Digitalizzazione, trascrizione, citazione: le fonti testuali per le pubblicazioni digitali

«Dig*Italia*» 2-2023
DOI: 10.36181/digitalia-00084

Elisa Bastianello
Bibliotheca Hertziana – Istituto Max Planck per la storia dell'arte, Roma

Le digitalizzazioni dei libri stanno riscuotendo un enorme successo negli ultimi anni, permettendo di superare le limitazioni all'accesso ai libri fisici più rari. Ma, a ben guardare, una riproduzione digitale delle pagine di un libro non è necessariamente accessibile e ritrovabile. Per questo motivo, la Bibliotheca Hertziana, tra le altre istituzioni, sta investendo molte risorse nella trascrizione del contenuto delle digitalizzazioni, con il supporto di nuove tecnologie come il riconoscimento neurale del testo. I punti cruciali includono l'uso di standard condivisi come IIIF e TEI XML e piattaforme open source come TEI Publisher per garantire l'accessibilità e la conservazione a lungo termine dei contenuti digitali.

Lo stato dell'arte

Negli ultimi decenni l'accesso alle fonti bibliografiche, intese come libri antichi, ha cessato di essere limitato dalla possibilità di consultazione dei codici originali in biblioteca, o dall'esistenza di riproduzioni anastatiche o edizioni critiche. La realizzazione di campagne di digitalizzazione, con scansioni a risoluzione sempre maggiore permette ora allo studioso, come anche al lettore comune, di consultare manoscritti e antichi stampati e confrontarli tra di loro nella comodità del suo ufficio o della sua casa. Si tratta di un patrimonio molto disomogeneo, che include scansioni di microfilm, come quelle della fase iniziale di Gallica[1], scansioni parziali e a bassa risoluzione o alta compressione, fino alle altissime risoluzioni dei progetti più recenti con informazioni colorimetriche e metadati di presa degli ultimi anni. Anche le modalità e le piattaforme di visualizzazione sono molto diverse e includono pagine web con elenchi dei file PDF scaricabili da parte dell'utente, visualizzatori proprietari a cui è demandato il compito di impedire al lettore di scaricare le immagini molto più che non quello di semplificare la sua ricerca, come la storica piattaforma del Ministero della Cultura e dell'Istituto centrale per il catalogo unico, Internet Culturale[2], ma anche visualizzatori dotati di spazi di lavoro per la manipolazione, l'annotazione e il confronto delle immagini come quelli usati dalla Estense Digital Library (EDL) e le sue storie[3].
Quest'ultima opzione è finalmente possibile perché recentemente gli istituti di conservazione stanno iniziando a utilizzare degli standard per la condivisione delle immagini che, come nel caso del IIIF[4], semplificano il riuso delle stesse in altri ambiti anziché scoraggiarlo. Non solo, alcune tecnologie come la visualizzazione piramidale delle immagini[5] permettono di ridurre notevolmente i tempi di accesso alle im-

[1] <https://gallica.bnf.fr/>. Gallica è la biblioteca digitale della Bibliothèque Nationale de France.
[2] https://www.internetculturale.it/.
[3] https://edl.cultura.gov.it/.
[4] International Image Interoperability Framework, <https://iiif.io/>.
[5] Babcock Kelli — Rachel Di Cresce, *Impact of International Image Interoperability Framework (IIIF) on Digital Repositories*, in: *New top technologies every librarian needs to know*, a cura di K. J. Varnum, Chicago: ALA Neal-Schuman, 2019, p. 181–196.

magini stesse, in quanto non è mai necessario scaricare l'intera immagine ad alta risoluzione, ma solo i dettagli via via che si consultano, senza che il lettore ne sia rallentato. Un ulteriore vantaggio di questo standard è, come dice il nome del protocollo, l'interoperabilità, ovvero la possibilità di mettere a disposizione le immagini, e in particolare le sequenze delle immagini, perché possano essere utilizzate con visualizzatori indipendenti da quello originale della biblioteca digitale di partenza, consentendo al lettore di confrontare le pagine di più volumi, annotarle e riassemblarle in storyboard[6].

Per quanto, come recita l'adagio, una immagine valga più di mille parole, l'esistenza di queste scansioni di fonti, anastatiche digitali, non garantisce l'accessibilità ai contenuti dei documenti digitalizzati. Nel caso delle pagine di un codice manoscritto, per esempio, l'interpretazione del segno grafico e più semplicemente la lettura del contenuto è demandata al lettore che deve necessariamente avere le competenze, per esempio, di paleografia, esattamente come nel caso del manoscritto originale. La situazione non migliora del tutto nel caso dei testi a stampa antichi, la cui messa in pagina, insieme all'uso di abbreviature e legature, rende la lettura disagevole e faticosa ad occhi abituati a caratteri tipografici molto più regolari e nitidi di quelli di un incunabolo o di una cinquecentina, senza nemmeno considerare le difficoltà derivanti dall'uso dei caratteri gotici e di un linguaggio desueto quando non straniero. La trascrizione dei documenti da parte di esperti richiede tempi molto lunghi ed è stata, in generale, riservata alle sole opere di particolare interesse e pregio, con una copertura a macchia di leopardo per epoca, genere e area geografica di provenienza e conservazione.

Un vantaggio ulteriore della disponibilità delle fonti digitalizzate è che semplifica, e anzi per certi aspetti impone al ricercatore, l'onere della verifica delle fonti in prima persona. Se in passato le difficoltà di accesso dei testi originali rendevano accettabile la citazione attraverso la letteratura critica, di cui si davano per buone interpretazioni e trascrizioni, verificabili solo con peregrinazioni attraverso gli istituti di conservazione, l'esistenza dell'anastatica digitale non può essere ignorata, sia per la necessità di informare il lettore della sua esistenza, sia perché la comunità scientifica può molto più facilmente del passato verificare la veridicità delle affermazioni riferite. Questo meccanismo sta permettendo in più occasioni di porre fine a decenni di citazioni erronee, ma risulta molto oneroso per lo studioso che deve confrontarsi con degli strumenti di ricerca e citazione diversi da quelli della filologia tradizionale appresa, con cui la maggior parte dei ricercatori si è formato.

Esistono altresì delle limitazioni: un primo limite è dato dalla mancanza di un repertorio unico aggiornato delle fonti digitalizzate. Per scoprire se di un dato manoscritto o di un volume esistono delle scansioni sul web non basta la ricerca sui motori di ricerca. Le biblioteche digitali dei diversi istituti, infatti, non sono consorziate in un catalogo unico (sebbene l'ultima versione di WorldCat[7] abbia migliorato moltissimo la reperibilità di digitalizzazioni per i contenuti indicizzati in biblioteche afferenti di tutto il mondo), ma vanno ricercate in modo indipendente in quanto raramente i metadati delle digitalizzazioni sono esposti ai motori di ricerca. Limitandoci al solo ambito italiano, i manoscritti digitalizzati sono disponibili, per ragioni storiche, in parte su Manus Online[8] e in parte su Nuova Biblioteca Manoscritta[9]. In molti casi i codici si trovano solo attraverso la segnatura e non attraverso titolo e autore, come succede sia su Internet Culturale[10] (Fig. 1), che su Gallica o sulla biblioteca digitale della Biblioteca Vaticana[11], con il risultato che cercando per autore non sempre si arriva al contenuto quando non si abbia a disposizione la segnatura in partenza.

[6] Vedi in particolare il visualizzatore Mirador <https://projectmirador.org/> e sistemi di annotazione come Annotorius <https://github.com/annotorious/annotorious> di Recogito <https://recogito.pelagios.org/> e Annonatate <https://annonatate.fly.dev/>.

[7] https://worldcat.org/.

[8] https://manus.iccu.sbn.it/.

[9] https://www.nuovabibliotecamanoscritta.it/.

[10] https://www.internetculturale.it/.

[11] https://digi.vatlib.it/.

Questo senza arrivare al caso di codici di piccole biblioteche che si possono trovare solo se il sito dell'istituto di conservazione include una sezione di biblioteca digitale.

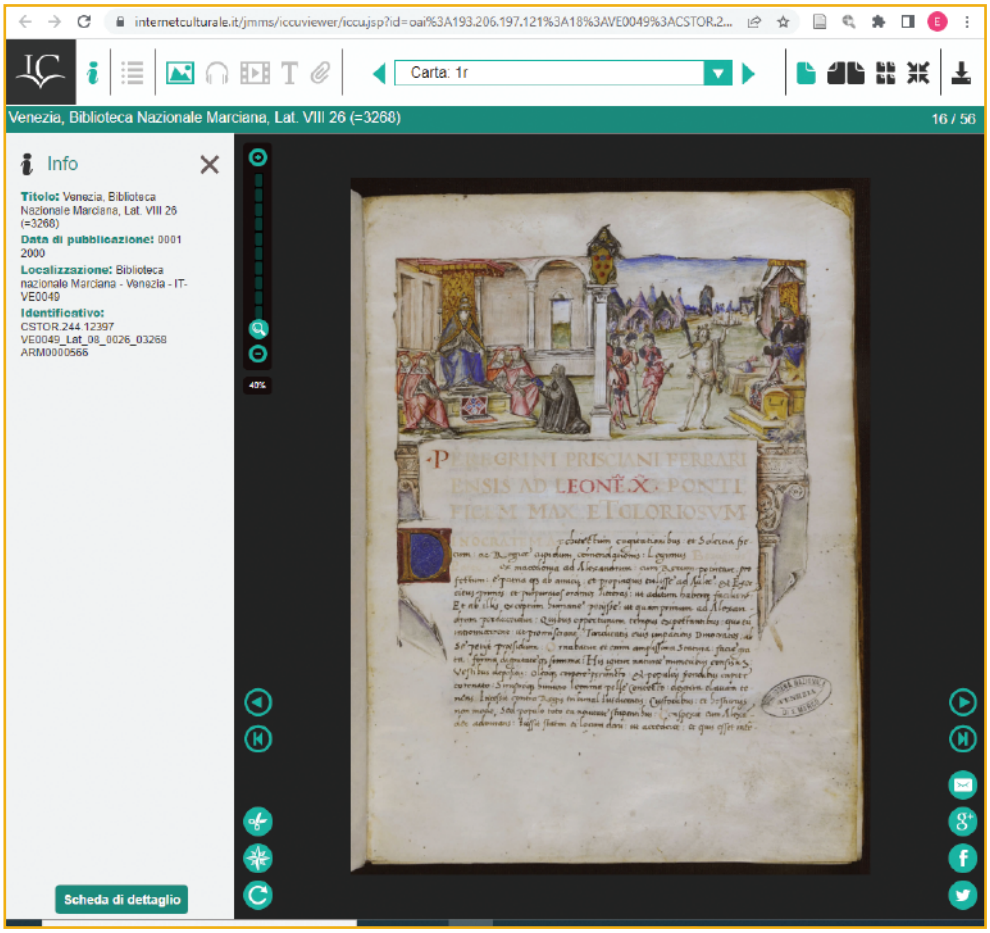

Figura 1. *Esempio di pagina dal sito Internet Culturale, dove l'autore e il titolo del documento non sono parte dei metadati e l'accesso è solo al documento e non alla pagina*

Fortunatamente anche in questo caso l'uso del protocollo IIIF permette di creare dei repertori come per esempio nel caso del progetto Biblissima[12] o nel progetto Open della stessa EDL, ma anche qui il reperimento dipende interamente dai metadati associati.
La difficoltà successiva per il ricercatore riguarda la citazione corretta e univoca della fonte. Infatti, mentre quasi tutti i cataloghi digitali permettono di citare l'intero volume, non è sempre agevole puntare al passo corretto o quantomeno alla carta. Nel caso di un codice cartaceo con la sola indicazione della segnatura e carta siamo sufficientemente esperti per muoverci tra le pagine, scorrendo rapidamente i numeri delle carte (se presenti); per farlo in digitale dobbiamo sperare che la cartulazione sia facilmente accessibile dall'interfaccia – che spesso si limita ad indicare il numero progressivo della presa fotografica – e che i tempi di caricamento e scorrimento siano accettabili. Ma anche quando la consultazione è rapi-

[12] https://biblissima.fr/.

da, non è detto che sia facilmente citabile in nota se non proprio con la forma tradizionale: alcune interfacce non permettono di citare in modo unico e persistente i documenti digitali (non sono stati assegnati identificativi PURL[13]) e ancora meno permettono di fare riferimento a una pagina specifica all'interno del codice oggetto di citazione. Questo significa che il lettore, ammesso che raggiunga la digitalizzazione delle pagine, debba procedere con la ricerca del contenuto citato per procedere con la verifica.

Il progetto HumanitiesConnect

Il progetto *HumanitiesConnect* della Bibliotheca Hertziana – Istituto Max Planck per la storia dell'arte nasce inizialmente dall'idea di creare una rivista digitale open access innovativa di ambito umanistico, con la possibilità di integrare annotazioni semantiche e connessioni tra i contenuti. La necessità di semplificare nei contenuti degli articoli i riferimenti alle fonti digitali ha imposto la presa di coscienza del panorama generale delle biblioteche digitali, alla ricerca di soluzioni che permettessero il riuso in modo semplice dei contenuti digitalizzati nelle nuove pubblicazioni. Per questo motivo lo sguardo è andato immediatamente alle risorse interne e alla loro ottimizzazione per renderle facilmente trovabili, accessibili, interoperabili e riutilizzabili come previsto dai principi dei FAIR data[14].

La Bibliotheca Hertziana da oltre un decennio ha iniziato un progetto di scansione del fondo dei libri antichi (Rara), attualmente disponibile secondo il protocollo IIIF nella biblioteca digitale[15]. Un primo progetto che includeva anche alcune trascrizioni scientifiche era stato fatto una decina di anni fa all'interno del progetto ECHO[16], coordinato dall'Istituto Max Planck per la storia della scienza (MPIWG), ma l'onere di trascrizione e di mantenimento dell'intera piattaforma si era rivelato eccessivo.

Anche se si tratta di testi a stampa, le tecnologie normalmente applicate per il riconoscimento automatico dei testi, cioè l'OCR (Optical Character Recognition) non funzionano in modo accettabile sui font antichi. Questo perché l'OCR opera analizzando i singoli caratteri e usa il confronto con un vocabolario per migliorare l'accuratezza delle trascrizioni e ridurre gli errori di lettura, causati per esempio dai difetti di stampa o di scansione. È facile intuire come queste caratteristiche si prestino poco alla tipografia antica, ricca di abbreviature e legature, e ancora meno a contenuti la cui ortografia, come nel caso dell'italiano fino almeno al XIX secolo, non è ancora stabilizzata, nemmeno all'interno dello stesso testo. Se per un libro recente possiamo attenderci dall'OCR una accuratezza ben sopra il 90-95%, essa scende drasticamente nel caso di testi pubblicati prima del XVIII secolo, rendendo la trascrizione inutile non solo ai fini dell'accessibilità, ma perfino della più basilare ricercabilità di parole del testo[17].

Da alcuni anni sono state sviluppate tecnologie basate sull'apprendimento automatico (*machine learning*) per il riconoscimento dei testi, in particolare quelli manoscritti, note come HTR[18], che sono in grado di migliorare l'accuratezza delle trascrizioni ampliando il contesto di riferimento dal singolo carattere all'intera riga[19]. Queste stesse macchine neurali possono essere usate anche sui testi a stampa, e anzi

[13] Persistent Uniform Resource Locator.

[14] Findability, Accessibility, Interoperability, and Reuse of digital assets <https://www.go-fair.org/>, v. anche Mark D. Wilkinson et al., *The FAIR Guiding Principles for scientific data management and stewardship*, «Scientific data» 3 (2016). DOI: 10.1038/sdata.2016.18.

[15] DLib <https://dlib.biblhertz.it>.

[16] https://echo.mpiwg-berlin.mpg.de/.

[17] Ryan Cordell, *"Q i-jtb the Raven": Taking Dirty OCR Seriously*, «Book History» 20 (2017) 1, p. 188–225. DOI: 10.1353/bh.2017.0006.

[18] Handwritten Text Recognition. Alcuni esempi di motori neurali per il riconoscimento dei testi sono HTR+ sviluppato da CITlab: <https://www.mathematik.uni-rostock.de/en/forschung/projekte/citlab/> (proprietario) e Pylaia: <https://github.com/jpuigcerver/pylaia> (open source).

[19] Joe Nockels et al., *Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research*, «Archival Science» 22 (2022) 3, p. 367–392. ‹https://link.springer.com/article/10.1007/s10502-022-09397-0›, DOI: 10.1007/s10502-022-09397-0.

stanno prendendo sempre più piede[20], al punto che si preferisce definirle motori OCR neurali, per evitare un'accezione legata al solo testo manoscritto. L'utilizzo diretto dei motori neurali richiede delle buone conoscenze informatiche, dimestichezza con la linea di comando e server dedicati. Per chi non ha a disposizione le infrastrutture e le competenze informatiche necessarie esistono fortunatamente delle piattaforme che offrono l'uso dei motori installati nei loro server attraverso una interfaccia in grado di guidare l'utente in quasi tutte le fasi del processo. In particolare dal 2016 (originariamente grazie ad un finanziamento europeo) è disponibile la piattaforma Transkribus[21], gestita dalla cooperativa sociale europea READ-COOP[22].

Il primo passo del progetto HumanitiesConnect è stato proprio quello di utilizzare le macchine di OCR neurale attraverso Transkribus per generare una prima trascrizione del contenuto di tutte le digitalizzazioni esistenti (fino al 2020) appartenenti alle collezioni digitali della Bibliotheca Hertziana di Roma, del Kunsthistorisches Institut in Florenz (KHI) e del Max Planck Institute for the History of Science (MPIWG)[23] di Berlino. Si tratta di oltre 3.800 volumi di storia dell'arte e della scienza, per un totale di oltre 1.300.000 pagine. Sono stati utilizzati diversi modelli neurali, ottimizzati rispetto al contenuto e all'epoca per le trascrizioni, in formato PAGE XML[24], che sono consultabili sulla speciale interfaccia di lettura e ricerca

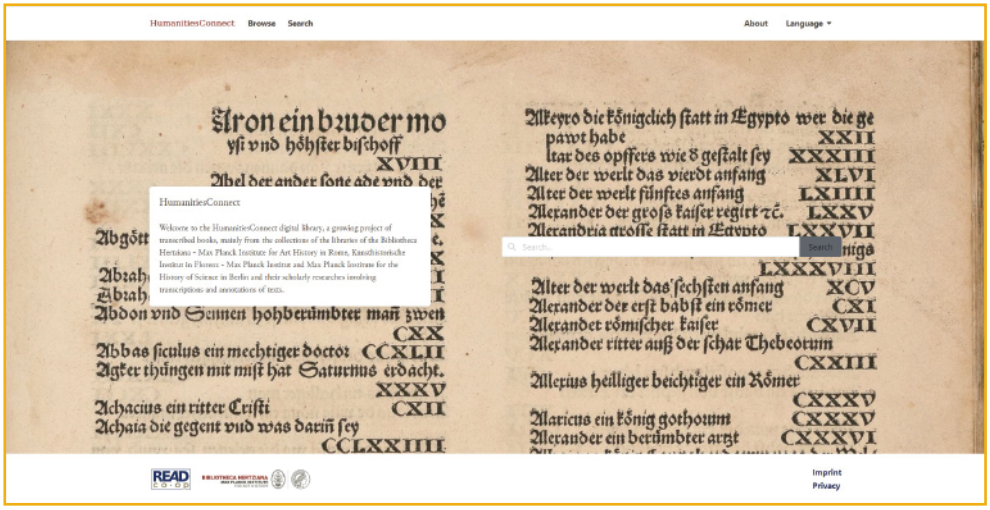

Figura 2. *Pagina iniziale della piattaforma Read&Search del progetto HumanitiesConnect*

[20] Phillip Benjamin Ströbel — Simon Clematide, *Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images*, Digital Humanities Conference 2019, e Phillip Benjamin Ströbel — Simon Clematide — Martin Volk, *How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR*, in: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, a cura di N. Calzolari et al., Marseille: European Language Resources Association, 2020, p. 3551–3559.

[21] <https://transkribus.eu/Transkribus>. Vedi Philip Kahle et al., *Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents*, in: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR),* IEEE, 2017, p. 19–24. Un altro esempio di uso delle macchine neurali applicate all'HTR è il progetto è eScriptorium <https://escriptorium.fr/>, che a differenza di Transkribus è open source e non interamente gestito lato server, quindi con la necessità di maggiori risorse tecnologiche e umane per l'implementazione ed il mantenimento a lungo termine.

[22] https://readcoop.eu/.

[23] Per questi due ultimi istituti l'indirizzo della biblioteca digitale è <https://dlc.mpg.de/>.

[24] <https://github.com/PRImA-Research-Lab/PAGE-XML>. Per approfondimenti: Stefan Pletschacher — Apostolos

Read&Search[25]. La piattaforma, in fase di *beta testing*, è disponibile online[26] ed è predisposta per espandersi con l'inclusione di nuove collezioni da altre biblioteche e istituti o curate da ricercatori (Fig. 2).
La piattaforma *Read&Search* nasce per la consultazione di manoscritti archivistici, con scarsa necessità di metadati bibliografici oltre alla segnatura della busta. Per adattarla alla consultazione e alla citazione di testi a stampa sono stati necessari dei cambiamenti a livello strutturale della piattaforma Transkribus, in particolare in relazione alla gestione dei metadati del documento. Al momento i metadati che sono associati alla singola pubblicazione sono solo una esportazione dei campi più comuni (autore, titolo, segnatura, anno e luogo di edizione, lingua) a partire dal catalogo bibliotecario in uso nell'istituto, Kubikat[27], alla data del conferimento delle digitalizzazioni (2020). Si tratta di un limite, dato che correzioni o revisioni del catalogo non sono aggiornate in tempo reale, che sarà probabilmente risolto nella prossima versione della piattaforma. I metadati stessi non sono visibili insieme al documento, ma possono essere utilizzati per filtrare le ricerche. Uno dei principali vantaggi di questa piattaforma è quello di poter cercare all'interno del testo del singolo documento, ma anche contemporaneamente di tutti i testi trascritti (Fig. 3). Per ovviare alle limitazioni date dall'ortografia antica, come per esempio l'uso del segno "u" consonantico per cui ora usiamo il segno "v", è stata introdotta la ricerca "sfocata"[28] che permette di stabilire un grado di "distanza" tra il termine cercato e quello trovato (per esempio cercando "colonna" posso trovare anche "colonne", "colomna", "columna" e così via), regolabile per ridurre il rumore di fondo, ovvero la presenza di un alto numero di risultati non coerenti (nell'esempio precedente troverei anche "corona" usando i parametri meno restrittivi).

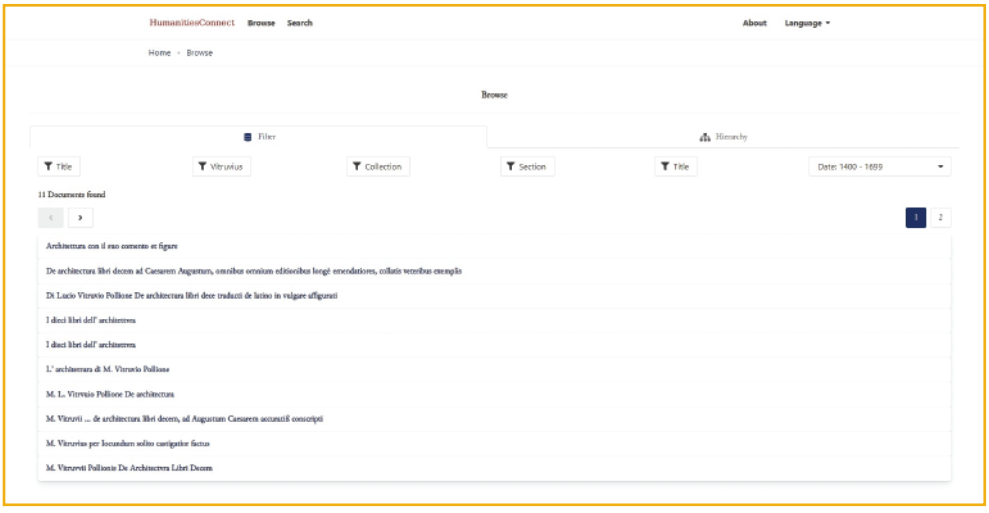

Figura 3. *Filtri di ricerca della piattaforma Read&Search*

Altri filtri riguardano la biblioteca di provenienza (Collection), autore (Author), titolo (Title), sezione (Section) e la data di pubblicazione. Al momento ogni volume può appartenere ad una sola sezione, ma in futuro saranno possibili più sezioni che fungeranno anche da percorsi tematici nella raccolta.

Antonacopoulos, *The PAGE (Page Analysis and Ground-Truth Elements) Format Framework*, in: *20th International Conference on Pattern Recognition,* IEEE, 2010, p. 257–260.

[25] https://readcoop.eu/readsearch/.

[26] https://transkribus.humanitiesconnect.pub.

[27] https://www.kubikat.org.

[28] Fuzzy search <https://it.wikipedia.org/wiki/Logica_fuzzy>.

Questo progetto non solo è una forma di riuso delle fonti digitali già esistenti, ma si presta per sua natura al riuso. A tutti i volumi nella raccolta è stato attribuito un numero identificativo che permette di raggiungere sempre univocamente il volume, che compare anche nel *manifest* IIIF all'interno della biblioteca digitale, in modo da semplificarne la citazione. Inoltre è facilmente citabile la singola pagina in modo univoco utilizzando direttamente l'URL della pagina visualizzata. Nello sviluppo in corso questo sistema di identificazione univoca delle pagine sarà reso ancora più esplicito e facilmente citabile. Anche la trascrizione espansa delle abbreviature semplifica la ricerca (per esempio "a a" diventa "anima"), ma per non limitare lo studio filologico, è stata anche sviluppata una versione della macchina neurale in grado di addestrare non solo il testo accessibile, ma il testo originario con annotata l'espansione[29] (Fig. 4).

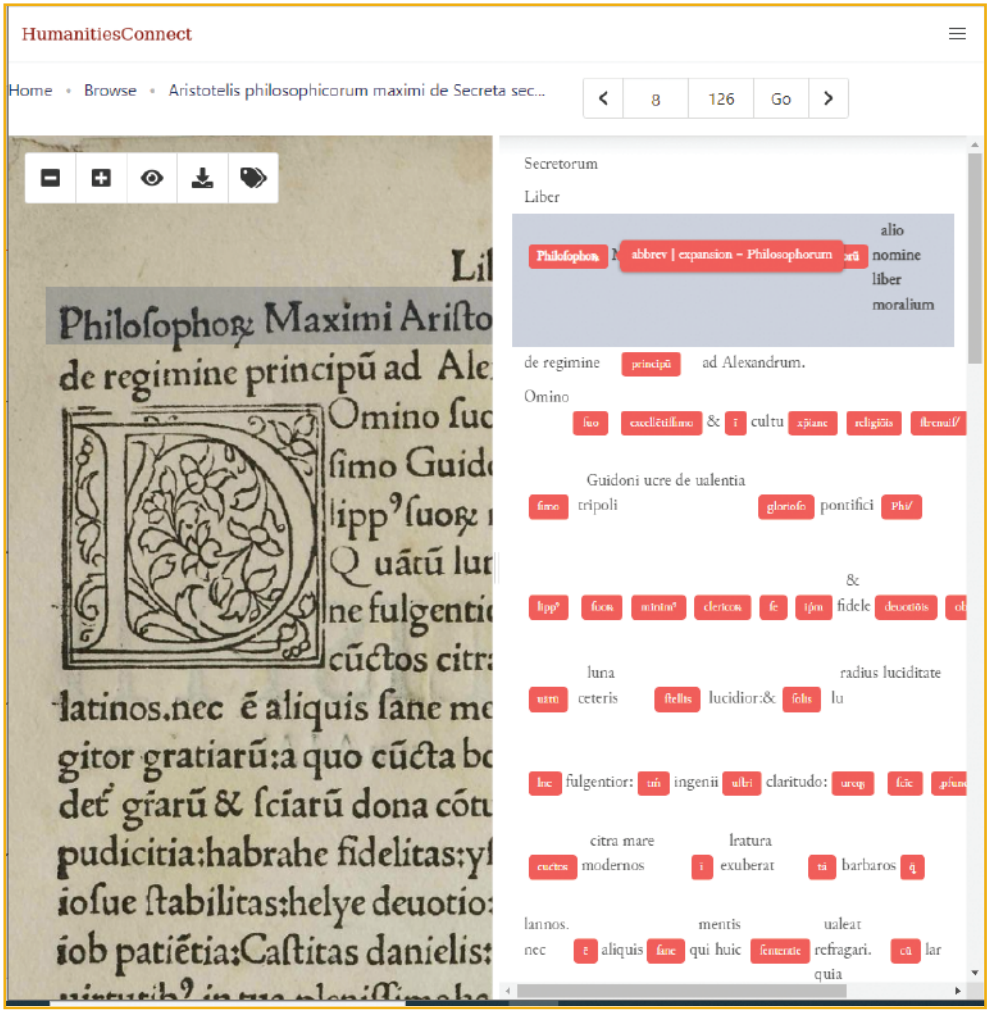

Figura 4. *Visualizzazione delle espansioni di abbreviature nella piattaforma Read&Search*

[29] In questo caso con termine annotazione si indica la presenza di tag nel contenuto XML che vanno ad indicare l'espansione delle abbreviature secondo specifiche compatibili con la codifica TEI XML (<abbr>a a</abbr> <expan>anima</expan>).

Tra le possibilità di riuso dei contenuti delle trascrizioni, la piattaforma Transkribus mette a disposizione anche un API[30] che garantisce l'accesso ai dati anche in modo automatico, per esempio per operazioni di analisi dei testi con gli strumenti della linguistica computazionale. Inoltre, sebbene le trascrizioni neurali siano generalmente sufficienti per la ricerca e la citazione, l'interfaccia di trascrizione (e in particolare la nuova versione su web[31]) prevede la possibilità di correggerle manualmente, sia da parte di ricercatori interessati a produrre edizioni critiche, che all'interno di progetti di *citizen science*[32] che permettono la collaborazione di più persone con diversi livelli di competenza, con il vantaggio che le correzioni possono essere utilizzate per addestrare modelli neurali di riconoscimento ancora più accurati.
La fase in corso di sviluppo della piattaforma prevede un miglioramento dell'usabilità dell'interfaccia *Read&Search*, soprattutto per quanto riguarda la connessione con il catalogo bibliografico attraverso l'uso di identificatori connessi tra loro tramite un *resolver* e i metadati bibliografici.
Dal punto di vista contenutistico oltre alle trascrizioni è possibile annotare la struttura del documento, identificando per esempio i titoli dei capitoli e dei paragrafi, i numeri di pagina e le note a piè di pagina. Anche il modello del layout di pagina del documento può essere addestrato, usando motori neurali che riconoscono le aree del testo e attribuiscono automaticamente i tag strutturali. Un documento così trascritto e segmentato può essere convertito in una vera e propria edizione digitale. Un esempio di questo riuso è quello del progetto della ristampa digitale dell'opera *Raphael in Early Modern Sources – 1483-1602* di John Shearman[33] (Fig. 5).

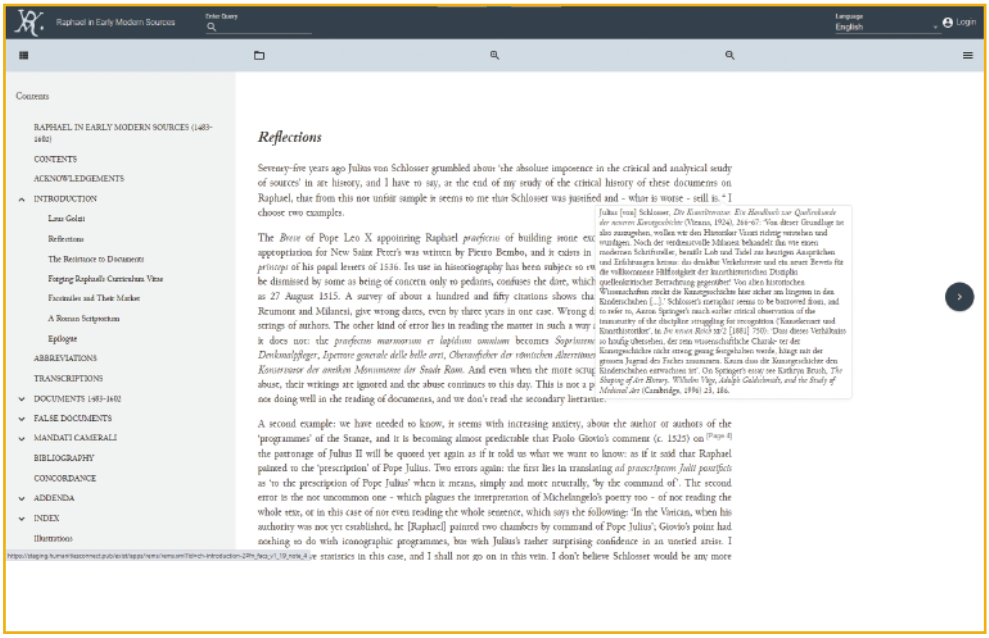

Figura 5. *Pagina dalla ristampa digitale REMS*

[30] Application programming interface. La documentazione di accesso si trova su <https://readcoop.eu/it/transkribus/docu/rest-api/> e richiede di essere abilitati almeno in lettura alla consultazione delle collezioni in formato nativo.
[31] https://app.transkribus.eu/.
[32] https://it.wikipedia.org/wiki/Citizen_science.
[33] John K. Shearman, *Raphael in early modern sources. (1483 - 1602)*, New Haven: Yale University Press, 2003.

Si tratta di due volumi di pubblicazione recente, la cui prima tiratura è andata esaurita ormai da alcuni anni e di cui non è stato possibile rintracciare i file digitali di stampa. Per questa ristampa digitale è stata fatta una nuova scansione delle oltre 1.700 pagine di testo. Il progetto prevedeva sin dall'origine una vera edizione digitale online in formato TEI XML[34] (Text Encoding Initiative) e non un semplice pdf con il testo riconosciuto dietro le immagini da scaricare. Questo implicava che oltre al testo era importante riconoscere anche la forma del contenuto, in particolare la struttura e le varianti grafiche dei caratteri, come corsivo o apice. Anche in questo caso Transkribus ha permesso sia la creazione di un modello di riconoscimento del testo che quello della struttura, grazie ad una macchina neurale per l'analisi del layout di pagina detta P2PaLA[35] integrata nella piattaforma. Data la tipologia del contenuto, costituito da un corpus di oltre mille documenti storici trascritti, commentati e dotati di informazioni di corredo, lo sforzo per la marcatura (*tagging*) manuale di un centinaio di documenti per creare il *ground truth* era ampiamente giustificato dai vantaggi derivati dalla creazione di un modello in grado di distinguere, per esempio, titolo, abstract, trascrizione, segnatura, commento critico, nota bibliografica e note a piè di pagina di ogni singolo documento. Dato che le macchine neurali usabili con Transkribus sono al momento in grado di analizzare solo le immagini e non il contenuto testuale, è stata necessaria una correzione manuale delle marcature dei paragrafi assegnate automaticamente dalla intelligenza artificiale – in particolare per identificare i paragrafi che continuavano attraverso le pagine – che è durata circa tre mesi[36]. In una edizione tradizionale con trascrizione OCR la revisione dei contenuti e della struttura sarebbe stata interamente manuale e avrebbe richiesto ancora più tempo. La marcatura del testo e della struttura è un passaggio fondamentale per poter convertire il contenuto in TEI XML. È stato infatti messo a punto un flusso di lavoro per ottenere il documento TEI strutturato – con suddivisione dei capitoli, riallineamento delle note a piè di pagina con i rispettivi numeri di riferimento nel testo, formattazione corretta degli stili di carattere - direttamente dal PAGE XML della pagina tramite trasformazioni XSLT, a partire da una versione personalizzata di PAGE2TEI[37]. La piattaforma di visualizzazione del testo è disponibile online in anteprima su una interfaccia di lettura basata su TEI Publisher[38]. Il vantaggio principale di questa trasformazione è la possibilità di sfruttare la ricerca a testo intero, selezionando però l'ambito della struttura (solo le trascrizioni o solo il commento) e utilizzando stringhe avanzate o ricerca di prossimità (fuzzy search), cosa che non sarebbe stata possibile con un PDF reso ricercabile con OCR.

Nello sviluppo in corso della piattaforma, la possibilità di attribuire ai paragrafi un valore strutturale permetterà di filtrare la ricerca distinguendo, all'interno della stessa piattaforma Read&Search, se il testo cercato è parte del corpo del testo, o di una nota di commento, o infine, di un titolo. Inoltre, standardiz-

[34] https://tei-c.org/

[35] Page to PAGE Layout Analysis, sviluppato dal progetto <https://github.com/lquirosd/P2PaLA>. Sui dettagli tecnici di questo progetto v. anche Elisa Bastianello — Reto Baumgartner, *L'applicazione del riconoscimento testi neurale per la realizzazione di ristampe digitali*, in: *La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD*, a cura di E. Carbé et al., Siena: Associazione per l'Informatica Umanistica e la Cultura Digitale, 2023, p. 15–23.

[36] Vorrei qui ringraziare Viviana Nocerino, Iolanda Pagano e Andrea Pecorella per aver brillantemente completato questo tedioso compito.

[37] Dario Kampkaspar <https://github.com/dariok/page2tei>. Il processo di conversione è stato messo a punto in collaborazione con Reto Baumgartner ed è disponibile open source <https://github.com/biblhertz/trans2tei>.

[38] <https://www.teipublisher.com>. La ristampa digitale è disponibile in formato beta all'indirizzo <http://rems.humanitiesconnect.pub>. Per una panoramica sull'uso di TEI Publisher nelle edizioni digitali rimando a Elisa Bastianello, *Dalla digitalizzazione all'edizione digitale: i progetti Digital Publishing della Bibliotheca Hertziana*, in: *Digital Humanities 2022. Per un confronto interdisciplinare tra saperi umanistici a 30 anni dalla nascita del World Wide Web*, a cura di M. Di Maro, V. Merola, T. Nocita, Roma: L'Erma di Bretschneider, 2023, p. 143–160.

zando i nomi degli elementi strutturali rispetto a specifiche funzioni sarà possibile migliorare l'esportazione di queste trascrizioni in formato TEI XML con dei modelli condivisi.
La tecnologia del riconoscimento neurale dei testi (a stampa e manoscritti) può essere facilmente applicata alle collezioni di altri istituti di conservazione: la piattaforma è in grado di importare direttamente i volumi in formato IIIF resi disponibili per applicare i modelli esistenti e ottenere, con poco sforzo, un riuso di miliardi di immagini di documenti già scansionate ma non ricercabili o riutilizzabili. La stessa piattaforma di lettura *Read&Search* è modulare e permette di includere raccolte tematiche, come nel caso della raccolta di trascrizioni del progetto Rara Magnetica[39], o collezioni di biblioteche interessate a partecipare al progetto. Sebbene esista un numero crescente di modelli pubblici con cui effettuare le trascrizioni automaticamente, e in particolare per la stampa antica e moderna nelle principali lingue europee esistono numerosi modelli affidabili[40], per i testi manoscritti in generale vale la pena addestrare un modello specifico per lingua, epoca e area geografica di riferimento, dato che i grandi modelli hanno una minore accuratezza a livello di trascrizione dei testi.

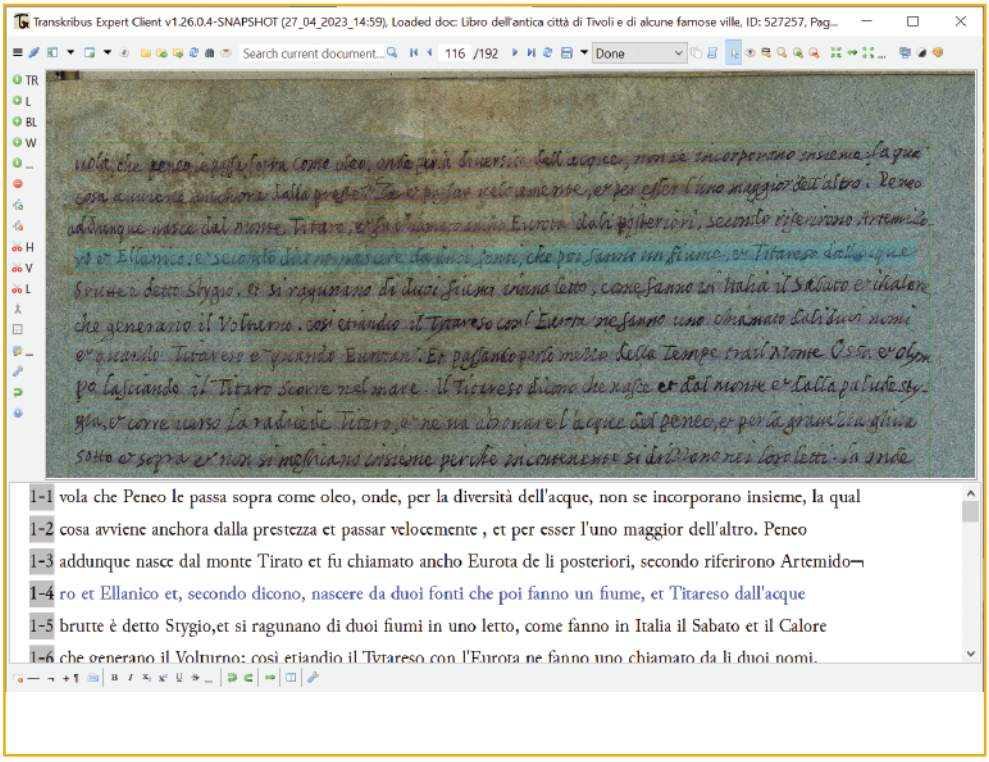

Figura 6. *Esempio di trascrizione in Transkribus di una pagina dell'Enciclopedia del Mondo Antico di Pirro Ligorio*

[39] Il progetto Rara Magnetica <https://www.raramagnetica.de/>, curato dal dott. Christoph Sanders, è nato all'interno del gruppo di ricerca *Visualizing Science in media revolutions*, diretto da Sietzke Fransen <https://www.biblhertz.it/3262783/research-group-fransen>.

[40] Per esempio Transkribus Team, *Transkribus Print M1 [ID 39995]*. Transkribus - Pylaia text recognition, 2022: <https://readcoop.eu/model/transkribus-print-multi-language-dutch-german-english-finnish-french-swedish-etc/>; Stefan Zathammer, *Noscemus GM 5 [ID 37664]*. Transkribus - Pylaia text recognition, 2022: <https://www.uibk.ac.at/projects/noscemus/>.

Un esempio di questa possibilità è quello relativo al riuso delle immagini dei manoscritti dell'architetto e antiquario rinascimentale Pirro Ligorio resi disponibili online dall'Archivio di Stato di Torino[41]. Dall'autunno 2021 è in corso un progetto congiunto con l'Universität Freiburg, l'Università di Napoli, l'Université de Rouen-Normandie, lo stesso Archivio di Stato di Torino e altri istituti per creare un'edizione digitale dei manoscritti dell'Enciclopedia del Mondo Antico[42]. In questo caso la piattaforma Transkribus è stata utilizzata dal team per impostare la trascrizione del testo e l'annotazione dei contenuti. In particolare, grazie all'opera di Giorgia Agostini, è stato creato e reso pubblico il primo modello neurale per il riconoscimento della mano di Ligorio[43], che verrà utilizzato per semplificare la fase di prima trascrizione dei volumi (Fig. 6). Le trascrizioni automatiche hanno come scopo primario rendere accessibili dei contenuti, e saranno rese disponibili su una sezione dedicata della piattaforma *Read&Search*, mentre specifici approfondimenti curati dai ricercatori del progetto saranno esportati in formato TEI XML.

In definitiva l'adozione di protocolli condivisi dalla comunità scientifica come IIIF per la interscambiabilità delle immagini, PAGE e TEI XML per la codifica delle trascrizioni e piattaforme open source sostenute da una grande comunità, insieme allo sviluppo sempre maggiore di macchine neurali e piattaforme basate su progetti di intelligenza artificiale in grado di rispondere a specifiche richieste della ricerca scientifica in ambito umanistico, si pone come la base di una nuova generazione di progetti digitali, che non solo riusano quanto prodotto negli ultimi decenni, ma ne semplificano la ricerca, l'accessibilità e il riuso, garantendone la permanenza a lungo termine.

Book digitizations are enjoying a tremendous success in recent years, making it possible to overcome the limitations on access to the rarest physical books. But, upon closer inspection, a digital reproduction of the pages of a book is not necessarily accessible and findable. For this reason, the Bibliotheca Hertziana, among other institutions, is investing lot of resources in transcribing the content of digitizations, with the support of new technologies such as neural text recognition. Crucial points include the use of shared standards such as IIIF and TEI XML and open source platforms such as TEI Publisher to ensure long-term accessibility and preservation of digital content.

[41] I volumi sono disponibili nella biblioteca antica digitale all'indirizzo: <https://archiviodistatotorino.beniculturali.it/dbadd/dlredi.php?i=23>.

[42] Il progetto nasce dalla conferenza *Ligorio Digitale. Idee e prospettive per un'edizione digitale dei manoscritti di Pirro Ligorio*, a cura di F. Rausa, A. Schreurs-Morét e G. Vagenheim, Roma, Bibliotheca Hertziana – Istituto Max Planck per la storia dell'arte, 26-27 ottobre 2020 e lo stato di avanzamento è stato recentemente oggetto di una specifica sezione del convegno *Pirro Ligorio e l'Italia. Antichità locali e cultura antiquaria*, a cura di M. Guarente e A. Di Tuccio, Università degli studi di Napoli Federico II, 22-23 maggio 2023.

[43] Giorgia Agostini, *Ligorio 0.3 [ID 42105]*. Transkribus - Pylaia text recognition, 2022: <https://readcoop.eu/model/ligorio-0-3/>. Il sito del progetto di dottorato è <https://lidiws-limes.cfs.unipi.it/>.

L'ultima consultazione dei siti web è avvenuta nel mese di dicembre 2023