StatusQuaestionis
Rivista di studi letterari, linguistici e interdisciplinari

Federica Perazzini
"Sapienza" Università di Roma

Words, Bytes and Numbers: le Digital Humanities "viste da vicino"

Abstract

This article attempts an updated, exhaustive definition of the now fashionable, as well as deceptive, idea of the Digital Humanities by outlining a short history of the discipline and providing examples of its methods and results. Far from being an instrument subservient to traditional humanist knowledge, the Digital Humanities are now a fully-fledged, independent science that is strongly rooted in empirical data. This article highlights the application, and the potential, of the Digital Humanities in relation to the analysis of single texts as well as of broad textual corpora, and highlights how the Digital Humanities could radically transform the field, and the practices, of literary studies.

> Dopo aver contato gli aggettivi e soppesato le righe e misurato le rime,
> il formalismo o si arresta in silenzio con l'aria di chi non sa più cosa fare di sé
> o emette una generalizzazione a sorpresa che contiene il cinque per cento di formalismo e il novantacinque per cento della più acritica intuizione.
> Lev Trotskij, *Letteratura e Rivoluzione*

Quando un anno fa, appena discussa la tesi di dottorato, mi venne chiesto di scrivere un articolo sulle Digital Humanities, devo confessare che giudicai la faccenda con una certa leggerezza. In fin de conti avevo appena terminato uno studio di 320 pagine di esperimenti computazionali su un intero genere letterario, il romanzo gotico inglese, e un articolo sulla medesima disciplina, appunto le vecchie *humanities computing* ribattezzate poi *digital humanities*, non poteva cogliermi impreparata.

Eppure, ogni volta che mi accingevo a concludere, ma diciamo pure iniziare, un qualsiasi contributo di divulgazione scritta riguardo «this thing called Digital Humanities», come sarcasticamente rimarcò William Deresiewicz,[1] l'effettiva quanto imbarazzante difficoltà d'esecuzione pratica dell'impresa sembrava sabotare in partenza ogni tentativo. "Blocco dello scrittore"? Forse. Mera inettitudine dello stesso? Anche questo molto probabile. Volendo però uscire da una logica di vittimismo autoreferenziale, credo che una spiegazione plausibile alla suddetta difficoltà possa, in realtà, ricondursi a due ordini di ragioni; una intrinseca all'oggetto d'osservazione e l'altra intrinseca al soggetto osservatore.

In questo senso, riconoscere alle Digital Humanities uno statuto di disciplina ambigua, dalla natura eufemisticamente sfuggente e omnicomprensiva,[2] è assolutamente inevitabile, così come attribuire alle recenti tendenze iper-relativistiche della critica contemporanea la responsabilità della paralisi dello studioso, ormai impossibilitato – se non direttamente incapace – di fornire definizioni univoche ed esaustive dei fenomeni che di volta in volta si trova ad analizzare, è luogo teorico, se non acclarato, almeno tristemente consolidato.[3]

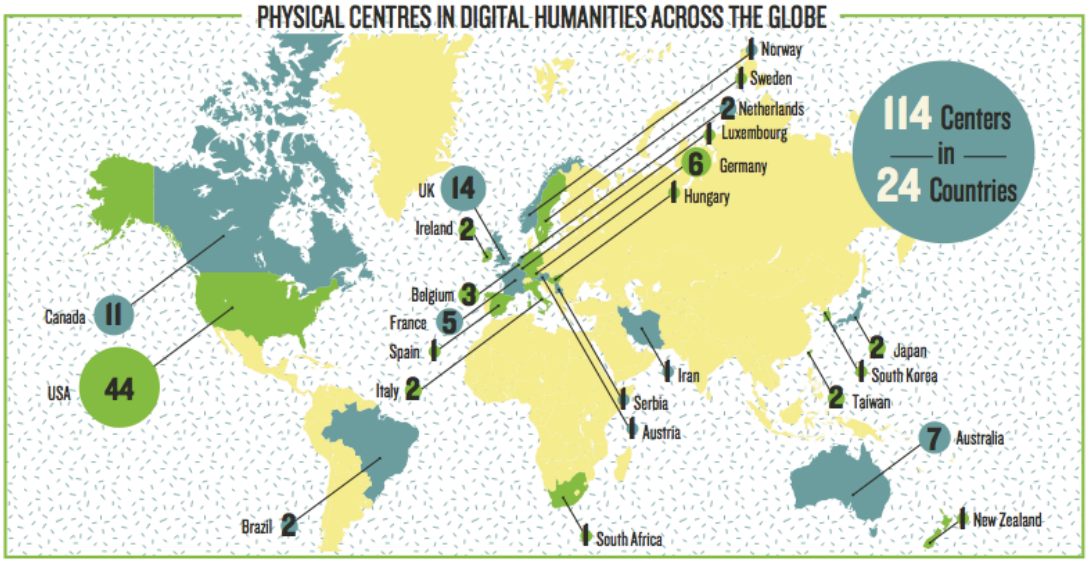
---

[1] Nel 2008, lamentando il tipo di tendenza nella pubblicazione di annunci di lavoro nella MLA job list, William Deresiewicz dichiarava: «There are postings here for positions in science fiction, in fantasy literature, in children's literature, even in something called 'digital humanities.'» http://digitalscholarship.wordpress.com/2008/10/18/digital-humanities-jobs/

[2] Segno distintivo dell'identità inclusiva delle Digital Humanities è infatti "l'ombrello" scelto dalla Alliance of Digital Humanities Organization (ADHO) come simbolo dell'associazione.

[3] Non più tardi di due mesi fa, nel Settembre del 2013, mi trovavo alla conferenza ACLAX (American Comparative Literature Association Examine) convocata per fare un bilancio degli ultimi dieci anni di lavori e collaborazioni e, contemporaneamente, per fare il punto sullo stato dell'arte della disciplina *Comparative Literature* nel mondo accademico. Ebbene, nonostante una ben collaudata organizzazione annuale di conferenze e una definizione del campo di studi assolutamente consolidata, alla ACLAX si discute ancora su cosa vada ad indicare l'etichetta *Comparative Literature*: una disci-

Come uscire dunque da questa empasse e proporre una visione quanto più fedele dello *status quaestionis* delle Digital Humanities? Probabilmente cominciando dal basso: dalle radici – tra l'altro tutte italiane – di una disciplina che pur avendo ormai conquistato il mondo, almeno in termini di diffusione geografica di centri di ricerca e dipartimenti, sembrerebbe non averlo mai del tutto convinto dei suoi "prodigi".



Mappa delle DH nel mondo a cura di Melissa Terras (UCL London, 2011)[4]

Dopo la ricostruzione diacronica di più di cinquant'anni di storia e iniziative, dalle radici passerò quindi all'esposizione delle componenti pratiche della ricerca nelle DH, concentrandomi sia sulla definizione

---

plina a tutti gli effetti? Un "method of enquiry"? Ora, capisco che l'atto di nominazione e definizione di un qualsiasi oggetto, soprattutto un oggetto di studio, inneschi anche un processo di cristallizzazione del tutto irreale rispetto alla natura e, se mi si consente, alla verità dell'oggetto stesso, ma resistere a priori all'atto di definizione per il relativistico timore di svilire o smarrire la complessità dell'oggetto temo sia altresì erroneo.

[4] Dal post di Melissa Terras, *Infographic, Quantifying the DH* pubblicato sul suo Blog: http://melissaterras.blogspot.it/2012/01/infographic-quanitifying-digital.html

dall'oggetto di studio, ossia il testo elettronico codificato, sia sulle possibili applicazioni di alcuni strumenti e metodi computazionali nell'ambito degli studi letterari, ossia gli esperimenti veri e propri. Inoltre, nel ripercorrere le fasi di trasformazione del testo da ente analogico a dato digitale, mi soffermerò sulle conseguenze epistemologiche che tale passaggio già e sempre comporta. Terminata l'esposizione di tali componenti e fasi, che definiremo come "pre-costruttive" della disciplina, effettuerò una panoramica sugli studi sperimentali più celebri e riusciti generati dall'intersezione tra metodi digitali e storia letteraria. Inevitabilmente, l'articolo si concluderà con una serie di osservazioni di natura, ancora una volta, epistemologica riguardo l'occasione da parte delle Digital Humanities di apportare della nuova conoscenza all'interno dell'odierno panorama di studi letterari, identificando nei concetti di "misurabilità" e "falsificabilità" delle unità d'analisi le condizioni di possibilità per una svolta paradigmatica di proporzioni epocali nella ricerca umanistica.

1. *Humanities Computing o Digital Humanities? Evoluzioni di un'intersezione*

Come condiviso dalla maggior parte degli studi in materia, data d'origine della tradizione dell'informatica umanistica è il 1949, anno in cui il progetto *Index Thomisticus* di Padre Busa vede la luce. L'idea dell'avanguardistico gesuita di Gallarate era appunto quella di produrre un indice di concordanze lemmatizzate di tutte le parole presenti nel corpus testuale di Tommaso d'Aquino e altre opere correlate.[5] Al di là degli

[5] Nelle parole di Susan Hockey «[Busa] wanted to produce a "lemmatized" concordance where words are listed under their dictionary headings, not under their simple forms. His team attempted to write some computer software to deal with this and, eventually, the lemmatization of all 11 million words was completed in a semiautomatic way with human beings dealing with word forms that the program could not handle. Busa set very high standards for his work. His volumes are elegantly typeset and

effettivi meriti teorici o pratici insiti nell'iniziativa del Busa, il punto è che essa fu la prima ad uscire dalla solitudine della pratica filologica conquistando il colosso informatico dell'epoca: l'IBM. Non solo, infatti, Thomas J. Watson, fondatore della suddetta società, si espose in prima persona per finanziare l'*Index* del Busa, progetto tra l'altro privo di qualsivoglia margine di lucro, ma a partire dagli anni sessanta il marchio IBM si fece garante e patrono del "dialogo" interdisciplinare tra *information sciences* e studi lingusitico-letterari, inaugurando un ciclo di conferenze internazionali incentrate sul tema del *Literary Data Processing* (la prima nel 1964). Fu proprio grazie a questa serie di incontri che tutte le diverse personalità avvicinatesi nel tempo a questo tipo di studi[6] ebbero modo di condividere le proprie idee e perplessità dando così vita a quella che di lì a poco sarebbe diventata la comunità scientifica delle *Humanities Computing.*

La nascita e l'adozione del termine *Humanitites Computing* quale etichetta identificativa di un campo di studi tanto variegato e a tratti ancora confuso necessita qui di una breve digressione. La maggior parte della ricerca prodotta nell'intervallo temporale compreso tra gli anni settanta e l'inizio dei novanta, infatti, rimase per lo più focalizzata sull'implemento

---

he would not compromise on any levels of scholarship in order to get the work done faster. He has continued to have a profound influence on humanities computing, with a vision and imagination that reach beyond the horizons of many of the current generation of practitioners who have been brought up with the Internet. A CD-ROM of the Aquinas material appeared in 1992 that incorporated some hyper-textual features ("cum hypertextibus") and was accompanied by a user guide in Latin, English, and Italian» in *A Companion to Digital Humanities*, Blackwell Publishing, Oxford, 2004.

[6] Roy Wisbey e la sua serie di concordanze per i testi alto germanici (successivamente pubblicato nel 1975 col titolo *Concordances to the Early Middle High German Biblical Epic*), Stephen Parrish sulla poesia di Yeats (*A Concordance to the Poems of W.B. Yeats*, Cornell University Press, 1963) o il reverendo Morton con i primi studi di *authorship attribution* per le lettere di San Paolo (A.Q. Morton, "Paul, the Man and the Myth, A Study in the Authorship of Greek Prose" in *Scottish Journal of Theology*, London, 1966, pp. 218-235).

delle metodologie e degli aspetti tecnico procedurali atti alla creazione di progetti in ambito archivistico conservativo (*data-oriented*),[7] ma anche allo sviluppo di *software* per l'analisi linguistica dei corpora di volta in volta generati (progetti *tool-oriented*). Ecco dunque che attraverso la scelta dal termine *computing,* a fianco delle intramontabili *Humanities,* i "padri fondatori" della disciplina vollero appunto sottolineare la dimensione prettamente procedurale, per non dire algoritmica, intrinseca alla nuova scienza della testualità digitale.

Inoltre, sull'onda del grande fermento ideologico e culturale che investì il mondo universitario e intellettuale negli stessi anni, l'esigenza di rafforzare i legami all'interno della nascente comunità scientifica delle *Humanities Computing* divenne sempre più pressante; una necessità a cui il mondo accademico anglosassone rispose con un serrato programma biennale d'incontri e conferenze[8] e un numero sempre crescente di *societies*, *alliances*, e *associations*, tra cui l'importantissima *Association for Literary and Linguistic Computing*, fondata al King's College nel 1973, e *l'Association for Computers and Humanities* del 1978.

Un ulteriore salto evoluzionistico all'interno della disciplina si ebbe poi a partire dalla seconda metà degli anni ottanta quando l'introduzione e la rapidissima diffusione del *personal computer* liberò di fatto gli studiosi dal vincolo della ricerca sui *mainframes* accademici inaugurando un'era di progettualità individuali *customized* e *customizable.* Questa radicale

---

[7] Tra le varie iniziative, d'incredibile impatto fu senz'altro la creazione *dell'Oxford Text Archive* (OTA) nel 1976, raccolta di testi e articoli della più svariata natura al servizio dell'accademia e della ricerca e, sempre nello stesso periodo e sempre ad Oxford, *dell'Oxford Concordance Program*, un software distribuito solo a partire dal 1982 che permetteva di generare liste, concordanze, co-occorrenze e indici da testi e corpora di qualsiasi lingua ed alfabeto.

[8] Tra queste ricordiamo l'incontro inaugurale a Cambridge dal titolo "Computers and the Humanities" (1970), seguito poi dalle conferenze di Edimburgo (1972), Cardiff (1974), Oxford (1976), Birmingham (1978) e di nuovo Cambridge (1980).

trasformazione dei tempi e delle modalità d'investigazione critica risultò in una moltiplicazione quasi esponenziale dei materiali e degli archivi elettronici disponibili; un fenomeno che in breve tempo impose all'attenzione della comunità delle *humanities computing* il problema di creare degli standard per la codifica dei documenti elettronici attraverso un linguaggio di *MarkUp* condiviso. Il *MarkUp* (in italiano "linguaggio di marcatura") può essere definito come un insieme di regole che descrivono i meccanismi di rappresentazione strutturale, semantica o di presentazione di un testo. Esistono due tipi di linguaggi di marcatura e, nella fattispecie, il *MarkUp* procedurale che indica le procedure di trattamento del testo aggiungendo le istruzioni che devono essere eseguite per visualizzare la porzione di testo referenziata, e il *MarkUp* descrittivo che invece lascia la scelta del tipo di rappresentazione da applicare al testo al software che di volta in volta lo riprodurrà (come XML, HTML, etc.). A questo ultimo tipo appartiene lo *Standard Generalized MarkUp Language* (SGML), primo schema generale per la definizione del linguaggio di codifica testuale apparso nel 1986, e il suo successivo perfezionamento rappresentato dalla *Text Encoding Initiative* (TEI).

Tornerò sulla TEI e la codifica nel paragrafo successivo. Per il momento, con un ultimo balzo temporale arriviamo alla seconda metà degli anni novanta e all'analisi dell'ultima rivoluzione che ha investito le *Humanities Computing* trasformandole nelle odierne Digital Humanities: l'avvento e la diffusione di massa di Internet e del World Wide Web.

In questa seconda denominazione della metodologia, infatti, la polisemia dell'aggettivo "digital", in sostituzione del più rigido "computing", va proprio a privilegiare l'essenza più generale e omnipervasiva delle cosiddette culture digitali indiscriminatamente disponibili in rete.

Riflesso immediato della diffusione e della domesticizzazione capillare del World Wide Web non è dunque solo lo smisurato "allargamento" della comunità scientifica delle *Humanities Computing*, ormai aperta anche a esponenti di ambiti non immediatamente riconducibili alla linguistica o alla letteratura, ma anche un miglioramento della possibilità di

comunicazione all'interno di essa. In quest'ottica, la maggiore possibilità di dialogo e condivisione, così come il conseguente incremento della quantità e diversificazione dei materiali immessi in rete, generarono una vera e propria emergenza organizzativa per quanto riguardava la categorizzazione di tale mole di dati. Grandi conquiste degli ultimi quindici anni di ricerca nelle Digital Humanities sono quindi gli archivi tematici, le biblioteche digitali, i *database* (*open-source* o a sottoscrizione) che sia a livello accademico che privato regolamentano il flusso continuo di materiali. Ad oggi, secondo uno studio *dell'Aspen Institute of Communication and Society*,[9] il risultato di questa monumentale opera di digitalizzazione ammonta ormai a circa 500 mila miliardi di megabytes. "The Age of Big Data" è infatti il termine utilizzato per identificare questa nostra prima decade del terzo millennio, un'epoca di informazioni e culture della più variegata natura che attendono solo di essere esplorate.

Come resistere a una sfida di tale entità gnoseologica? Come riuscire a non perdersi nei meandri di una disciplina ormai talmente versatile da investire e influenzare praticamente tutti i campi del sapere umanistico, dalla storia alle arti grafiche, dall'archeologia alla geografia politica, dalla letteratura al teatro? Consapevole dei pericoli che un'appropriata ricognizione critica potrebbe comportare ho altrettanto consapevolmente scelto di evitare l'ammaliante impresa di sondare l'oceano di alterità delle culture digitali concentrandomi piuttosto su quanto accade esclusivamente all'interno del mio campo di specializzazione: le scienze del testo.

Come preannunciato in apertura dell'articolo, nel prossimo paragrafo mi occuperò dunque di illustrare le fasi pre-costruttive della ricerca delle Digital Humanities applicate agli studi letterari analizzandone unità mi-

9 D. Bollier, *The Promise and Peril of Big Data*, The Aspen Institute Publication Office, Washington DC, 2001. Anche in http://citpsite.s3.amazonaws.com/events/bigdata/Aspen-Big_Data.pdf

nime d'analisi e procedure di preparazione dati per poi concludere con una ricognizione delle possibili forme d'investigazione.

2. *Dalle parole ai bytes: la rinascita del testo da analogico a digitale*

In informatica umanistica, perché questa è la più corretta traduzione italiana delle DH, l'insieme di operazioni atte alla trasformazione degli oggetti culturali da un formato analogico a uno digitale viene definito codifica: un atto che, prendendo in prestito le parole di Louis Althusser, realizza un importante slittamento paradigmatico dal mondo degli *oggetti reali* a quello degli *oggetti della conoscenza.* Se per il filosofo francese, infatti, la conoscenza di un oggetto reale è impossibile se non attraverso la creazione di un concetto dello stesso in termini di oggetto di conoscenza, un modello fluido e modificabile su cui operare, tale formulazione sembra rivelarsi assolutamente valida anche e soprattutto per gli oggetti culturali. Un libro, ad esempio, è un oggetto reale e il testo, sua controparte concettuale, costituisce l'unità minima d'analisi negli studi letterari. Ora, sia che esso si presenti nella consueta forma cartacea, sia che ci appaia nelle recenti versioni elettroniche per la lettura su schermo (doc, pdf, epub o mobi), il libro rimane, di fatto, un oggetto reale. Diviene classificabile quale oggetto di conoscenza solo quando, una volta digitalizzato, questo viene codificato attraverso l'apparato meta-linguistico descrittivo di tags e metadati del *MarkUp*. È dunque l'atto di codifica ciò che trasforma un testo elettronico di format generico (rtf, txt, html, ecc.) in un modello operativo a configurazione standard, ultimando così il passaggio epistemologico del documento da oggetto reale a oggetto della conoscenza.

Come accennato nel paragrafo precedente, infatti, la creazione di uno schema di codifica unificante con cui rappresentare la struttura semantica delle diverse tipologie di testualità digitale è da sempre scopo e compito della TEI (Text Encoding Initiative). La TEI è un consorzio di istituzioni internazionali di ambito linguistico-letterario nato al termine della confe-

renza al Vassar College (New York) nel 1987 al fine di sviluppare e mantenere una serie di linee guida di alta qualità per la codifica di testi umanistici. Il consorzio ha attualmente sede presso l'*Institute for Advanced Technology in the Humanities* della University of Virginia e ancora oggi lavora per fornire "la grammatica" dell'organizzazione semantico-strutturale dei documenti digitali così da migliorarne portabilità, archiviazione e gestione. All'inizio dell'avventura TEI, l'SGML (Standard Generalized MarkUp Language) venne utilizzato come linguaggio di marcatura conforme, mentre negli anni successivi al 1994 questo fu sostituito dal più aggiornato e versatile XML (Extensible MarkUp Language). Come deducibile dai rispettivi acronimi, sia SGML che XML possono essere considerati più o meno verosimilmente linguaggi di *MarkUp*: sistemi di metalinguaggio descrittivo atti a demarcare ed etichettare le singole parti di cui un documento elettronico è composto. In maniera del tutto pratica, quindi, l'utilizzo del linguaggio di marcatura diviene altresì cruciale per ciò che riguarda la preparazione di testi digitali da sottoporre ad analisi computazionale in quanto, nel momento in cui il documento viene processato dal computer, è proprio il sistema di tags e simboli descrittivi del *MarkUp* che comunica alla macchina come classificare e di conseguenza "trattare" i diversi elementi del testo. Per maggior chiarezza, ecco un esempio pratico di codifica secondo TEI Guidelines con i metadati iniziali relativi a *David Copperfield*:

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model
href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng"
type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model
href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng"
type="application/xml"
        schematypens="http://purl.oclc.org/dsdl/schematron"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="copperfield">
<teiHeader>
```

```
<fileDesc>
<titleStmt>
<title>David Copperfield</title>
<author>Charles Dickens</author>
</titleStmt>
<publicationStmt>
<p>A partial electronic edition based on a
<ref target="http://www.gutenberg.org/">Project Gutenberg</ref> transcription.</p>
</publicationStmt>
<sourceDesc>
<p><ref target="http://www.gutenberg.org/cache/epub/766/pg766.txt">Gutenberg
plain-text version</ref> of David Copperfield.</p>
</sourceDesc>
</fileDesc>
</teiHeader>
```

Senza voler scendere in dettagli troppo tecnici, credo comunque che una breve spiegazione degli elementi chiave della codifica TEI sia importante. Prima tra tutti la dichiarazione XML: un'istruzione speciale con cui s'identifica la tipologia del documento e la specifica versione di XML utilizzata di modo che il software che dovrà leggere ed elaborare il documento possa avere sufficienti informazioni e non generare bug o errori di compatibilità. Sempre all'interno della dichiarazione, l'attributo opzionale <encoding> segnala il tipo di codifica utilizzata nel documento, in questo caso la "UTF-8", mentre nella riga successiva l'elemento radice <TEI.2>, punto di partenza della nidificazione gerarchica di ogni documento XML, va a indicare le sezioni fondamentali dell'intestazione, <TeiHeader>, e del corpo del testo <text = "body">. Infine, il documento si compone di diversi marcatori strutturali atti a classificare le sezioni virtuali per la collocazione di immagini e contenuti, come nel caso dei o i paragrafi <p>.

```
<body>
<div>
<head>CHAPTER 1. I AM BORN</head>
<p>Whether I shall turn out to be the hero of my own life, or whether that station will be held by anybody else, these pages must show. To begin my life with the beginning of my life, I record that I was born (as I have been informed and believe) on a Friday, at twelve o'clock at night. It was remarked that the clock began to strike, and I began to cry, simultaneously.</p>
<p>In consideration of the day and hour of my birth, it was declared by the nurse, and by some sage women in the neighborhood who had taken a lively interest in me several months before there was any possibility of our becoming personally acquainted, first, that I was destined to be unlucky in life; and secondly, that I was privileged to see ghosts and spirits; both these gifts inevitably attaching, as they believed, to all unlucky infants of either gender, born towards the small hours on a Friday night.</p>
<p>I need say nothing here...<note rend="footnote">There are many more paragraphs in chapter 1.</note></p>
</div>
</body>
</text>
</TEI>
```

Esempio di testo codificato secondo TEI Guidelines (2):
Corpo del testo capitolo primo in *David Copperfield*

Al momento, sia i testi acquisiti attraverso costose collezioni private (HathiTrust Digital Library, ProQuest collections, The University of Oxford Text Archive) o scaricati gratuitamente da piattaforme e biblioteche *open source* (Internet Archive, Google books, Gutenberg Project) vengono provvisti solo di pochi essenziali tag e metadati e questo perché ogni ricerca, ogni investigazione critica, ha potenzialmente bisogno di un apparato descrittivo diverso in base al tipo di aspetti e unità d'analisi a cui si è interessati in virtù del tipo di studio da intraprendere. Lasciato quindi alla discrezione del singolo, il processo di marcatura semantica tramite aggiunta di tag personalizzati ha in sé un forte carattere interpretativo che

può essere espresso sia a livello "manuale", ossia con singole *entries* ad opera del ricercatore stesso, sia a livello automatico attraverso uno specifico script.[10] È chiaro, dunque, quanto ogni atto di codifica, ogni passaggio del testo da ente analogico a dato digitale, non debba essere inteso come una mera sequenza meccanica di operazioni preliminari a un esperimento, bensì come un processo di profonda ridefinizione dell'oggetto culturale sia a livello paradigmatico che ermeneutico.

Sulla base delle suddette premesse non ci resta che avventurarci in una ricognizione dell'odierno panorama di ricerche e sperimentazioni delle *Digital Humanities* in campo letterario, cercando al contempo di determinare quanto e se sussistano i termini affinché questa disciplina possa effettivamente contribuire in modo inedito ed esclusivo all'avanzamento degli studi letterari.

3. *Dal qualitativo al quantitativo: micro e macro analisi computazionali*

Quando si entra nell'ambito della metodologia quantitativa applicata alla ricerca letteraria, il critico non può prescindere da una semplice quanto vincolante premessa: tutto ciò che nelle scienze umanistiche è e può essere intuitivo, a partire dal concetto stesso di testo, in informatica deve, al contrario, divenire qualcosa di formalizzabile e formalizzato. Tale assunto emerge con particolare intensità da un recente articolo di Franco Moretti, da poco incluso nella serie di Pamphlet dello *Stanford Literary Lab*, intitolato *Operationalizing*. Dalle parole del celebre comparatista, possiamo definire l'operazionalismo quale "aim of measurement", un'intenzione di misurazione laddove l'importanza del concetto di misurabilità, preso in prestito dalla fisica e dalle *hard sciences* in generale, si configura quale

[10] Attraverso programmi di parsing e taggatura automatica come POS o Morph Adorner.

«means to make a concept actual»[11] attraverso una verifica di corrispondenza tra teoria e realtà. Con una sottile quanto impressionante equivalenza, Moretti illustra come la trasformazione di un concetto in una serie di operazioni quantificabili, e soprattutto testabili, non si configuri solamente quale condizione di possibilità per l'applicazione di strumenti computazionali alla ricerca letteraria (appunto la formalizzazione di cui sopra), ma anche quale essenza di una nuova teoria letteraria non più teleologicamente "theory-driven" ma "data-driven"; una modalità d'investigazione critica in cui il rapporto dialettico tra teoria e realtà diviene inevitabilmente oggetto di continua negoziazione, data appunto la natura non più elusiva della verifica empirica.[12]

In questo senso, iniziare a sfruttare appieno le potenzialità connaturate al mezzo informatico e abbracciare più ampie prospettive critiche[13] per Moretti significa spostarsi dall'usuale piano analitico-qualitativo del *close-reading*, in cui lo studioso si concentra nel dettaglio sulle singole parti componenti un certo fenomeno, ad un piano sintetico-quantitativo di *distant reading* o *macro-analysis*, in cui il tutto è maggiore della forma delle sue parti e l'oggetto di studio viene inteso ed esaminato nel suo insieme. Ovviamente, la possibilità del tutto inedita di esaminare l'enorme «empirical potential of the digital universe» in questo secondo orizzonte paradigmatico è un qualcosa che fa tremare le fondamenta della teoria letteraria.

Sì, perché se è vero che a livello di approccio analitico-qualitativo le Digital Humanities vantano una lunga tradizione di progetti ed esperi-

---

[11] «Operationalizing means building a bridge from concepts to measurement, and then to the world. In our case: from the concepts of literary theory, through some form of quantification, to literary texts.» F. Moretti, *Operationalizing,* "Pamphlets of the Stanford Literary Lab" 6, p. 1.

[13] Precisamente dall'articolo "World Literature" apparso sulla *New Left Review* nel 2000 e poi nel più celebre *Maps Graphs and Trees* del 2005.

menti di *close reading*, sia per quanto riguarda il campo delle *Digital Scholarly Editions* sia per le realizzazioni di studi critici di varia natura,[14] i contributi inseribili all'interno del modello sintetico-quantitativo della *distant reading* computazionale sono, al contrario, ancora pochi.

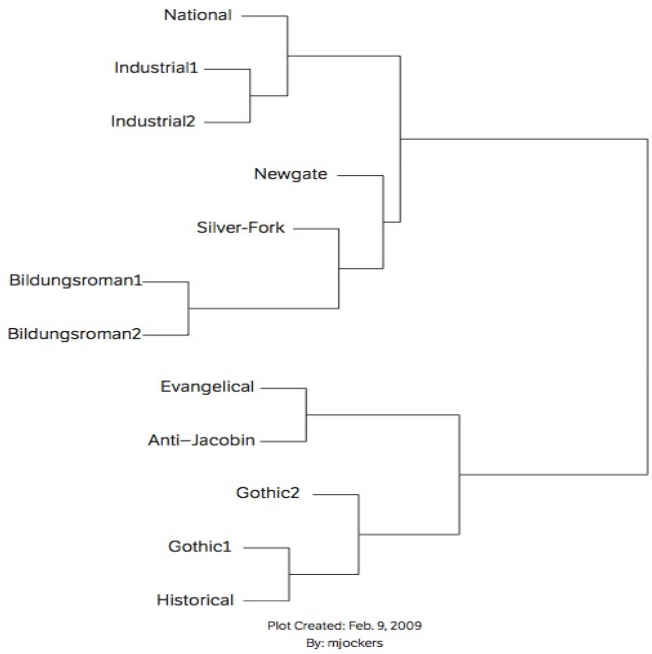
In quest'ultima sezione, procederò alla presentazione di progetti di ricerca realizzati attraverso i metodi e gli strumenti della macro-analisi. Il fine è quello di illustrare quali vantaggi questo tipo di ricerca comporta e in che modo essa possa dimostrarsi in grado di apportare nuova conoscenza nel campo delle scienze umane.

Partiamo dunque dalle origini, dallo stesso Moretti, e dagli avanguardistici lavori del suo *Stanford Literary Lab*; un gruppo di lavoro giovane e coeso che dal 2010 è impegnato nella ricerca sperimentale in campo letterario e che ogni anno desta sempre maggior curiosità e interesse accogliendo *visiting students* e *scholars* da tutto il mondo. I suoi contributi di ricerca sono accessibili online, direttamente dal sito http://litlab.stanford.edu/, sotto forma di pamphlet; una forma di pubblicazione assolutamente appropriata per una serie di libelli che, con un misto di passione incosciente e rigore scientifico, mirano a intaccare l'idea stessa di principi critici assoluti.

---

[14] Per Digital Scholarly Editions s'intendono tutti quei progetti il cui scopo è produrre una "rappresentazione" critica di documenti storici. In questo senso ogni modellizzazione e riproduzione digitale di documenti esistenti (da qui il significato dell'aggettivo "storici") non può prescindere da una procedura di transmedializzione (trasformazione da un medium all'altro) ottenuta attraverso l'utilizzo dello schema di codifica TEI. Tra i più celebri esempi di tali studi troviamo il Walt Whitman Archive della University of Nebraska-Lincoln (http://www.whitmanarchive.org/) e The William Blake Archive (http://www.blakearchive.org/blake/) ad opera di Ed. Morris Eaves, Robert N. Essick, e Joseph Viscomi. Per quanto riguarda, invece, i risultati prodotti dall'applicazione del text-mining su singole opere o comunque su piccoli corpora a criterio autoriale, è importante nominare i grandi studi monografici di J. F. Burrows sulla stilometria della Austen del 1987 (*Computation into Criticism: A Study of Jane Austen's Novels*) o quello di Jonathan Hope sull'authorship del corpus shakespeariano del 1994 (*The Authorship of Shakespeare's Plays: A Socio-linguistic Study*).

Quantitative Formalism è il titolo dell'esperimento inaugurale che ha visto impegnata la squadra di Moretti in collaborazione con Michael Witmore dalla University of Wisconsin. In questo pioneristico lavoro, datato 2009 ma pubblicato solo a partire del Gennaio 2011, si è testata l'ipotesi di una procedura informatica *unsupervised* (senza intervento umano) per il riconoscimento e la classificazione dei sottogeneri romanzeschi attraverso le MFW (most frequent words). Il campione di riferimento era composto da 48 romanzi appartenenti a 12 generi differenti: ognuno dei testi rappresentava un ben acclamato classico del genere d'appartenenza, per esempio *The Monk* per il gotico, *Daniel Deronda* per il romanzo di formazione o *Hard Times* per il romanzo industriale.



Con grande sorpresa degli addetti ai lavori, il computer riuscì a operare una *clusterizzazione* automatica perfetta. Perfetta in quanto totalmente coincidente alla classificazione della critica letteraria tradizionale da cui si era partiti. Ulteriore fonte di stupore, però, riguardava il "come" la macchina fosse giunta al medesimo risultato dell'uomo. Infatti, nel commen-

tare il brano a più alto coefficiente "gotico" selezionato dal computer, Sarah Allison, co-autrice del pamphlet, osserva:

> The gothic of Docuscope was different from that of "Humanscope" (as she called it): it was not the same gothic we saw. For us, that page was gothic because of the subdued terror and the archway, the ruin and apprehension and the limbs that trembled—not because of the "he" "his" "him" "had" "was" "struck the" and "heard the" which caught Docuscope's attention.[15]

Docuscope, il programma di MFW utilizzato, era stato in grado di classificare come appartenente al genere gotico un brano tratto da *A Sicilian Romance* (1790) della Radcliffe – romanzo gotico per antonomasia – ma ci era arrivato attraverso degli indicatori (articoli, pronomi e tempi verbali) totalmente diversi da quelli abitualmente considerati dall'uomo (elementi tematici come terrore, passaggi sotterranei ed eroine tremanti).

In termini di conclusioni questo esperimento aveva dimostrato essenzialmente due cose. La prima riguardava senz'altro la capacità da parte della macchina di "riconoscere", e quindi raggruppare, romanzi in base al genere senza alcun intervento umano. La seconda, conseguente alla prima e dalla portata del tutto inaspettata, riguardava il fatto di aver portato alla luce l'esistenza di un substrato di micro-unità formali e formalizzabili quali articoli, prefissi, pronomi e preposizioni che, al pari di più evidenti aspetti semantici, era in grado di fornire un netto segnale di genere.[16]

Altro contributo di assoluto pregio e interesse ad opera dello Stanford Literary Lab è poi il quarto pamphlet dal titolo: *A Quantitative Literary*

---

[15] S. Allison, R. Heuser, M. Jockers, F. Moretti, M.Witmore. *Quantitative Formalism: An Experiment*, "Pamphlets of the Stanford Literary Lab" 1.

[16] Queste stesse premesse sono poi diventate la base teorica e pratica su cui ho strutturato la mia tesi dottorale, pubblicata nel 2013 col titolo *Il gotico @ distanza, nuove prospettive nello studio dei generi letteraria*.

History of 2,958 Nineteenth-Century British Novels, the Semantic Cohort Method.[17]

Nello specifico, il *semantic cohort method* consisterebbe nell'applicazione di un generatore di campi semantici empirici, uno *script* affettuosamente ribattezzato *correlator* che a ogni parola chiave inserita collega tutti gli altri lemmi che, all'interno del corpus di riferimento, condividono la sua stessa frequenza d'uso e linea di tendenza cronologica. In questo modo, pur non trattandosi di un campo semantico in senso tradizionale, il *correlator* può comunque definirsi un generatore di campi semantici di tipo empirico proprio per questa sua capacità di mettere in relazione parole in base ad un criterio di contiguità e probabilità di co-occorrenza.

Scopo di questo quarto esperimento era dunque l'esplorazione del lessico del romanzo inglese dell'Ottocento in vista di una possibile "misurabilità" e misurazione dei suoi "major shifts". A livello di dati tangibili questi si tradussero in un chiaro declino dei *fields* quali "moral evaluation" e "abstract values", in favore di un altrettanto netto decollo del campo semantico delle "hard seed words" (ossia parole che condividono la stessa frequenza e tendenza cronologica d'occorrenza della parola chiave "hard", suddiviso in *action verbs*, *body parts*, *colors*, *numbers*, *locative definers* e *physical adjectives*).

[17] R. Heuser, L. Le-Khac. *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method* ,"Pamphlets of the Stanford Literary Lab" 4.

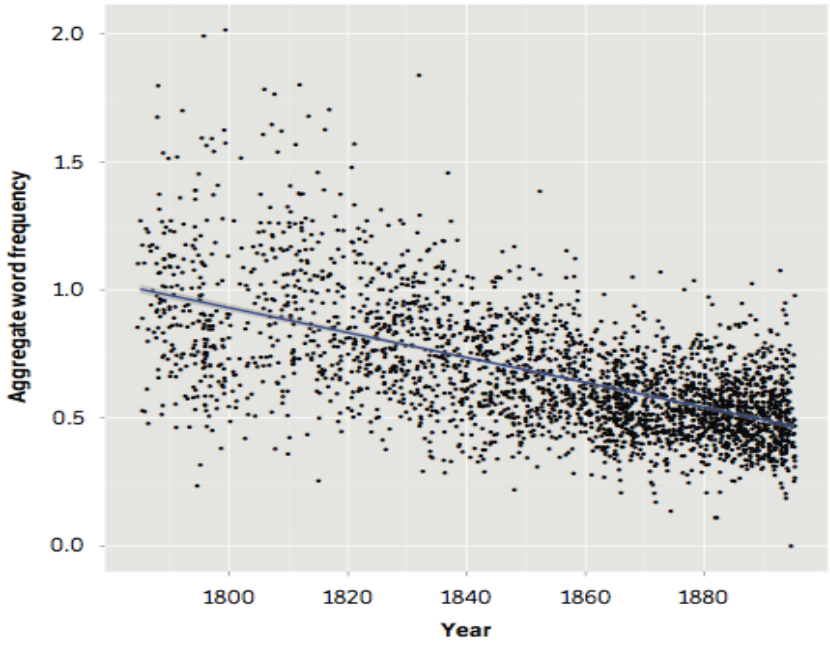


Figure 8: Aggregate term frequencies of the abstract values fields combined in novels, 1785-1900.

Field	[A] Percent of words in corpus	[B] Number of words after OED (stage 1.3)	[C] Number of words after filtering (stage 1.4)	[D] Average correlation coefficient	[E] Median correlation p-value
Action Verbs	1.99%	257	248	73%	.742%
Body Parts	0.65%	147	111	71%	.773%
Colors	0.13%	96	46	57%	6.16%
Locative Prepositions	1.09%	28	27	74%	.499%
Numbers	0.37%	46	44	73%	.679%
Physical Adjectives	0.20%	32	32	79%	.227%
Collectively	4.43%	606	508	71%	1.51%

Table 2: Magnitude, number of words, and correlation values for the hard seed fields.

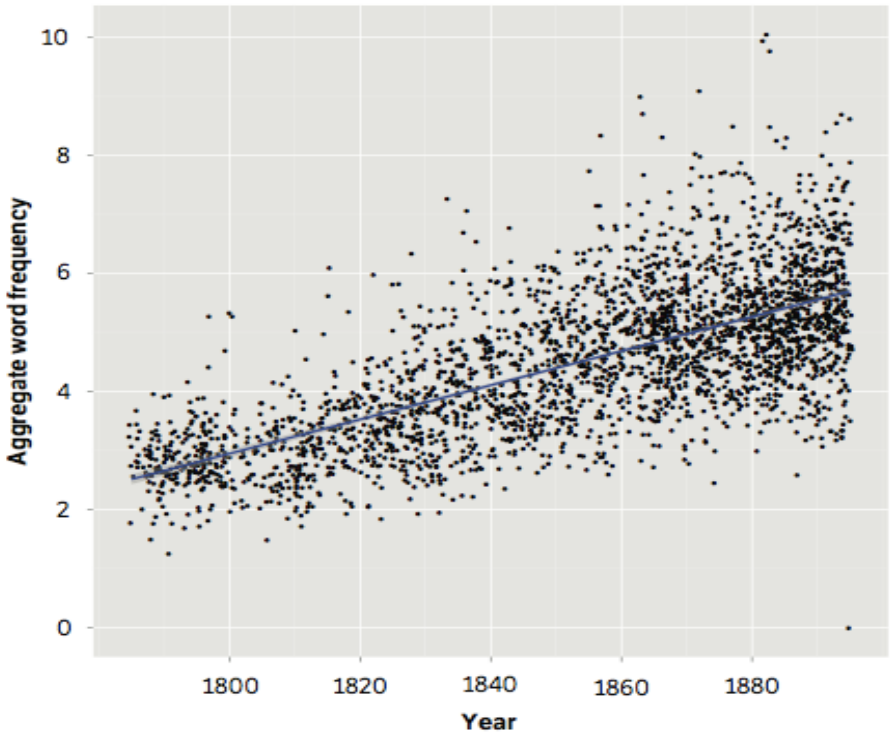


Figure 15: Aggregate term frequencies of the hard seed fields combined in novels, 1785-1900.

Uscendo per un momento dalle aule alternative dell'accademia californiana, ma senza allontanarsi – ahimé – dai confini a stelle e strisce, ci spostiamo verso il mid-west (nella fattispecie a Urbana-Champaign nell'Illinois e all'università Nebraska-Lincoln) dove la ricerca letteraria quantitativa su larga scala trova nelle personalità di Ted Underwood e Matthew Jockers altri importanti punti di riferimento. Entrambi giovani (o comunque sotto i cinquant'anni, come la locuzione "young scholar" ormai suggerisce), con un ottimo comando dei principali linguaggi di programmazione per il data-mining (tra cui Python, PHP e R) ed entrambi animati da un'utopica[18] inclinazione alla condivisione e alla divul-

[18] Con un briciolo di sano scetticismo, bisogna tenere sempre a mente che mai come in questa disciplina i dati, sia d'origine (ossia i corpora digitalizzati) che d'elaborazione (le tabelle e gli z-score), sono tutto. Diffiderei di chiunque in possesso di tali dati de-

gazione di strumenti e materiali, smania per altro del tutto connaturata allo stereotipo del *DH scholar* che i due professori sfogano nei rispettivi blogs.[19] Underwood e Jockers hanno di recente pubblicato due testi fondamentali nel campo della macro-analisi letteraria in cui accanto alla teoria si forniscono anche degli interessanti esperimenti pratici. In *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies,* ad esempio, Ted Underwood affronta empiricamente il problema della periodizzazione del romanzo inglese portando nuove evidenze circa l'emergere dei differenti sotto-generi della prosa. Parimenti, nel suo primo studio di divulgazione sulla disciplina, appunto intitolato *Macroanalysis Digital Methods and Literary History*, Matthew Jockers presenta una vasta panoramica di esperimenti di stilometria (idoletti autoriali, scritture di genere, classificazione di temi sulla base della nazionalità autoriale, analisi di tendenze lessicali su piano cronologico) e *topic modeling* dimostrando cosa un database di circa 3000 opere di finzione e delle buone "domande" critiche possano generare.

Interessante notare come, proprio a dimostrazione del carattere ancora inedito di questo tipo di studi, fin dalle prime pagine del suo libro, il critico senta in qualche modo il bisogno di giustificare al lettore, e anche in maniera piuttosto articolata, l'esistenza stessa dell'approccio macroanalitico agli studi letterari. In particolare, utilizzando la metafora dell'economia quale disciplina dipendente da entrambe le sue componenti per l'appunto micro e macro economiche, Jockers rinuncia all'occasione di poter legittimare *tout-court* la valenza della macro-analisi computazionale come paradigma critico a sé stante portando avanti piut-

---

cidesse di metterli indiscriminatamente a disposizione del pubblico bruciando, di fatto, la possibilità di un lavoro inedito a priori.

[19] Quello più istituzionale di Matthew Jockers (http://www.matthewjockers.net/) e quello più poetico di Ted Underwood intitolato *The Stone and the Shell* (http://tedunderwood.com/).

tosto un discorso incentrato sull'aspetto sinergico di "interplay" tra le due metodologie di "close" e "distant" reading.

The approach of the study of literature that I am calling macroanalysis, instead of distant-reading (for reasons explained below) is in general ways akin to the social science of economics or, more specifically, macroeconomics. Before the 1930s there wasn't a defined field of "Macroeconomics." There was, however, microeconomics, which studies the economic behavior of individual consumers and individual businesses. As such, microeconomics can be seen as analogous to the study of individual texts via "close-readings" of the material. […]

Micro-oriented approaches to literature, highly interpretative readings of literature, remain fundamentally important. Just as microeconomics offers important perspectives on the economy. It is the exact interplay between macro and micro scale that promises a new, enhanced and perhaps even better understanding of the literary record. The two approaches work in tandem and inform each other. Human interpretation of the "data", whether it be mined at macro or micro level, remains essential. While methods of enquiry, of evidence gathering, are different, they are not antithetical, and they share the same ultimate goal of informing our understanding of the literary record.[20]

Con il dovuto rispetto per l'*escamotage* tanto originale quanto "politically correct" di Jockers, ben conscio che quando si parla di *distant reading* il livello di polemiche da ogni versante della critica internazionale può arrivare a picchi insostenibili, ipotizzare un "interscambio" o "sinergia" tra due orizzonti epistemologici differenti – non quindi due approcci, ma due paradigmi – sarebbe davvero difficile.

[20] M. Jockers, *Macroanalysis: Digital Methods and Literary History*, University of Illinois Press, Urbana, 2013, p. 35.

4. *Conclusione polemica*

Se in tanti anni di passione per la letteratura e le vite che essa può raccontare ho capito che gli intrecci più riusciti, le trame più belle, nascono da un atto di trasgressione, ebbene la storia delle Digital Humanities nasce, al contrario, dal tentativo di ricomporne una: l'atto sovversivo di quel lontano 30 novembre 1609 in cui Galileo Galilei puntando al cielo notturno il suo cannocchiale non solo scoprì il volto imperfetto della luna, ma cancellò d'un colpo il tradizionale percorso congiunto dei saperi umanistico-scientifici. In questa prospettiva, le evidenze empiriche delle osservazioni galileiane costituiscono una sorta d'improrogabile atto di scissione tra *hard sciences* and *soft sciences* in cui le misurazioni sperimentali del metodo scientifico deprivano la filosofia e le altre scienze umane di un necessario tratto di oggettività.

Ecco dunque che le Digital Humanities, viste soprattutto dal punto di vista degli approcci macro-analitici, cercano di guadagnare l'importante statuto di punto di fusione tra due mondi da tempo inconciliabili e poco importa se i progetti e gli esperimenti finora esposti convincano chi legge di aver effettivamente apportato un contributo inedito di "nuova conoscenza" nel campo degli studi letterari. Qualora, infatti, la risposta dovesse propendere per il no e quindi il lettore giudichi che, in realtà, le Digital Humanities propongano una modalità di ricerca che non faccia altro che confermare qualcosa di già noto e divulgato dalla critica tradizionale, allora ci si dovrebbe porre il problema di come considerare tali convalide. Sono esse una forma mascherata di fallimento della ricerca quantitativa rispetto a quella qualitativa? E su quali basi è possibile stabilire questo? Quali sono gli strumenti per valutare l'apporto della ricerca quantitativa nel panorama acquisito degli studi letterari? Evitando di cedere a facili allarmismi, spesso risultanti in altrettanto facili generalizzazioni, credo sia

opportuno considerare tali interrogativi, ancora una volta, da un punto di vista epistemologico identificando nei principi di "falsificabilità"[21] e, con un orribile traduzione del termine morettiano, "operazionabilità" della ricerca letteraria di tipo quantitativo la cifra della sua validità.

Tristemente, per quanto convincente questa argomentazione potrà finalmente risultare, continuano comunque a echeggiare nella mia testa parole come quelle di Thomas Rommel, uno dei massimi esperti europei di DH, secondo cui «literary computing has, right from the very beginning, never really made an impact on mainstream scholarship», o come quelle di Stephen Ramsay, associate professor all'università di Lincoln-Nebraska:

> The digital revolution, for all its wonders, has not penetrated the core activity of literary studies, which, despite numerous revolution of a more epistemological nature, remains mostly concerned with the interpretative analysis of written cultural artifacts. Texts are browsed, searched, and disseminated by all but the most hardened Luddites in literary studies, but seldom are they transformed algorithmically as a means of gaining entry to the deliberately and self-consciously subjective act of critical interpretation.[22]

Se, a oggi, perfino nell'opinione dei più illustri esponenti della comunità scientifica delle Digital Humanities queste devono ancora dimostrare sul campo la possibilità di apportare un plusvalore nella ricerca umanistica sulla base della loro capacità di penetrazione nella soggettività dell'atto interpretativo umano, allora il problema è davvero complesso e forse irrisolvibile. Parafrasando il celebre aforisma di Albert Einstein sul genio, "se si giudica un pesce dalla sua abilità di arrampicarsi sugli alberi, quello

---

[21] Nonostante l'apparente paradosso, infatti, per essere controllabile e quindi considerabile scientifica, ogni ricerca deve contenere nelle sue premesse le condizioni di possibilità affinché un secondo esperimento possa dimostrarla integralmente falsa.

[22] Ramsay S., "Algoritmic Criticism" in *A Companion to Digital Humanities*, ed. Schreibman S., Siemens R., Unsworth J., Blackwell Publishing, Oxford, 2004.

passerà tutta la vita a credersi stupido," ovvero se continuiamo a chiedere alle Digital Humanities, ma soprattutto ad aspettarci da loro, risultati ascrivibili ad una sfera ermeneutica del tutto umana, esse non ci serviranno mai a nulla.

Comunque, volendo chiudere con una nota di speranza e in guisa di una previsione per il futuro, credo che da qui ai prossimi cinque anni, la grande partita per la legittimazione delle Digital Humanities nell'olimpo dei saperi accademici si giocherà nell'ambito degli strumenti e dei progetti di ricerca.[23] Essere preparati a investire in sperimentazioni e analisi dagli esiti non sempre sicuri, evitando di cedere alle facili lusinghe di un'applicazione della disciplina al solo campo dell'archiviazione, della conservazione o delle *scholarly editions* sarà fondamentale. Se è vero, infatti, che biblioteche virtuali e altre forme di custodia del patrimonio in formato digitale sono sempre progetti piuttosto proficui, e non solo a livello intellettuale, è altrettanto vero che questi si rivelerebbero uno spreco in termini di opportunità qualora i dati non potessero essere esplorati e studiati. Sarà dunque di vitale importanza che, per quell'epoca, la futura così come la presente generazione di umanisti digitali impari finalmente a porre ai testi e ai computers le domande giuste. D'altronde, come Matthew Jockers stesso racconta nel suo blog:

> I came to utilize computation in my research not because the siren's song of revolution was tempting me away from my dusty, tired, and antiquated approaches to literature. Rather, computational tools and statistical methods simply offered a way of asking and exploring the questions that I (and others such as those pictured above) have about the literary field. What has changed is not the object of study but the nature of the questions. So, the answer to my colleague who asked what is needed to "break into this field of Digital Humanities" is simply this: "questions, you need questions."

---

[23] In questo ambito, oltre agli Stati Uniti, la ricerca made in Canada è davvero avanti. Da oltre vent'anni l'università di Alberta e Victoria lavorano sulla costruzione di tools e piattaforme gratuite per text mining. Un esempio tra tutti è il progetto Text Analysis PORtal (http://www.tapor.ca/).

Bibliografia

A.A.V.V. (2009), *What's in a word list, Investigating Word Frequency and Key word extraction*, Ashgate, Faraham.

Bollier D. (2001),*The Promise and Peril of Big Data,* The Aspen Institute Publication Office, Washington DC.

Burrows J. F. (1987), *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford University Press, New York.

Heuser R., Le-Khac L. (2012), *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, "Pamphlets of the Stanford Literary Lab" 4.

Jockers M. (2013), *Macroanalysis: Digital Methods and Literary History*, University of Illinois Press, Chicago.

Moretti F. (2011), *Quantitative Formalism: an experiment*, 2011, "Pamphlets of the Stanford Literary Lab" 1.

Moretti F. (2014), *Operationalizing,* 2014, "Pamphlets of the Stanford Literary Lab" 6.

Schreibman S., Siemens R., Unsworth (eds.) (2004), *A Companion to Digital Humanities*, J., Blackwell Publishing, Oxford.