Progetto di un Osservatorio Permanente sulle Edizioni Digitali di autori Italiani (OPEDIt).

Prime indagini sulle pratiche di digitalizzazione e sull'autorevolezza dell'edizione di testi letterari italiani in formato elettronico

Michelangelo Zaccarello

I. Premessa

A partire dagli anni Novanta, le nuove opportunità di accesso ai testi hanno propiziato l'allestimento di archivi e biblioteche digitali che fanno della grande quantità di testi disponibili il loro punto di forza (ad es. *Project Gutenberg*; *Archive.org*; *Wikisource*); la gratuità di tali iniziative non deve ingannare circa la natura commerciale dei siti coinvolti, il cui profitto è dato dalla pubblicità ed è quindi proporzionale al numero dei contatti (McGann 2014). Non è un caso che la maggior parte dei testi caricati su tali piattaforme sia libero da ogni diritto d'autore, a conferma di una filosofia d'uso che potrebbe definirsi *low risk-low gain*: poiché il guadagno non deriva direttamente dai diritti su singole opere ma dalla mole di uten-

ti della piattaforma (attraverso la pubblicità derivante dai molti contatti), occorre (a) immettere in rete il massimo numero di testi battendo sul tempo le realtà concorrenti; (b) coinvolgere il massimo numero di utenti, facendo leva sulla loro limitata consapevolezza. Con finalità più vicine alla ricerca scientifica sui testi, altri siti web hanno puntato sulla consultazione integrale o selettiva dei *corpora* (*Letteratura italiana Zanichelli*; *Biblioteca italiana telematica*): su tali piattaforme, il testo può essere tanto analizzato (indicizzato, interrogato, esportato verso diverse piattaforme), quanto rappresentato, per comparire in varie forme di **output* ritenute utili (da leggere, pubblicare, stampare).

Nella maggior parte delle iniziative citate, l'immissione dei testi in rete non avviene mediante immissione/revisione manuale con codifica testuale, modalità che è propria delle più qualificate iniziative di *Digital Humanities* (si pensi alla banca dati TLIO/Tesoro della Lingua italiana delle Origini sviluppato dall'Opera del Vocabolario), bensì mediante scansione con *software* per il riconoscimento ottico dei caratteri grafici (OCR = *Optical Character Recognition*); questi impiegano nella maggior parte dei casi una tecnologia che analizza il contrasto fra le linee scure dei caratteri e il fondo bianco (*stroke edge technology*), poi confronta l'immagine che ne risulta a set predefiniti di possibili «letture» (generati per mezzo di dizionari integrati). In presenza di fonti cartacee invecchiate o danneggiate, o di stampa con caratteri di forma non convenzionale, i margini d'errore degli OCR possono diventare un vero problema per l'affidabilità del testo (Zaccarello 2017, pp. 154-155).

II. La digitalizzazione a mezzo OCR

Spesso ignorata o sottovalutata negli studi umanistici, la mediazione dei software OCR ha un impatto estremamente vasto e importante sulla qualità dei testi che vengono digitalizzati, specie nel contesto di un fenomeno che nei parametri quantitativi i suoi principali obiettivi: si cerca di digitalizzare il massimo numero di testi, impiegando anche scansioni già disponibili sul Web, e di gestire tutte le fasi dell'archiviazione in modo automatizzato, limitando al massimo la revisione manuale dei testi e pubblicando le opere con informazioni (metadati) lacunosi, scorretti o persino assenti. In un sistema altamente codificato come la pagina scritta, è intuitivo che

la stessa migrazione di opere scritte dal contesto stampato alla 'liquidità' del testo espone a notevoli rischi nella navigazione del testo, specie perché il prodotto 'grezzo' della scansione viene spesso pubblicato senza specifica revisione, e gli *e-text* in rete – spesso ripresi in altri archivi – si diffondono molto rapidamente presso varie fasce di utenza (si vedano i sondaggi di Kichuk 2015, che ritiene «the frequency of disorienting transformations resulting from OCR technology as significant and a reflection of an alarmingly casual indifference to accuracy and authenticity», p. 67).

Il rapido progresso informatico non promette di risolvere questi aspetti, perché la ricerca sui software OCR risponde a logiche e finalità spesso ben diverse da quelle degli studi umanistici: più che per rispondere a requisiti di esattezza, gran parte dei software OCR – usati anche per la scansione e l'archiviazione di documenti – sono sviluppati e migliorati alla ricerca di una maggiore velocità con cui le immagini possono essere processate, e della lettura di alfabeti non latini (arabo, russo, cinese ecc.). Anche gli indici di affidabilità sono intesi in modo diverso: quelli normalmente disponibili sono infatti tarati su moderna documentazione cartacea, stampata con font attuali, mentre le fonti cartacee degli *e-texts* sono spesso edizioni invecchiate e/o copie ingiallite o danneggiate dal tempo (ad esempio, Tanner – Muñoz – Hemy Ros 2009). Inoltre, questo stesso studio ha dimostrato come il dato significativo per l'affidabilità degli *e-text* non sia la quantità di caratteri erronei per pagina, normalmente valutata in ambito informatico, ma la quantità di parole erronee, ambito in cui l'accuratezza risulta sempre inferiore e che – naturalmente – ha importanti ripercussioni sui risultati di ricerche e indicizzazioni eseguite sul testo digitalizzato. D'altra parte, quest'ultima tipologia d'impiego – con accesso puntuale e parcellizzato al testo – è responsabile del fatto che la massima parte delle scorrettezze presenti negli *e-texts* vengono riconosciute molto difficilmente, perché al testo si arriva attraverso i motori di ricerca, con risultati che molto spesso non vengono allargati oltre il loro contesto minimo (Vaidhyanathan 2011).

Anche quando non possano essere scaricati interamente, l'interrogabilità dei testi sui motori di ricerca presuppone non solo la digitalizzazione delle fonti cartacee mediante OCR, ma la loro copia integrale sui relativi server. Poiché – com'è noto – le leggi internazionali sul copyright proteggono le opere dalla riproduzione integrale, eseguita con qualunque mezzo, il fatto che tutti i *digital repositories*, *libraries* e *data bases* si fon-

DOI: https://doi.org/10.13130/2499-6637/9491

dino al contrario sulla riproduzione integrale delle opere configura una colossale infrazione del copyright, dato che nella vastissima scala della *mass digitization* è impensabile non solo ottenere i diritti, ma identificarne gli eventuali beneficiari: negli USA, sono già state minacciate colossali class actions, e nel 2011 un tentativo di accordo extra-giudiziale fra Google e il Sindacato degli autori e degli editori americani è stato rigettato dal Giudice Chin (http://www.chronicle.com/article/Judge-Rejects-Settlement-in/126864/).

Il comune denominatore delle diverse anomalie fin qui sommariamente indicate è che la pubblicazione su web non sembra parificata alle forme tradizionali di editoria, né indirizzata a pratiche consuete di lettura; come è stato evidenziato da giuristi specializzati, la *mass digitization* si presenta come un'archiviazione di massa realizzata per l'accesso computerizzato; essa non è finalizzata ad offrire libri alla lettura, né principalmente intesa per sostenere le metodologie tradizionali della ricerca: «In this context, computers discover information that human intelligence can never extract, such as quantitative and qualitative data on 'cultural trends' over centuries and across languages, migration of 'ideas' from one place to another, evolutions of linguistic phenomena, influences of one author on another, and vice versa. In turn, this information is linked back to the individual item—book, image, sound recording, video—in order to create an empowered reading environment, from which data about reading habits and, in general, behaviours associated with the experience of content are extracted and processed. It is no surprise to read the following reported words of an anonymous Google engineer: '*we're not scanning all those books to be read by people. We're scanning them to be read by Artificial Intelligence*'» (Borghi -Karapapa 2013, p. 14, mio il corsivo).

III. Uno sguardo dall'Italia

A fronte di questa problematica situazione, la massiccia fruizione degli *e-texts* offerti da varie piattaforme avviene in modo sostanzialmente inconsapevole, con testi scaricati e indirizzati a vari usi (lettura, consultazione, persino insegnamento scolastico e universitario) senza il minimo 'controllo di qualità' e, data la contrazione dei tempi di fruizione, senza il contraddittorio di altre versioni o il riscontro sulla fonte cartacea da cui il testo

proviene. Per il contesto italiano, manca una tradizione di studi e sondaggi paragonabile a quella che – in tempi comunque recente – si è resa disponibile per la pubblicazione di testi in lingua inglese; dato l'uso dilagante degli *e-texts* in una quantità di ambiti e presso varie tipologie di lettori, questo ritardo sarebbe di per sé già un motivo importante per avviare alcune verifiche, necessariamente parziali, sull'effettiva qualità dei testi circolanti.

Una prima serie di sondaggi, svolti in varie tesi di laurea triennale e magistrale, ha evidenziato problemi di ampia portata, a partire dalle fonti cartacee oggetto di digitalizzazione: ove dichiarata, spesso quest'ultima non coincide con l'edizione più autorevole o accreditata, ma con quella più facilmente reperibile e libera da vincoli legali di riproduzione: nella migliore delle ipotesi, ciò significa perpetuare la *vulgata* moderna dell'opera, ma non è raro il caso in cui il testo proviene da edizioni ben più invecchiate, specie quelle la cui riproduzione digitale è liberamente scaricabile in rete da siti come *Google Books*; in bianco e nero o a colori, tali scansioni aggiungono ai difetti del testo quelli conseguenti al tempo e all'usura, con la forma obsoleta dei caratteri, la carta ingiallita o macchiata e persino vari segni a penna che creano enormi problemi ai già fallaci *software* OCR. Come si è detto, questi ultimi operano con dizionari integrati che tendono a ricondurre le forme arcaiche e le varianti grafiche alla equivalente forma moderna, con massiccia perdita di dati caratterizzanti la lingua antica e casi frequenti di fraintendimento sostanziale.

Tutto ciò ha l'effetto combinato che da un lato i testi più accessibili (e gratuiti) sono quelli di peggiore qualità, dall'altro la loro stessa diffusione li promuove in certo modo a standard di riferimento, almeno per l'uso non specializzato. Il lettore, del resto, sviluppa un accesso sempre più frettoloso e superficiale ai testi, senza interrogarsi sulla relativa qualità e affidabilità (è il *lettore Google* descritto da Italia 2016). Pubblicate con metadati lacunosi, imprecisi o affatto mancanti, gli *e-texts* sono in massima parte privi di quel contesto di legittimazione che, nell'editoria cartacea, è il risultato di secoli di tradizione metodologica, dibattito scientifico e ricezione di lettura. È infatti ben noto che, nell'editoria cartacea, l'autorevolezza di un testo è sancita dalle principali collane che pubblicano edizioni critiche, con particolare riferimento ad autori classici della letteratura italiana: sottoscritte per abbonamento dalle biblioteche italiane e straniere, tali pubblicazioni godono di una risonanza, per così dire, istituzionale, che si concretizza in

DOI: https://doi.org/10.13130/2499-6637/9491

recensioni, commenti e citazioni, elementi che equivalgono a ciò che per i settori scientifici è la valutazione bibliometrica, la quantificazione dell'*impact factor* di una ricerca.

Da simili fenomeni, osservabili in scala vastissima, scaturisce una prima considerazione generale: l'esponenziale proliferazione di versioni digitalizzate (*digitized*) ha immesso un'enorme quantità di inesattezze nelle opere letterarie antiche, derivanti da (a) l'inadeguatezza della fonte cartacea prescelta, talora non dichiarata; (b) le condizioni materiali della copia sulla quale è stata condotta la scansione; (c) gli ampi margini di errori dei *software* OCR. A tutto ciò si aggiunge il funzionamento di questi ultimi, che implica una sistematica uniformazione e modernizzazione delle forme che non corrispondono all'uso registrato nei principali dizionari in uso. Ciò giustifica da un lato la forte persistenza delle tradizionali edizioni critiche nell'uso specialistico, dall'altro l'incidenza ancora limitata delle citazioni di versioni digitali negli studi umanistici in genere, dovuta alla generale diffidenza degli studiosi (Suković 2009). Nell'accertata necessità di continuare a pubblicare i nostri autori antichi secondo criteri di fedeltà almeno sostanziale ai documenti, è dunque lecito affermare che la moltiplicazione esponenziale dei testi disponibili sul *World Wide Web* influisce sul contesto italiano in misura maggiore che in altri contesti nazionali.

IV. L'utilità di un Osservatorio permanente sulle pratiche editoriali scientifiche e sull'autorevolezza dell'edizione di testi letterari italiani nel contesto digitale.

Pur nella necessaria concisione, le brevi considerazioni sopra esposte delineano un quadro attuale di fruizione dei Classici della letteratura italiana con tratti altamente problematici: il minore utilizzo delle biblioteche a fronte della comoda consultazione *on line*, il costo elevato delle edizioni critiche più autorevoli, il tasso – limitato o minimo – con cui vengono divulgate le più aggiornate acquisizioni della filologia e la conseguente scarsa consapevolezza generale della problematicità implicita nella fissazione di un testo 'autorevole' (e della relativa diffusione nel mercato editoriale) sono fattori che possono vanificare decenni di cure filologiche rivolte alle opere più importanti della nostra letteratura. Da tempo sto raccogliendo, con l'aiuto di alcuni laureandi dell'Ateneo veronese, esempi tratti da opere cru-

ciali delle Tre Corone (per Dante, la *Commedia* e il *Convivio*; per Petrarca, il *Canzoniere* e i *Triumphi*; per Boccaccio, la *Fiammetta* e il *Decameron*): in questa sede, mi limito ad alcuni esempi tratti da quest'ultimo.

Nella popolare versione PDF di *LiberLiber*, il *Decameron* è tratto da un'edizione (Branca 1951) precedente al riconoscimento dell'autografo berlinese (Hamilton 90) e pertanto basata sul ben noto *Codice Martelli* (Laur. XLII 1): ne deriva una miriade di varianti formali spesso improntate alla modernizzazione di tratti arcaici (*adormentò* / *addormentò*; *prencipe* / *principe*; *sagliendo* / *salendo*; *camiscia* / *camicia*: esempi tratti da IV 1 e 2), ma anche gravi equivoci sostanziali: il raro aggettivo *cassesi* (dall'arabo *qasīs* 'prete cristiano') è risolto in uno strano *c'ha Ascesi*; la mediazione OCR è poi evidente in vari casi di errata segmentazione *quali* > *qua li*; *fare* > *fa re* (tutti gli esempi da IV 2-3). Rivolta a un pubblico più esigente, la versione che compare nell'archivio digitale del sito *Decameron Web* di Brown University, codificata in XML, è invece basata sulla più autorevole edizione critica allestita da Branca 1992, ma non vi mancano errori anch'essi probabilmente introdotti in fase di scansione OCR: nella famosa novella boccacciana di madonna Lisetta, il soprannome *Baderla* attribuito alla protagonista perde la maiuscola e diventa un verbo, condizionale in forma arcaica, *baderia* (*Decameron*, IV 2 24).

Alcuni sondaggi su opere petrarchesche e dantesche hanno evidenziato scenari analoghi: è emblematico in tal senso il caso del verso conclusivo del *Triumphus Cupidinis* di Petrarca (IV 166), «che *il piè va innanzi* e l'occhio torna adietro», in corrispondenza del quale la versione in formato PDF del sito *LiberLiber* ha un assurdo *che il più va innanzi*. Conseguenze serie può avere anche la sistematica omissione dei segni diacritici, ad esempio per indicare l'integrazione di un salto meccanico nel testo base: «cioè nel verso ch'è lo secondo di *questa parte* [e lo terzo della canzone; e poi quello che dicea la parte che vincea, cioè nel verso ch'è lo terzo di *questa parte*] e lo quarto della canzone» (Ageno 1990, III, p. 23).

Di fronte a tali fenomeni, di vastissima incidenza, è opportuno – anzi urgente – che organi deputati alla riflessione metodologica sui testi e all'allestimento di edizioni critiche elaborino delle linee guida per la diffusione in rete dei Classici della nostra letteratura. Nell'ovvia impossibilità di revocare la circolazione in rete – ormai decennale – di testi gravemente corrotti, sarebbe inoltre opportuno che tali organi definissero i requisiti

DOI: https://doi.org/10.13130/2499-6637/9491

minimi di affidabilità per le risorse digitali, mettendo in guardia gli utenti dal rischio dell'apparente 'autorevolezza' di testi di grande diffusione. Per i suoi fini statutari, un ruolo di primo piano nel promuovere tale consapevolezza dovrebbe essere assunta dalla nostra Commissione per i Testi di Lingua, fondata nel 1860 appunto col nobile «fine di reperire e diffondere, con la pubblicazione, le opere degli scrittori italiani del Trecento e del Quattrocento» (http://www.commissionetestidilingua.it); per gli stessi motivi, e per le forti implicazioni linguistiche delle dinamiche sopra richiamate, è desiderabile che un ruolo altrettanto importante di elaborazione e promozione di *standard* condivisi per le edizioni digitali di testi italiani antichi sia svolto dall'*Opera del Vocabolario Italiano* (http://www.ovi.cnr.it), il più importante centro di ricerca lessicografico sull'italiano antico, basato sull'immissione digitale diretta di testi antichi, in versione codificata e riveduta (non di rado migliorata) sulle fonti cartacee.

In sintesi, sarebbe opportuno unire i nostri sforzi in un Osservatorio integrato permanente: sull'esempio di quanto già istituito a Milano da parte del gruppo di studio sulle edizioni critiche, un analogo gruppo pisano-bolognese potrebbe svolgere un'analoga 'mission' su altre forme di pubblicazione digitale, che dagli standard dell'ecdotica tradizionale dovrebbero partire ma che si indirizzano a tipologie di fruizione e pratiche di lettura o di ricerca più generali e non necessariamente 'scientifiche': è infatti in questo ambito che la svolta digitale ha introdotto i mutamenti più radicali, attraverso i meccanismi già sommariamente elencati della *mass digitization*. Si tratta di un forte mutamento di prospettiva, dato che l'immissione dei testi appare sostanzialmente indirizzata alla gestione dati: come ancora sottolineano Borghi e Karapapa 2013, «While in both the analogue and the digital world works have been used primarily *as works* —namely as expressions addressed by the author to the public— in the mass digital environment works are primarily used *as data*» (p. 16).

Una prima rete di sondaggi può essere avviata formalizzando collaborazioni da tempo in atto, e in particolare con il coinvolgimento di: Università di Grenoble Alpes (E. Pierazzo); Università di Losanna (L. Tomasin); Mazarikova Univerzita di Brno (P. Divizia); Opera del Vocabolario italiano (L. Leonardi, P. Larson); Università di Pisa (M. Tavoni, M. Zaccarello). Il consorzio risultante è soggetto idoneo alla partecipazione a bandi competitivi, italiani o internazionali, mentre l'adesione dei singoli Atenei s'in-

tende naturalmente senza oneri, con l'obiettivo scientifico di perseguire le seguenti finalità primarie:

a) Nel perimetro operativo dell'Osservatorio e dei singoli Soggetti aderenti, avviare ricerche sul campo sull'ampia tematica delle edizioni digitalizzate (mediante tesi di laurea e dottorato, progetti di ricerca, pubblicazioni scientifiche), per svolgere verifiche sui testi reperibili in rete di opere dei primi secoli della Letteratura italiana, verificarne l'affidabilità e promuoverne possibili revisioni.

b) Con particolare riferimento a Dante, Petrarca e Boccaccio (ma con la prospettiva di coinvolgere i Classici italiani fino all'Ottocento), salvaguardare l'integrità e correttezza delle opere della Letteratura italiana dalle insidie della digitalizzazione e della diffusione *on line*, promuovendo verifiche a campione (ad esempio, mediante tesi di laurea) e/o gruppi di ricerca che valutino sia la fedeltà degli *e-texts* alle rispettive fonti cartacee, sia la generale validità – formale e sostanziale – del testo in essi contenuto.

c) Diffondere parametri, criteri e modelli per la definizione di 'buone pratiche' editoriali, con particolare riferimento alla circolazione di testi in rete, e alla scelta delle relative fonti. Si tratta di un'iniziativa che – nel contesto di lingua inglese – è stata da tempo intrapresa ad opera della prestigiosa Modern Language Association, e che ha dato origine a un lucido documento che ha presto assunto un elevato valore paradigmatico, almeno fra gli addetti ai lavori, il *White Paper of the Modern Language Association's Committee on Scholarly Editions*.

d) Elaborare un protocollo di 'certificazione della qualità' dei testi digitalizzati: articolato in più livelli, esso dovrebbe assegnare a ogni risorsa digitale esaminata un chiaro indice di affidabilità (*reliability index*), sul modello di quanto sta avvenendo nel contesto anglo-americano. Senza alcuna pretesa di imporre standard qualitativi, tale protocollo verrebbe incontro alle esigenze dei molti lettori avvertiti ma non filologicamente preparati, che desiderano un accesso consapevole a testi ragionevolmente affidabili nella loro forma e sostanza.

e) Con il consorzio di cui al punto (a), partecipare a bandi competitivi, italiani o europei, che possano reperire le risorse – umane e finanziarie – necessarie a promuovere e diffondere quanto elencato ai punti a-d su una scala proporzionata alla vastità dei fenomeni sopra descritti, e possibilmente a istituire un gruppo permanente di consulenza filologica 'al pubblico'

DOI: https://doi.org/10.13130/2499-6637/9491

sul Web, sull'esempio di quanto sviluppato con notevole successo, anche mediatico, dall'Accademia della Crusca col servizio *La Crusca per voi*.

michelangelo.zaccarello@univr.it

Riferimenti bibliografici

Dante Alighieri, *Convivio*, a cura di Franca Brambilla Ageno, 3 voll., Firenze, Le Lettere (Società Dantesca italiana. Edizione Nazionale), 1995.

Barbara Bordalejo, *The Texts We see, the Works We Imagine. The Shift of Focus of Textual Scholarship in the Digital Age*, «Ecdotica», VII, 2010, pp. 64-76.

Maurizio Borghi e Stavroula Karapapa, *Copyright and Mass Digitization*, Oxford, Oxford University Press, 2013.

Giovanni Boccaccio, *Decameron*, a cura di Vittore Branca, Firenze, Le Monnier, 1951.

Giovanni Boccaccio, *Decameron*, a cura di Vittore Branca, Torino, Einaudi, 1992.

Paola Italia, *Il lettore Google*, «PEML. Prassi ecdotiche della modernità letteraria», 1, 2016, pp. 1-12.

Diana Kichuk, *Loose, Falling Characters and Sentences: The Persistence of the OCR Problem in Digital Repository E-Books*, «Portal: Libraries and the Academy», 15/1, 2015, pp. 59-91.

Jerome J. McGann, *A New Republic of Letters. Memory and Scholarship in the Age of Digital Reproduction*, Cambridge (Mass.), Harvard University Press, 2014.

Suzana Suković, *References to e-texts in academic publications*, «Journal of Documentation», 65/6, 2009, pp. 997-1015.

Simon Tanner, Trevor Muñoz e Pich Hemy Ros, *Measuring Mass Text Digitization Quality and Usefulness*, «D-Lib Magazine», a. 15, nn. 7/8, 2009, al link: http://www.dlib.org/dlib/july09/munoz/07munoz.html

Siva Vaidhyanathan, *The Googlization of Everything (and Why We Should Worry)*, Berkeley, University of California Press, 2011.

White paper = Considering the Scholarly Edition in the Digital Age: A White

Paper of the Modern Language Association's Committee on Scholarly Editions, memorandum redatto dal *Committee for Scholarly Editions della Modern Language Association*, al link https://scholarlyeditions.mla.hcommons.org/2015/09/02/cse-white-paper.

Michelangelo Zaccarello, *L'edizione critica del testo letterario. Primo corso di filologia italiana*, Milano, Mondadori Education, 2017.

DOI: https://doi.org/10.13130/2499-6637/9491