I giornali storici nell'era digitale. Dal file immagine alla ricerca full text

Katalin Szabò
Biblioteca Provinciale Dr. F. Teßmann, Bolzano

Le raccolte di giornali storici sono una fonte importante per ricercatori delle branche più disparate. Ma la loro consultazione spesso è scomoda e gli originali si deteriorano facilmente. Digitalizzazione e messa in linea contribuiscono alla loro conservazione e facilitano l'accessibilità e la diffusione, ma non li rendono automaticamente più fruibili. Un file immagine messo in rete permette di consultare il giornale da qualsiasi parte del mondo, ma non dà la possibilità di cercare nomi, luoghi o eventi concreti all'interno di una testata. Negli ultimi anni in Europa sono state avviate numerose iniziative di digitalizzazione dei periodici, con approcci e obiettivi molto diversi: dalla digitalizzazione di massa all'elaborazione approfondita di singole raccolte specifiche. La Biblioteca Provinciale Dr. Friedrich Teßmann[1] di Bolzano nel 2006 avviò la digitalizzazione del suo fondo di giornali del Tirolo storico, che attualmente comprende 47 testate prevalentemente in lingua tedesca. Decise di non limitarsi al posseduto ma compì lo sforzo di integrarlo, dove possibile, con le edizioni mancanti, in modo da poter offrire alla propria utenza collezioni possibilmente complete dell'intera vita di un periodico. Per questo motivo avviò cooperazioni con altre biblioteche dell'Alto Adige, del Trentino e del Tirolo del Nord. Gli accordi prevedevano la fornitura degli originali da parte delle istituzioni partner, che in cambio ottenevano una copia digitale delle edizioni messe a disposizione. In una prima fase la digitalizzazione si limitava alla produzione di semplici file immagine, resi accessibili dalle postazioni di lavoro della biblioteca. Ma l'obiettivo era sviluppare un software per la consultazione online, che permettesse la ricerca per data, sul modello del portale ANNO[2] della Österreichische Nationalbibliothek (Biblioteca nazionale austriaca). Una prima versione andò in linea nel 2008 e per la prima volta (nei limiti di quanto previsto dalla normativa sul diritto d'autore) la raccolta di giornali storici era consultabile anche via internet.

Questa prima fase di cooperazione più informale tra biblioteche e archivi di regioni limitrofe, precedette di qualche anno la cooperazione ufficiale nata con la creazione dell'Euroregione alpina Tirolo-Alto Adige-Trentino, che nel 2010 portò alla sottoscrizione di un accordo di cooperazione tra cinque biblioteche scientifiche dell'Euregio, volto a coordinare, tra le altre cose, anche le attività di digitalizzazione.

Da un contesto spiccatamente locale e regionale, dal 2012 la Biblioteca Provinciale va ad inserirsi in un contesto ben più ampio, diventando uno dei 19 partner iniziali di Europeana Newspapers[3], progetto europeo, il cui scopo è quello di raccogliere in Europeana[4] ampie col-

[1] http://www.tessmann.it/.
[2] http://www.anno.onb.ac.at/.
[3] http://www.europeana-newspapers.eu/.
[4] Uno dei progetti culturali europei più ambiziosi con l'obiettivo di rappresentare online il patrimonio culturale europeo, http://www.europeana.eu/.

lezioni di giornali storici europei e renderli accessibili attraverso un unico portale. La partecipazione fu possibile grazie a 1,5 milioni di pagine di giornali storici digitalizzati, che la Biblioteca Provinciale ha fatto confluire nel progetto. Uno degli obiettivi di Europeana Newspapers è la creazione di strumenti di ricerca per rendere più fruibile quell'immenso patrimonio culturale, attraverso l'introduzione di modalità di ricerca per data, per lingua, per paese o per titolo di giornale, la trasformazione dei file immagine in file di testo ricercabili, il riconoscimento automatico delle entità nominali o il riconoscimento strutturale delle pagine di giornale. Al termine del progetto (prolungato al 31 marzo 2015), l'utente avrà a sua disposizione oltre 30 milioni di pagine di quotidiani digitalizzati, di cui oltre 10 milioni di pagine ricercabili *full text,* provenienti da 25 biblioteche di 23 paesi europei, raggiungibili dal sito di Europeana oppure da quello di The European Library[5].

Da queste esperienze, la Biblioteca Provinciale Dr. F. Teßmann ha sviluppato l'idea di organizzare un convegno sulle tematiche connesse alle raccolte digitali di giornali storici e agli strumenti di *post-processing* disponibili per migliorare i risultati di ricerca. Durante la giornata, dal titolo "I giornali storici nell'era digitale. Dal file immagine alla ricerca full text", tenutasi in data 27 ottobre 2014 a Bolzano, 9 esperti da 4 paesi europei (Germania, Austria, Italia e Inghilterra) hanno approfondito una serie di argomenti che ora si passeranno brevemente in rassegna.

Nella relazione introduttiva, Dave Thompson, *digital curator* presso la Wellcome Library di Londra, ha sintetizzato il tipico workflow di un grande progetto di digitalizzazione, evidenziando i punti critici e gli errori da evitare. Innanzitutto ha sottolineato l'importanza di formulare obiettivi chiari e una strategia in sintonia con la filosofia istituzionale, senza perdere mai di vista la cosa più importante, ovvero le esigenze dell'utente finale. Secondo Thompson, la digitalizzazione non è un'attività relegata ai tecnici, ma va considerata come attività sociale che coinvolge l'intera istituzione, richiede un insieme di competenze diverse e va organizzata attraverso processi standardizzati, continuamente controllati e rivisti per assicurare un continuo miglioramento.

Il secondo relatore, Günter Mühlberger, del reparto per la digitalizzazione e l'archiviazione elettronica dell'Università di Innsbruck, DEA[6], che sin dagli anni Novanta si occupa del riconoscimento ottico dei caratteri OCR (Optical Character Recognition) e del layout OLR (Optical Layout Recognition), ha collaborato in numerosi progetti europei come IMPACT (Improving Access to Text) e Europeana Newspapers. Per quest'ultimo, il DEA ha elaborato otto milioni di pagine di giornale. Contrariamente all'opinione diffusa che il trattamento OCR sia troppo costoso, Mühlberger ribadisce che con 1-2 centesimi per pagina, i costi del riconoscimento ottico sono di gran lunga inferiori a quelli dell'acquisizione dell'immagine o *image capturing,* che si aggira intorno ai 35 centesimi. La ricerca a testo pieno ormai costituisce uno strumento indispensabile per accedere ai contenuti delle raccolte digitali. Anche se l'OCR applicato a periodici storici, per via della carta di scarsa qualità e spesso in cattivo stato di manutenzione (e in particolar modo per i caratteri gotici in uso nella stampa tedesca fino al XX sec.), spesso non raggiunge livelli di precisione altissimi, migliora comunque la fruibilità delle raccolte permettendo la ricerca a testo pieno. Un altro aspetto che costituisce una sfida particolare è la struttura grafica molto complessa dei giornali, se paragonata a quella di un libro, con caratteri di diverse grandezze nei titoli o negli annunci, l'impaginazione a colonne o la presenza di immagini e didascalie. Per questo presso il DEA è stato sviluppato Structify, uno strumento per individuare

[5] http://www.theeuropeanlibrary.org/tel4/.
[6] http://www.uibk.ac.at/germanistik/dea/.

in maniera automatica le cosiddette unità di contenuto, come possono essere oltre agli articoli ad esempio le previsioni del tempo, gli annunci di lavoro, la pubblicità o i necrologi. La possibilità di effettuare ricerche limitate a determinate unità di contenuto potrebbe restringere ulteriormente il campo e portare a risultati di ricerca più stringenti. Si tratta di un programma open source, in grado di creare o correggere metadati strutturali, la cui versione attuale può essere scaricata dal sito di progetto Europeana Newspapers[7].

All'Accademia Europea (EURAC) di Bolzano sono in corso ricerche per correggere i risultati OCR con tecnologie per il trattamento e la comprensione automatica del linguaggio naturale. Nell'ambito del progetto OPATCH[8], Piattaforma aperta per la pubblicazione e l'analisi di documentazione testuale storica, Michel Généreux si occupa di metodologie di linguistica computazionale applicate alla correzione automatica degli errori di riconoscimento OCR. I primi risultati, presentati in occasione del convegno di Bolzano, sono incoraggianti: «We can expect to improve our dictionary coverage so that very noisy OCR-ed texts (i.e. 48% error with distance of at least three to target) can be corrected with accuracies up to 20%. OCR-ed texts with less challenging error patterns can be corrected with accuracies up to 61% (distance two) and 86% (distance one)».

Clemens Neudecker della Staatsbibliothek zu Berlin (Biblioteca di Stato di Berlino) e coordinatore del progetto Europeana Newspapers, ha confermato l'importanza di supportare le ricerche per nome NER (Named Entity Recognition), poiché da uno studio[9] sul tipo di ricerche maggiormente effettuate dagli utenti quando navigano nelle raccolte di giornali, emerse che 9 ricerche su 10 riguardano nomi di persone o luoghi. Un'analisi presso la Österreichische Nationalbibliothek (Biblioteca nazionale austriaca) su 200.000 *queries* di utenti conferma questo dato: soltanto un quarto circa dei termini si riferiva a vere e proprie *key word searches*, mentre un quarto delle ricerche riguardava i luoghi e ben la metà i nomi di persona. Per questo motivo anche Europeana Newspapers ha dedicato molta attenzione al riconoscimento delle entità nominali, lavorando sul miglioramento della performance del software open source Europeananp-ner[10], originariamente sviluppato dall'Università di Salford (Manchester), che è stato rielaborato per adeguarlo a un contesto multilingue.

Patrizia Rossi del CSI-Piemonte[11] ha presentato la BDIG (Biblioteca Digitale dell'Informazione Giornalistica), che ha l'obiettivo di rendere fruibili le edizioni storiche dei giornali piemontesi al fine di valorizzarle, conservarle e renderle disponibili al pubblico. Primo nucleo e fiore all'occhiello della BDIG è l'Archivio Storico La Stampa[12], iniziativa realizzata in collaborazione con la Società Editrice La Stampa, la Compagnia di San Paolo e la Fondazione Cassa di Risparmio di Torino. Il prossimo traguardo della BDIG è l'Archivio Storico dei Periodici Piemontesi, che il CSI-Piemonte sta realizzando basandosi sul riutilizzo dell'infrastruttura

[7] http://www.europeana-newspapers.eu/public-materials/tools/.

[8] OPATCH - Open Platform for Access to and Analysis of Textual Documents of Cultural Heritage; partner di progetto sono oltre all'EURAC e la Biblioteca Provinciale Dr. F. Teßmann anche l'Institut für Corpuslinguistik und Texttechnologie (Istituto di linguistica dei corpora e tecnologie del linguaggio dell'Accademia delle Scienze dell'Austria).

[9] Paul Gooding, *Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online*. In: *Digital Humanities 2014*, Lousanne, 8 luglio 2014, «dh2014.org».

[10] La Koninklijke Bibliothek (Biblioteca Nazionale dei Paesi Bassi) ha implementato le lingue olandese, tedesco e francese; il software è scaricabile da https://github.com/KBNLresearch.

[11] CSI-Piemonte è l'ente strumentale della Pubblica Amministrazione Regionale in campo informatico e telematico, http://www.csipiemonte.it/web/it/.

[12] http://www.archiviolastampa.it/.

software dell'Archivio Storico La Stampa e con la migrazione dei periodici locali piemontesi già digitalizzati disponibili nella Biblioteca Digitale Piemontese[13]. Obiettivo ultimo sarà la creazione dell'Archivio Storico dell'Editoria Piemontese (ASEP) pensato come un archivio federato dell'Archivio Storico dei Periodici Piemontesi e dell'Archivio Storico La Stampa, con lo sviluppo di un motore trasversale in grado di ricercare in entrambi gli archivi.

Andrea Bolioli della CELI Srl di Torino[14], ha illustrato alcuni aspetti tecnici riguardo la trasformazione di contenuti testuali in dati analizzabili. Innanzitutto si è occupato del riconoscimento automatico delle entità nominali per l'Archivio Storico La Stampa. Sono stati annotati automaticamente circa 5 milioni di articoli di giornale, per estrarre le citazioni di persone, luoghi, organizzazioni e gli autori degli articoli. Ma per consentire l'annotazione automatica (supportata da regole linguistiche o *pattern matching*), il programma è stato addestrato attraverso l'annotazione manuale di un campione significativo di pagine tramite un'interfaccia web collaborativa. Un'altra operazione molto complessa e dispendiosa è stata la mappatura automatica delle pagine per individuare la posizione dei titoli e il numero e la larghezza delle colonne dei singoli articoli. Nel corso di quasi 150 anni il quotidiano ha modificato di frequente la sua strutturazione arrivando a impaginazioni sempre più complesse, per cui sono state necessarie diverse operazioni manuali di verifica e correzione. Sono stati inoltre corretti manualmente tutti i titoli principali, che per la misura dei caratteri utilizzati generalmente non vengono riconosciuti correttamente dai programmi OCR, per renderli accessibili per le ricerche *full text*. Infine, inserendo un termine di ricerca, con i dati estratti è possibile creare delle infografiche di vario tipo. Questa visualizzazione delle informazioni consente all'utente di paragonare i dati e di captare le relazioni tra di essi, migliorando significativamente l'usabilità dell'archivio digitale.

Ad approfondire il tema della visualizzazione delle informazioni è stato Andrea Marchetti, tecnologo presso l'IIT (Istituto di Informatica e Telematica) del CNR di Pisa[15]. Le tecnologie web di *information extraction* e visualizzazione aprono nuove possibilità per rappresentare contenuti di collezioni altrimenti poco accessibili e accattivanti. Mappe geografiche, linee temporali, *word clouds* o infografici sono soltanto alcune fra le tante possibilità di visualizzazione. Un esempio di come materiale archivistico possa essere rappresentato in modo consono alle abitudini visive di oggi è il progetto ClaviusOnTheWeb[16]. Al progetto hanno collaborato l'Istituto di Linguistica computazionale e l'IIT del CNR di Pisa, i quali hanno sviluppato strumenti specifici per l'annotazione linguistica, lessicale e semantica dei testi, per la visualizzazione delle annotazioni e per i *linked data*. Il progetto ha l'obiettivo di valorizzare tramite web un archivio di manoscritti conservato dalla Pontificia Università Gregoriana, e in modo particolare un insieme di circa 300 lettere inviate a Cristoforo Clavio (1538-1612), matematico e astronomo gesuita. Il risultato è un sito web che applica i concetti dello *storytelling*, la rappresentazione tramite *timeline* della biografia di Clavius e mappe interattive per visualizzare i suoi viaggi.

Maurizio Messina, direttore della Biblioteca Nazionale Marciana di Venezia, nel suo intervento ha parlato invece della conservazione digitale, un aspetto fondamentale che tutti gli archivi e biblioteche grandi e piccole che sia-

[13] http://www.regione.piemonte.it/TecaRicerca/home.jsp.

[14] Società che progetta e sviluppa software basato su tecnologie semantiche e *natural language processing*, http://www.celi.it/index.shtml.

[15] http://www.iit.cnr.it/.

[16] http://www.claviusontheweb.it/.

no devono affrontare, ma che difficilmente sono in grado di affrontare da soli. La conservazione digitale richiede infrastrutture adeguate, possibilmente esterne e centralizzate non solo a livello italiano ma europeo, con protocolli, procedure e politiche comuni, garantendo l'interoperabilità e l'apertura dei dati. Richiede un insieme di servizi che siano disponibili in rete, di cui alcuni, come l'autenticazione federata o l'assegnazione di identificatori persistenti sono già disponibili[17], ma altri devono ancora essere creati, come ad esempio servizi *cloud* o *grid* chiavi in mano. Nel progetto europeo DCH-RP (Digital Cultural Heritage-Roadmap), sono state delineate le modalità di interazione fra le istituzioni culturali e le *e-infrastructures,* formulando piani d'azione per il breve, medio e lungo termine[18].

In chiusura è intervenuto Klaus Kempf della Bayerische Staatsbibliothek di Monaco (Biblioteca di Stato della Baviera) sul futuro ruolo delle biblioteche come gestori di collezioni sempre più orientate verso il digitale. Questo comporta che la biblioteca spesso non dispone più di un oggetto acquisito e quindi posseduto, ma soltanto di un diritto all'utilizzo, addirittura limitato nel tempo, che preclude le consuete modalità di utilizzo ampliato come il prestito a distanza e che richiede nuove riflessioni su prestazioni e servizi offerti dalle biblioteche. Ma a medio termine il mondo dell'informazione subirà cambiamenti ancora più radicali: il classico oggetto di raccolta basato sul testo stampato sarà sostituito da documenti veramente multimediali, o da risorse di rete. Di conseguenza, lo sviluppo sistematico e sostenibile delle proprie collezioni, ovvero la raccolta, la classificazione e la messa a disposizione di contenuti di qualsiasi tipo, richiederà una nuova forma di cooperazione, intesa in maniera più ampia, tra biblioteche di ogni tipo, ma anche con altri luoghi della memoria come archivi e musei. Lo sviluppo delle collezioni in futuro, secondo Kempf, sarà partecipato e condiviso. Da un lato crescerà l'importanza delle grandi istituzioni della memoria dove si concentrerà maggiormente l'attività di raccolta. Ma allo stesso tempo, anche le biblioteche più piccole possono aggiungere una tessera al grande mosaico delle collezioni digitali globali, ad esempio facendo confluire attraverso *repository* istituzionali in rete, le proprie collezioni documentali tipicamente locali e uniche nel loro genere. L'attività di raccolta e sviluppo dei fondi non sarà più il cuore ma soltanto uno fra tanti servizi offerti dalle biblioteche in base alle specifiche esigenze dei propri utenti[19].

Tutto il materiale relativo al convegno, comprese le videoriprese di tutti gli interventi e documentazione tecnica integrativa, sarà pubblicato in forma di e-book e messo a disposizione sul sito della Biblioteca Provinciale Dr. Friedrich Teßmann al fine di fornire una raccolta utile a chi, in futuro, intenda avviare un progetto di digitalizzazione. Da subito tutte le presentazioni sono disponibili su slideshare[20].

[17] Si pensi a Magazzini Digitali, il servizio nazionale coordinato di conservazione e accesso a lungo termine per le risorse digitali promosso dal MIBAC, Ministero dei beni e delle attività culturali e del turismo, attualmente in fase di sperimentazione con il servizio di deposito legale, <http://www.depositolegale.it/>.

[18] http://www.dch-rp.eu/index.php?en/115/roadmap-for-preservation.

[19] Kempf usa il termine di biblioteca ibrida, per approfondimenti si veda Klaus Kempf. *L´idea della collezione nell'età digitale: Lectio magistralis in Biblioteconomia*. Fiesole (Firenze): Casalini Libri, 2013, scaricabile da: <http://www.torrossa.it/pages/ipplatform/home.faces>.

[20] http://www.slideshare.net/Europeana_Newspapers.

L'ultima consultazione dei siti Web è avvenuta nel mese di dicembre 2014.