Umanistica Digitale – ISSN: 2532-8816 – n. 18, 2024
T.L. Gil – *Sui big data e il nepotivismo digitale nella ricerca storica*
DOI: http://doi.org/10.6092/issn.2532-8816/20365

Sui big data e il neopositivismo digitale nella ricerca storica[1]

Tiago Luís Gil

Universidade de Brasília. Brasília, Brasil
tiagoluisgil@gmail.com

Abstract

Il saggio discute l'avanzamento delle iniziative digitali nella ricerca storica, sottolineando in particolare l'aumento del numero di grandi *repository* e di collezioni archivistiche *online*. Tale scenario è visto come parte di una crescente attenzione ai *big data* che deve essere ancora discussa nella comunità disciplinare. Vengono presentati alcuni casi di progetti con un forte approccio empirista (come *Transkribus* e *Time Machine*) che sembrano prefigurare una presunta irrilevanza del metodo storico tramite il rilievo dato alla disintermediazione. L'articolo auspica di fornire spunti per il dibattito, indicando la formazione digitale come la condizione imprescindibile per orientare gli storici di fronte a un mare di incertezze e di algoritmi.

Parole chiave: big data, neopositivismo, storia digitale, strumenti digitali, fonti storiche digitali.

The text discusses the rise of digital initiatives in historical research, particularly emphasizing the increase in the number of large repositories and archival collections. This proliferation is seen as part of the "big data" trend that needs to be debated in the discipline. Some examples of projects with a strong empirical approach are presented (such as *Transkribus* and *Time Machine*), and how these initiatives are associated with discussions about the supposed irrelevance of theory. The article proposes ideas for debate, indicating digital training as a necessity in preparing historians in the face of a sea of uncertainties and algorithms.

Keywords: big data, Neopositivism, digital history, digital tools, digital historical sources.

Negli ultimi anni abbiamo assistito a un continuo aumento delle iniziative tese a "informatizzare" la ricerca storica, utilizzando gli strumenti e risorse digitali per i temi e gli approcci più diversificati. L'espansione del Web ha generato e genera ogni giorno un numero incommensurabile di documenti digitali, nativi digitali o digitalizzati. La valanga di dati è tale che sono ovviamente sorti diversi progetti per tentare di gestire questa immensità.[2]

Queste preoccupazioni si sono lentamente fatte strada nell'agenda della ricerca storica, sia nel trattamento dei dati *online* sia nell'uso delle fonti digitalizzate, che stanno diventando sempre più

[1] Articolo già uscito in portoghese, *Sobre big data e neo-positivismo digital na pesquisa em história,* sulla rivista e-A *Almanack (36, 2024)* https://doi.org/10.1590/2236-463336ep00124, tradotto da Tiago Luís Gil con la collaborazione di Enrica Salvatori.

[2] [18]; [10].

comuni in tutto il mondo.[3] L'agenda delle *big tech* è ormai una realtà anche per gli storici e il richiamo ai cosiddetti *big data* è costante, soprattutto in Europa e negli Stati Uniti. Dobbiamo abbracciare i *big data* come sostengono gli storici Guldi e Armitage?[4] Oppure dobbiamo essere cauti e valutare criticamente questi strumenti, come auspica la matematica e attivista Cathy O'Neil?[5] La nostra formazione di storici ci fornisce gli strumenti per farlo? Questo contributo porta alcuni elementi a un dibattito che sta prendendo piede in molti paesi.

C'era una volta... il futuro

Era il 2012 quando Frédéric Kaplan avviò la *Venice Time Machine*. Questo progetto internazionale da milione di dollari prevedeva la digitalizzazione di molti chilometri lineari di documenti dell'Archivio di Stato di Venezia (partner del progetto) per creare un «*multidimensional model of Venice and its evolution covering a period of more than 1000 years*».[6] Il progetto prevedeva l'installazione di grandi *scanner* nell'Archivio e il loro utilizzo da parte di ingegneri (come lo stesso Kaplan) per il successivo riconoscimento digitale automatico dei manoscritti. Attraverso elaborate risorse di programmazione, sarebbe stato possibile identificare i nomi di persone e luoghi, il che avrebbe portato, quasi come conseguenza naturale, alla scoperta di reti di relazioni e alla creazione (altrettanto "naturale") di ampi grafici di reti sociali, che mostrassero le connessioni (certamente semplici) tra le persone e i loro spazi di attività:

> *By combining this mass of information, it is possible to reconstruct large segments of the city's past: complete biographies, political dynamics, or even the appearance of buildings and entire neighborhoods. The information extracted from the primary and secondary sources are organized in a semantic graph of linked data and unfolded in space and time in an historical geographical information system*.[7]

Era una favola digitale. L'ingegneria informatica moderna stava cercando di avere successo dove innumerevoli storici avevano finora fallito: scansionare in modo esaustivo una lunghissima serie di documenti e leggerli. Non si trattava solo di digitalizzare, infatti, ma di riconoscere anche le varie grafie prodotte in varie lingue (ricordiamo che Venezia era una potenza commerciale che manteneva ambasciatori in terre lontane), tra cui il latino e il veneziano.

Non si trattava quindi solo di leggere il contenuto, ma di ottenere informazioni che ci permettessero di riconoscere i personaggi e le loro vite (ossia di estrarre dati significativi), che sarebbero state poi assemblate e raccontate (anche al grande pubblico) utilizzando moderni dispositivi informatici. La favola era ancora più fantastica perché i siti web del progetto utilizzavano con cura risorse grafiche come video, mappe a colori e grafici animati, e un impressionante (e impressionistico) TED Talk. Il futuro aveva finalmente raggiunto i territori del passato e gli storici stavano per diventare spettatori privilegiati - al pari degli altri utenti - della potenza di calcolo che gli ingegneri potevano mettere in campo.

[3] [10].

[4] [11].

[5] [28].

[6] [16].

[7] [16].

Non c'è stato alcun happy end.. Era il settembre 2019 e l'Archivio di Stato di Venezia emise un comunicato stampa per informare della sospensione degli accordi di collaborazione allora in vigore con l'*École Polytechnique Fédérale* (EPFL) di Losanna, dove Kaplan lavorava, e con l'Università Ca' Foscari di Venezia, che era anche partner. Il motivo era la mancanza di trasparenza sulle attività svolte dal *team* di Losanna, sulle procedure adottate e sui risultati ottenuti. L'Archivio di Venezia lamentava in particolare la gerarchia dei ruoli tra le istituzioni, l'esclusione dei propri tecnici dal lavoro, l'assenza di discussione sulle procedure di digitalizzazione e catalogazione e, infine, la mancanza di trasparenza sulle decisioni prese e sull'analisi dei primi risultati. Secondo l'allora direttore dell'Archivio, Gianni Doria, la decisione di terminare il lavoro fu reciproca dopo diversi tentativi di raggiungere accordi che specificassero una nuova politica di interazione. L'EPFL, tuttavia, sostenne che si trattava di una decisione unilaterale dell'Archivio e informò che avrebbe cercato nuovi dialoghi, che poi non si concretizzarono.[8]

La decisione motivata dell'Archivio di Venezia ha evidenziato qualcosa di più di una semplice *impasse* tecnica e politica sulla collaborazione reciproca. Alla fine del documento, infatti, viene fatta un'osservazione importante, che va a toccare al cuore d i termini della partnership, ribadendo la « *convinzione che non sia sufficiente digitalizzare i documenti, anche attraverso strumenti e algoritmi complessi, per comprendere la storia di Venezia*».[9] Si può leggere la chiosa come una nota di profumo e brio locale, ma anche come una palese resistenza a un progetto che tentava di colonizzare i documenti dell'archivio con armi sconosciute e una potenza di fuoco non del tutto spiegata.

Non si trattava solo di una questione di contratti o di mancanza di comunicazione. La posta in gioco era la nozione stessa di Storia nelle menti dei membri dei *team*. Il concetto di ricerca storica dell'Archivio era tradizionale, legata alla lettura lenta ed erudita delle fonti. La nozione di Kaplan, apparentemente più moderna, eliminava lo studioso e lo sostituiva con un algoritmo. Entrambe erano fondamentalmente empiriste, ma quella dell'EPFL non era aperta alla critica (delle fonti) e partiva dalla premessa che i dati si sarebbero organizzati da soli, indipendentemente dalla teoria, o meglio, i dati sarebbero stati visualizzati in un modo accettabile per tutte le possibili letture del mondo, sia per gli esperti che per il grande pubblico.[10]

Il progetto delle *Macchine del Tempo* ha preso negli anni successivi una strada diversa, che meriterebbe un'analisi dettagliata da svolgere in un'altra occasione. In ogni caso, vale la pena di sottolineare che le sue implicazioni politiche sono molto complesse, ben lontane dall'offrire una soluzione meramente tecnica ai problemi degli storici. Il progetto ha ora rivolto la sua attenzione alla creazione di hub della *Time Machine* in varie città europee, senza mai perdere di vista l'idea che gli algoritmi possano "amplificare" le fonti disponibili.[11]

[8] [3], p. 607.

[9] Archivio di Stato di Venezia, 2019. "*Con la convinzione che non basta digitalizzare i documenti, anche con l'uso di strumenti e algoritmi complessi, per capire la storia di Venezia*".

[10] La teoria, qui, non appare come una teoria specifica, come un particolare modello esplicativo, né come un campo disciplinare, ma come una componente fondamentale del processo di produzione della conoscenza.

[11] [31].

Il canto della sirena

Un noto passo dell'*Odissea* racconta di quando Ulisse passò con la sua nave davanti a un'isola intorno alla quale abbondavano le sirene, capaci di attirare le navi vicino agli scogli con il loro bel canto e di farle naufragare.[12] Per godersi il canto senza rischiare nulla, Ulisse coprì le orecchie dei suoi marinai e si legò all'albero della nave.

Transkribus, uno strumento di riconoscimento automatico dei manoscritti, senza dubbio appare agli studiosi delle fonti storiche è come una melodia irresistibile.. Nato da un progetto dell'Università di Innsbruck, è in grado di identificare tramite HTR (Handwriting Text Recognition) il testo in un documento scritto a mano, riconoscere le linee e infine trascriverlo identificando la "zona" dell'immagine (tramite le coordinate cartesiane dei pixel dell'immagine) che lo contiene.

Il progetto è partito dalla precedente esperienza di Günter Mühlberger nella digitalizzazione e nel riconoscimento dei caratteri con la tecnologia OCR (*Optical Character Recognition*) nei giornali tedeschi degli anni Novanta.[13] Da quel momento in poi, il team intorno a Mühlberger è passato a nuove sfide e all'inizio degli anni 2000 stava già lavorando a progetti di riconoscimento di manoscritti, anche se con vari insuccessi.[14]

È essenziale spiegare la differenza tra le due tecnologie. Mentre i sistemi OCR si basano sulla somiglianza delle lettere in base a una standardizzazione rigorosa, come quella della carta stampata, l'HTR non può contare sulla standardizzazione dei caratteri a causa della molteplicità delle forme di scrittura umana, anche quando si tratta della stessa "mano". Le soluzioni di riconoscimento della scrittura manuale devono essere molto più elaborate, il che implica la creazione di modelli di scrittura basati sull'addestramento dell'intelligenza artificiale.[15]

Nel 2013, le iniziative di Mühlberger hanno preso la forma di un progetto che mirava effettivamente a riconoscere i testi scritti a mano. Questo progetto *tranScriptorium* è stato la base per *Transkribus*, sviluppato successivamente in collaborazione con l'Università di Valencia. Lo strumento è stato mantenuto all'interno del progetto originale fino al 2016, quando è stata creata la cooperativa READ. Da allora, la cooperativa è responsabile della manutenzione e dello sviluppo di nuove funzionalità.

Transkribus è senza dubbio un buon strumento, suggestivo come il canto delle sirene. Il problema è sempre il rischio di naufragio. Transkribus non è uno strumento che di per sé alimenta l'ossessione empirista, ma certamente vi contribuisce. La tecnologia HTR, infatti, non può essere intesa come uno strumento neutro e ancor meno decontestualizzata da un mondo in cui la mole di dati è diventata effettivamente seducente.

Come per la *Venice Time Machine*, la maggior parte dei testi che presentano il progetto e i suoi *outputs* sono fortemente tecnici.[16] Un'altra parte significativa è presa dalla descrizione dei vantaggi e dall'elenco dei potenziali consumatori, in cui storici e pubblico in generale sono affiancati come se la lettura delle fonti antiche fosse la stessa o comunque molto vicina. Non si discute la visione

[12] Odissea XII.

[13] [30].

[14] [30].

[15] [15], pagg. 19-24.

[16] [30]; [15].

del mondo che guida (deve guidare) la lettura o la natura della domanda, o le motivazioni dietro la volontà di avere tanto materiale da analizzare. Avere la possibilità di cercare dati in milioni di documenti antichi sembra una necessità indiscutibile, di per sé evidente.

La voracità nel fagocitare e digerire documenti non è sempre sinonimo di empirismo esagerato. Per raccontare la storia di persone semplici su cui sono stati prodotti pochi documenti, dobbiamo leggere e scartare migliaia di potenziali fonti che potrebbero - forse - parlare di quella persona e del contesto in cui ha vissuto. Ne consultiamo molte ma ne usiamo, spesso, poche. Questo è dovuto al fatto che le fonti sulle persone del passato non sono state prodotte né conservate in modo omogeneo. Fonti primarie e secondarie differiscono per numero, quantità e qualità a seconda dell'oggetto che si vuole narrare. Sono del tutto diseguali e riflettono, nel complesso, ciò che è stato prodotto e ciò che è stato preservato in relazione a precise scelte e al caso. Questo implica che le persone comuni siano scarsamente descritte nelle fonti: se da un lato è più che ragionevole raccontare le loro storie, questo non si ottiene semplicemente aumentando il numero dei documenti consultabili. Tuttavia, la preoccupazione di includere persone meno famose nelle fonti storiche non è certo al centro di *Transkribus*. Non si discute sulla selettività della memoria e non ci si pone il problema se, l'aumento incontrollato non finirà per aumentare le disparità che già esistono nelle narrazioni storiche, dando ancora più importanza a coloro che sono già ben noti in quanto "menzionati più di frequente". Il motto del progetto, *unlock the past*, presentato sulla homepage e nella pubblicità del servizio è una metafora significativa: suggerisce che il passato è chiuso a chiave. Trattandosi di uno strumento di riconoscimento dei caratteri, il messaggio implicito è che il ricercatore debba solo leggere tutti i documenti per sbloccarlo. Il passato è intrappolato nella paleografia.

La nozione di *unlock* compare in altri contesti nella pagina del progetto. Quando si presenta l'insieme degli strumenti associati a *Transkribus*, la metafora riappare, con una leggera variazione: *unlock history*. Queste sono le *Features to unlock history*, descritte nella sequenza *AI Text Recognition; Custom AI Training; Field & Table Recognition; Powerful Text Editor; Publishing & Search Tool*. L'osservazione storica comporta l'apertura di documenti e la ricerca di parole chiave in fonti con testo già riconosciuto. L'idea del metodo storico, ossia dell'approccio critico, come strumento essenziale di decifrazione delle fonti (esegesi) non compare. Sarà la marea dei dati a guidarci.

Il concetto di *sbloccare la storia* è rafforzato dalle parole del creatore di *Transkribus,* Günter Mühlberger, in una breve intervista al suo team nel 2023. Secondo lui, «*there are still so many interesting documents out there waiting to be discovered: Exploring them with HTR will be a big boost to historical research*».[17] I documenti aspettano solo di essere scoperti, e incontrarne uno darà un notevole impulso alla ricerca storica. Questa concezione della storia affonda le sue radici nell'empirismo esplicito del XIX secolo. È l'uso dell'intelligenza artificiale con un'immaginazione romantica.

Il problema non è, ovviamente, la metafora. Ginzburg e Prosperi hanno usato la stessa metafora dello "unlocking" (*forzare*) in un libro pubblicato nel 1975, quando hanno cercato di decifrare le condizioni di produzione e le prospettive teologiche di un'opera del XVI secolo. Tuttavia, a quel tempo, la "chiave principale" per svelare il segreto era un concetto teorico, non un mucchio di manoscritti.[18]

[17] [30].

[18] [7].

Sul neopositivismo digitale

Le storie della *Venice Time Machine* e del *Transkribus* non sono casi isolati. Sono progetti vicini al campo degli storici e forse per questo ci sembrano più attraenti. La nostra vita digitale quotidiana è segnata da "soluzioni" digitali che organizzano milioni di dati e ci offrono risposte considerate "neutre" e accettabili senza alcun impegno critico e metodologicamente valido. I motori di ricerca ne sono un esempio perfetto. Ogni giorno li usiamo per trovare informazioni senza avere la minima idea delle decisioni che questi strumenti prendono (automaticamente, ma creati da un cervello umano) per dare priorità alle risposte, escludere o includere elementi e molte altre variabili.[19] De Certeau ci ha ricordato nel 1972 che la selezione dei dati è il primo momento in cui opera l'ipotesi scientifica.[20] Cosa facciamo ora quando un algoritmo seleziona i materiali con cui abbiamo a che fare nella ricerca e nella nostra vita?

> *Par sa puissance et son efficacité, par son ambiguïté, Google neutralise le sens critique qui nous permettrait de garder à l'esprit que lorsque nous y cherchons une information, le moteur de recherche mondial avance une représentation particulière de la réalité pour nous répondre, et non la réalité elle-même*.[21]

Questo appello alla tecnologia come risposta primaria ai problemi degli storici non è un una fissazione degli ingegneri che creano motori di ricerca, trascrittori automatici o "modelli multidimensionali" dell'evoluzione di un determinato luogo nel tempo. È un modo di pensare già penetrato profondamente nella nostra disciplina, non solo per la (sempre crescente) ampiezza degli strumenti digitali nella nostra vita quotidiana, ma anche per la vertiginosa osservazione delle collezioni *online* sempre più sovrabbondanti, siano esse nate digitali o ora digitalizzate. Molti storici sono euforici riguardo ai *big data* e alle soluzioni tecniche per analizzarli.[22] Alcuni sostengono che l'uso di grandi volumi di dati, combinato con nuove forme di visualizzazione, può essere dirompente e dare origine a nuove epistemologie.[23]

La programmazione è diventata uno strumento di ricerca ampiamente disponibile e stiamo assistendo alla nascita di diversi corsi, non solo nelle *Digital Humanities* ma anche nella stessa *Digital History*. La critica a questa impostazione, però, non può limitarsi al fatto che le cosiddette "scienze umane" sono difficili da classificare e che condividono epistemologie e pratiche solo parzialmente. Dobbiamo discutere i limiti e le opportunità della tecnica. [24]

Un esempio di questo approccio è stato un lavoro del 2014, *The History Manifesto,* che ha avuto ripercussioni significative e ha generato un ampio dibattito. La maggior parte del dibattito si è limitata a discutere la *longue durée* presentata dagli autori, Guldi e Armitage. Secondo loro, una nuova ricerca storica con una maggiore estensione cronologica avrebbe avuto un effetto benefico dopo anni di ricerche brevi, responsabili, tra le altre cose anche dell'allontanamento della storia da un pubblico potenzialmente significativo e della perdita di rilevanza degli storici come figure essenziali nella formazione dell'opinione pubblica. Tuttavia l'uso dei *big data,* che gli

[19] [27].

[20] [5].

[21] [27].

[22] [19]; [6]; [13].

[23] [26]; [19], p. 20-29.

[24] [27].

autori hanno sostenuto utilizzando strumenti digitali, non è stato sufficientemente discusso. Così oggi l'emergere dei big *data* appare quasi come la soluzione a vecchi problemi.[25]

Guldi e Armitage non sono soli. Ci sono una profusione di riviste specializzate che stanno nascendo sul tema delle *Digital Humanities* e della storia digitale. Non si tratta di pubblicazioni acritiche; infatti, lo spazio riservato alla discussione teorica esiste, ma è molto ridotto rispetto alle descrizioni e di ricette per guarnire l'appetitosa torta digitale. Nel 2007 è *apparso* il *Digital Humanities Quarterly*. Nel 2011 è apparso il *Journal of Digital Humanities*. Nel 2015 è stata la volta del *Journal of Open Humanities Data*. Nel 2017 è stata creata *Umanistica Digitale*, dall'altrettanto recente *Associazione per l'Informatica Umanistica e la Cultura Digitale* (2011). Nel 2019, l'*International Journal of Digital Humanities*. Nel 2020 sono apparse due nuove pubblicazioni: la francese *Humanités numériques* e l'italiana *Magazén: international journal for digital and public humanities*. Questo considerando solo le pubblicazioni con un focus integrale sul tema, senza considerare i vari *dossier* lanciati.

Dal 2010 si è assistito a un'impennata di pubblicazioni tecniche destinate alla ricerca umanistica. Uno dei primi esempi è *Macroanalysis: digital methods and literary history* (2011) *di Matthew Jockers*.[26] Nel 2014, Folgert Karsdorp ha presentato un'edizione virtuale (in un "quaderno virtuale") di un corso sul linguaggio di programmazione Python, con *Python Programming for the Humanities,* che ha portato alla successiva edizione di *Humanities data analysis: case studies with Python* (2021), in collaborazione con Mike Kestemont e Allen Riddell.[27] Nel 2018, Brian Kokensparger ha lanciato *Guide to Programming for the Digital Humanities*, anch'esso incentrato sul linguaggio Python.[28] Nel 2020, Jemielniak ha lanciato *Thick big data: doing digital social sciences*, dimostrando che la moda delle digital humanities si stava spostando verso i cosiddetti *big data*.[29] Esiste anche un'ampia gamma di corsi estivi, master e dottorati costruiti intorno a questo tema nuovo e caldo.

L'enfasi di tutti questi corsi e pubblicazioni è fondamentalmente incentrata sulla tecnica come strumento per generare grandi volumi di dati. Oltre ad avere uno scarso *appeal*, la discussione metodologica è generalmente trattata come un problema individuale, relativo alla singola analisi, che non sarebbe un male se il dibattito generale fosse vivace. Il problema, come già accennato, è che tutto questo movimento si svolge parallelamente a un altro grave processo: la negazione della teoria. Come già sottolineato da Pierre Mounier nel libro *The End of Theory* di Chris Anderson*,* l'autore afferma esplicitamente che la grande quantità di dati sempre più accessibili renderà presto obsoleto il metodo scientifico.[30] Non si tratta di un'affermazione scollegata dalla realtà e dall'ambiente in cui sono nate tutte le opere a cui abbiamo fatto riferimento, e nemmeno i corsi tecnici. Il progetto *Time Machine* non aveva una proposta strutturalmente distante da questa prospettiva.

La proposta di Anderson non è una posizione isolata. Molti ingegneri, matematici e statistici scommettono sulla capacità degli strumenti statistici di operare attraverso le correlazioni e generare nuova conoscenza. Sulla base di schemi specifici, sarebbe possibile - a loro avviso -

[25] [11].

[26] [14].

[27] [17].

[28] [20].

[29] [13].

[30] [1].

ottenere intuizioni «*born from the data*»[31] in una nuova era ove «*the volume of data, accompanied by techniques that can reveal their inherent truth, enables data to speak for themselves free of theory*».[32]

Mounier si concentra su un altro caso, quello della *culturomics*, come Google Ngram Viewer[33] e il recentissimo *Gallicagram*[34]. In questi programmi, milioni di libri sono organizzati e resi ricercabili a partire da unità di testo di base, gli *ngram*, che permettono di identificare la frequenza d'uso di termini specifici in determinati periodi in *corpora* testuali massicci. Mounier discute proprio l'esempio fornito da Michel *et al.*[35] del termine "Chagall" in milioni di libri in lingua tedesca, ove il declino di questo termine durante il periodo del nazismo sarebbe stato causato dalle limitazioni e dalle condanne agli artisti ebrei. Mounier sottolinea che questa associazione è stata possibile non grazie al grande volume di dati di Google Ngram Viewer, ma a una conoscenza pregressa, basata su opere non necessariamente digitali o quantitative. Ma come interpretiamo i risultati statistici di processi se non abbiamo contezza del contesto? ome interpreteremmo, ad esempio, un eventuale calo delle registrazioni di vendita negli uffici del registro di Venezia (per tornare al caso precedente)? Un calo delle imprese o delle registrazioni commerciali, giusto per dare due semplici possibilità? I caso Chagall conferma un assunto pregresso, non lo svela.

In fin dei conti, l'analisi dei big data non consente alcuna critica documentale perché non abbiamo idea di come sia stato effettivamente prodotto il materiale che analizziamo, dato che l'analisi si basa sulla relazione di dati grezzi . Comprendere l'esistenza di un documento storico, la sua creazione e la sua conservazione nel tempo, implica la delucidazione delle varie selezioni effettuate socialmente nel corso del tempo, che hanno fatto sì che alcune cose venissero ricordate e altre dimenticate. Gli archivi, come i libri nelle biblioteche (per tornare ai casi di *Google Ngram Viewer* e della *Venice Time Machine*), non sono campioni rappresentativi del mondo o del passato, ma costruzioni sociali generate da una società in evoluzione. Come ha detto Mounier «*Ce n'est pas l'absence d'exhaustivité des données qui compte, c'est le fait que leur constitution est le résultat d'une intention humaine*».[36]

La minaccia dei *big data* è ben visibile se la consideriamo, come abbiamo fatto qui, come una minaccia al metodo scientifico e alla definizione di teoria. Alla base di tutto questo c'è tuttavia un concetto purtroppo diffuso tra gli storici: la separazione tra teoria e tecnica. Questa nozione era alla base dell'esperimento della *Venice Time Machine*, ma, paradossalmente, è molto più diffusa e accettata negli studiosi tradizionali. È presente in tutti i manuali di programmazione per storici e nei corsi che abbiamo citato. È largamente accettata perché molti storici affidano acriticamente a strumenti digitali le loro attività di raccolta di documenti e, più recentemente, anche di elaborazione dei dati. L'adozione di un *software* che svolge parte del lavoro sembra a loro del tutto in linea con la ricerca come è sempre stata fatta. Ma i problemi sorgono eccome quando non sappiamo esattamente cosa fa il software.

[31] [18].

[32] [18].

[33] [23].

[34] [2]; [4].

[35] [23].

[36] [27].

Avvicinarsi alla conclusione (con un gioco)

Una delle tecniche più utilizzate nelle *Digital Humanities*, che appare con forza in diversi libri di storia digitale, è chiamata *Topic Modeling* (o Topic Extraction). Questa tecnica promette di identificare le parole chiave di un testo dal testo stesso, come se cercasse di indovinare l'argomento dello scritto senza bisogno che un essere umano lo legga. Si tratta di una premessa interessante, che potrebbe aiutare i ricercatori a scegliere le loro letture in base a un precedente filtraggio per parole chiave. L'idea di utilizzare le parole chiave è vecchia, ma è sempre dipesa da un'azione umana precedente, da parte dell'autore o dei professionisti della biblioteca. La promessa ora è quella di automatizzare questo compito, umanamente impossibile nell'attuale contesto di produzione frenetica e di pubblicazione accelerata.[37]

Il lettore che ci ha seguito fin qui è ora invitato a fare un gioco. Si immagini delle parole chiave per il testo appena letto. L'articolo pubblicato ha le parole chiave assegnate dall'autore, ma il lettore è invitato a valutarle, poiché questa decisione è sempre molto soggettiva. Fatte queste premesse, vediamo come due algoritmi hanno risolto lo stesso problema. Il primo è quello disponibile sul *sito web* "nocodefunctions.com"[38] , creato dal professore di economia della *Emlyon Business School* Clément Levallois, che dispone di alcuni strumenti digitali considerati eccellenti da alcuni studi.[39] Il risultato ha avuto diversi livelli di risposta, ma i primi due, i più affidabili, indicavano le seguenti espressioni:

> argomento 0: idea, idea più piccola, Venice time machine, analisi, google, documentario
> argomento 1: progetto, documenti, strumento, transkribus, fonti

Possiamo immaginare che il lettore sia frustrato: non comprende il motivo alla base di una tale selezione, anche se una descrizione del funzionamento dell'algoritmo si trova direttamente sul *sito web* di Levallois.

Prima di condannare (o dare ragione) ai risultati dell'algoritmo di Levallois, vediamo però un altro risultato prodotto da un algoritmo creato da chi scrive utilizzando il linguaggio di programmazione Python. Il *code* non tiene conto di tutto ciò che non è un sostantivo e cerca le parole più strettamente correlate tra loro - quelle che compaiono più spesso nella stessa frase - il che non significa che siano le più frequenti. Il risultato è stato un insieme di parole composto da "progetto", "strumento", "storico", "dati", "documento" e "storia".

Secondo il lettore quali dei due risultati sono migliori? Quali sono erronei? Quali riassumono meglio il testo? La difficoltà non sta nel giudicare l'efficienza dello strumento. Il punto è che il risultato avrà sempre da un lato un certo grado di verosimiglianza e dall'altro una buona dose di assurdità, e la sua classificazione come "buono" o "cattivo" è una decisione che spetta alla soggettività di ognuno. Per chi ha appena letto il testo, il giudizio è relativamente facile. E per coloro che non l'hanno letto? È per questi ultimi che è stata costruita l'analisi dei *big data* e questa tendenza non sembra rallentare. Siamo preparati a questa storia disintermediata? Siamo in grado di criticare strumenti sempre più diffusi? Stiamo discutendo di tutto questo nelle università e nei centri di ricerca? O l'analfabetismo digitale attualmente così diffuso tra gli umanisti resterà la migliore via d'uscita per fingere che il problema non sussista? ?

[37] [10].

[38] [21].

[39] [29]; [22].

Riferimenti

[1] Anderson, Chris. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, [*s. l.*]. Disponível em: https://www.wired.com/2008/06/pb-theory/. Acesso em: 14 mar. 2024.

[2] Azoulay, Benjamin; Courson, Benoît De. 2021. Gallicagram: un outil de lexicométrie pour la recherche. *SocArXiv*, [*s. l.*], 8 dez. 2021. doi: https://doi.org/10.31235/osf.io/84bf3.

[3] Castelvecchi, Davide. 2019. Venice 'Time Machine' Project Suspended amid Data Row. *Nature*, [*s. l.*], v. 574, n. 7780, 31 out. 2019, p. 607. doi: https://doi.org/10.1038/d41586-019-03240-w.

[4] Courson, Benoît De *et al.* 2023. Gallicagram : les archives de presse sous les rotatives de la statistique textuelle. *Corpus*, [*s. l.*], n. 24, 15 jan. 2023. doi: https://doi.org/10.4000/corpus.7944.

[5] De Certeau, Michel. 1978. A operação histórica. *In*: Le Goff, Jacques; Nora, Pierre (org.). *História:* novos problemas. São Paulo: Livraria Francisco Alves Editora.

[6] Ehrmann, Maud *et al.* 2021. Named Entity Recognition and Classification on Historical Documents: A Survey . *arXiv:2109.11406 [cs]*, 23 set. 2021. Disponível em: http://arxiv.org/abs/2109.11406. Acesso em: 14 mar. 2024.

[7] Ginzburg, Carlo; Prosperi, Adriano. 1975. *Giochi di pazienza*: un seminario sul Beneficio di Cristo. Vol. 258. Torino: Einaudi.

[8] "Google Books Ngram Viewer", 2024. https://books.google.com/ngrams/.

[9] Graham, Shawn; Milligan, Ian; Weingart, Scott. 2016. *Exploring Big Historical Data*: The Historian's Macroscope. [*S. l.*]: World Scientific Publishing Company.

[10] Graham, Shawn; Milligan, Ian; Weingart, Scott. 2016. *Exploring Big Historical Data*: The Historian's Macroscope. [*S. l.*]: World Scientific Publishing Company.

[11] Guldi, Jo; Armitage, David. 2014. *The history manifesto*. Cambridge, United Kingdom: Cambridge University Press.

[12] Homero. 1997. *Odisséia Em Versos*. São Paulo: Ediouro.

[13] Jemielniak, Dariusz. 2020. *Thick big data*: doing digital social sciences. New product. New York: Oxford University Press.

[14] Jockers, Matthew Lee. 2013. *Macroanalysis: digital methods and literary history*. Topics in the digital humanities. Urbana: University of Illinois Press.

[15] Kahle, Philip *et al.* 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *In*: Iapr International Conference on Document Analysis and Recognition (ICDAR), 14., 2017, Kyoto. *Anais* […].Kyoto: IEEE, 2017. p. 19-24. https://doi.org/10.1109/ICDAR.2017.307.

[16] Kaplan, Frédéric. 2015. The Venice Time Machine. *In*: Acm Symposium on Document Engineering, 2015, Lausanne Switzerland. *Proceedings* […]. Lausanne Switzerland: ACM, 2015. p. 73. doi: https://doi.org/10.1145/2682571.2797071.

[17] Karsdorp, Folgert; Kestemont, Mike; Riddell, Allen. 2021 *Humanities data analysis*: case studies with Python. Princeton: Princeton University Press, 2021.

[18] Kitchin, Rob. 2014. Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, [*s. l.*], v. 1, n· 1, 1 abr. 2014. doi: https://doi.org/10.1177/2053951714528481.

[19] Kitchin, Rob. 2006. Positivistic geography and spatial science". *In*: Kitchin, Rob. *Approaches to human geography*. London: Sage. p. 20-29. Disponível em: https://scholar.google.com/citations?view_op=view_citation&hl=pt-PT&user=Y_3-GBQAAAAJ&cstart=200&pagesize=100&sortby=pubdate&citation_for_view=Y_3-GBQAAAAJ:hEXC_dOfxuUC. Acesso em: 14 mar. 2024.

[20] Kokensparger, Brian. 2018. *Guide to Programming for the Digital Humanities*: Lessons for Introductory Python. New York: Springer.

[21] Levallois, Clément. 2018. *Nocode Functions*, 2024. Disponível em: https://nocodefunctions.com/. Acesso em: 14 mar. 2024.

[22] Levallois, Clement. 2013. Umigon: sentiment analysis for tweets based on lexicons and heuristics. *In*: International Workshop on Semantic Evaluation, Semeval, 2013, Atlanta. *Proceedings* […]. Atlanta: [*s. l.*], 2013. p. 414-417. Disponível em: https://scholar.google.com/scholar?cluster=9774134509927520 80&hl=en&oi=scholarr. Acesso em: 14 mar. 2024.

[23] Michel, Jean-Baptiste *et al*. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, [*s. l.*], v. 331, n. 6014, p. p. 176-182, 14 jan. 2011. doi: https://doi.org/10.1126/science.1199644.

[24] Ministerio Della Cultura. Archivio di Stato di Venezia. 2019. *Sospensione dei rapporti con EPFL su Time Machine*, 2019. Disponível em: https://www.archiviodistatovenezia.it/it/eventi/news/sospensione-dei-rapporti-con-epfl-su-time-machine.html. Acesso em: 14 mar. 2024.

[25] Moretti, Franco. 2024. *Atlas of the European Novel, 1800-1900*. Verso, 1999. Disponível em: https://books.google.com.br/books?id=ja2MUXS_YQUC. Acesso em: 14 mar. 2024.

[26] Moretti, Franco. 2005. *Graphs, Maps, Trees*: Abstract Models for a Literary History. Verso, 2005. Disponível em: https://books.google.com.br/books?id=YL2kvMIF8hEC. Acesso em: 14 mar. 2024.

[27] Mounier, Pierre. 2018. *Les humanités numériques*: Une histoire critique. Paris: Éditions de la Maison des sciences de l'homme, 2018. doi: https://doi.org/10.4000/books.editionsmsh.12006.

[28] O'neil, Cathy. 2016. *Weapons of math destruction*: how big data increases inequality and threatens democracy. 1st. ed. New York: Crown.

[29] Ribeiro, Filipe Nunes *et al.* 2024. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *arXiv*, [*s. l.*], 14 jul. 2016. Disponível em: http://arxiv.org/abs/1512.01818. Acesso em: 14 mar. 2024.

[30] Stauder, Florian. 2023. A Short History of Transkribus with Günter Mühlberger . *READ-COOP*, 22 fev. 2023. Disponível em: https://readcoop.eu/a-short-history-of-transkribus-with-gunter-muhlberger/. Acesso em: 14 mar. 2024.

[31] Time Machine Europe. 2024. *About Us*. Disponível em: https://www.timemachine.eu/about-us/. Acesso em: 14 mar. 2024.