposita, in modo che l'utente possa trovare aggregato tutto il materiale avente lo stesso argomento e al contempo conoscerne la provenienza.

5 Prospettive per il futuro

Come nella tradizione delle realizzazioni informatiche, il sistema PAD viene costantemente implementato per fornire nuove funzionalità e migliorare quelle attuali. La scelta di configurarsi come un progetto di piccola

scala, limitato a autori o a istituti culturali selezionati, permette di dedicare grande cura nel perfezionare le soluzioni tecniche, secondo le necessità dell'archiviazione e della descrizione. L'architettura di PAD è stata pensata per la conservazione, ma negli anni si è evoluta con la finalità di consentire lo studio del materiale conferito. La stretta collaborazione con gli scrittori conferenti garantisce il rispetto delle loro decisioni sul trattamento e la gestione del loro archivio e permette di andare incontro alle esigenze che essi manifestano. Seguendo le tendenze dei cambiamenti sociali nell'uso di internet e delle sue risorse, PAD sta sperimentando una funzionalità che permette la conservazione a lungo termine e la consultazione delle pagine personali di social network, come Facebook, Twitter o Instagram, dei canali YouTube e delle e-mail. Questa tipologia di documenti informatici, che potrebbero anche essere di rilevante importanza per studi futuri, sono per ora solo pensati in funzione della conservazione. Su richiesta esplicita di un autore, verranno raccolti i dati direttamente dai social e preservati. Con lo stesso criterio verrebbero trattate le e-mail, utilizzando criteri analoghi a quelli che negli archivi tradizionali si applicano ai carteggi e agli epistolari.

Attualmente è al vaglio la possibilità di affidare un'ulteriore copia degli archivi al progetto nazionale Magazzini Digitali, avviato nel 2006 dalla Fondazione Rinascimento Digitale, dalla Biblioteca nazionale centrale di Firenze e dalla Biblioteca nazionale centrale di Roma. La conservazione digitale assicurata dai depositi digitali affidabili o fidati (trusted or trustworthy digital repositories) di un servizio pubblico è una ulteriore garanzia che archivi digitali di autore così preziosi e che rischiano l'oblio vengano adeguatamente conservati nel lungo termine.

Bibliografia

Bergamin, G., and Messina, M. 2010. Magazzini digitali: dal prototipo al servizio. *DigItalia* 1, 115-122.

Black, Paul E. 2008. Levenshtein distance. *Dictionary of Algorithms and Data Structures [online]*. U.S. National Institute of Standards and Technology, retrieved 12/09/2019

Costa, M., Gomes, D., and Silva, M. J. 2017. The evolution of web archiving. *International Journal on Digital Libraries*, *18*(3), 191-205.

Klein, M., Shankar, H., Balakireva, L., and Van de Sompel, H. 2019. The Memento Tracer Framework: Balancing Quality and Scalability for Web Archiving. *International Conference on Theory and Practice of Digital Libraries* 163-176.

Masanès, J. 2006. Web archiving: issues and methods. *Web Archiving*. Springer, Berlin, Heidelberg.

Weston, P. G., Carbé E. and Baldini P. 2017. Hold it All Together: a Case Study in Quality Control for Born-Digital Archiving. *Qualitative and Quantitative Methods in Libraries* 5.3, 695-710.

Weston, P. G., Carbé E. and Baldini P. 2017. If bits are not enough: preservation practices of the original contest for born digital literary archives. *Bibliothecae. it* 6.1, 154-177.