DigitCult | Scientific Journal on Digital Cultures

Dieci anni di libri
Autobiografia per dati di un lettore forte, 2008-2017

Andrea Zanni
MediaLibraryOnLine

Abstract

Il presente articolo è un caso di studio riguardante l'analisi statistica dell'attività di un singolo lettore in dieci anni, partendo dai dati memorizzati sulle piattaforme di social reading e analizzandoli tramite una serie di strumenti, che vanno dal raffinamento, all'analisi statistica e alla visualizzazione. In un esempio di *data storytelling*, è stata compiuta una ricognizione autobiografica analizzando tutti i metadati bibliografici dei libri letti da gennaio 2008 a dicembre 2017, cercando di ricavare statistiche e grafici atti a fornire una fotografia dettagliata, seppur incompleta, della vita di un lettore forte.

Ten Years of Books
Statistical Analysis on the Activity of a Single Reader, 2008-2017

The article is an exploration regarding regarding the reading activity of a single reader in the last ten years. All the bibliographical metadata from books read from January 2008 to December 2017 have been analyzed, developing stats and charts in order to provide a detailed, albeit incomplete, overview of the life of a "strong reader".

Published 24 September 2018

Correspondence should be addressed to Andrea Zanni, MediaLibraryOnLine, via della vigna 2, Bastiglia (MO). Email: zanni.andrea84@gmail.com

DigitCult, Scientific Journal on Digital Cultures is an academic journal of international scope, peer-reviewed and open access, aiming to value international research and to present current debate on digital culture, technological innovation and social change. ISSN: 2531-5994. URL: http://www.digitcult.it

 doi:10.4399/97888255181535

Introduzione

Le statistiche della lettura

Il mondo delle statistiche che riguardano la lettura è ampio[1]: si suddivide principalmente in statistiche sui *libri*, spesso analisi commerciali che tengono traccia di quanti e quali libri vengono venduti nelle librerie, e statistiche sui *lettori*. Fra queste ultime, le analisi più famose e utilizzate sono quelle ISTAT, che usano i dati di vendita delle librerie[2] o più raramente i dati di prestito delle biblioteche[3], e che hanno un approccio quantitativo, cioè orientato a suddividere la popolazione in fasce di lettura per numero di libri: lettori forti, medi o deboli.

Questo tipo di analisi è importante ma *superficiale*: per ogni lettore conta il numero di libri letti (o comprati, o presi in prestito) durante un periodo limitato di tempo. Non c'è alcun tentativo di carotaggio *verticale* sulla storia dei lettori, cioè capire quanti e quali libri hanno letto prima, cercando di creare un contesto e analizzare l'attività di lettura a tutto tondo.

L'analisi della lettura è una attività complessa, perché la lettura è un sistema complesso (Faggiolani, Verna e Vivarelli 2017). Editori, distributori, librerie sono storicamente quelli che più investono in analisi e ricerche di mercato: essendo entità *for profit* sono obbligate ad operare sui *dati di vendita dei singoli libri*, elaborando strategie a breve termine per mantenere ciascuna il proprio business e generare profitti. Per questo agenzie quali Nielsen[4] o GFK[5] offrono a pagamento ricerche e analisi a livello settimanale.

La filiera del libro potrebbe godere di maggiori informazioni, perché a tutti gioverebbe una fotografia più dettagliata dell'intero *ecosistema* della lettura in Italia, considerando non solo i singoli libri ma anche i singoli lettori. Soprattutto, esiste una parte importante ma poco considerata della filiera che consiste di operatori pubblici, come scuole o biblioteche, o "privati" quali insegnanti e genitori, che possedendo minori risorse non può permettersi analisi di mercato, e possiede dunque minori informazioni.

Eppure molti dati sono in realtà presenti, anche se non sempre accessibili: storicamente, le biblioteche hanno conservato un'enorme mole di dati relativa ai prestiti di libri ai propri utenti, avendo quindi anche uno storico dell'attività di lettura di una comunità e dei singoli utenti. Per la privacy, questi dati molto spesso non sono pubblici, e anche quando lo sono (magari rilasciati come *open data* in qualche portale opportuno), lo sono in forma *aggregata*. Il dato aggregato è un dato utile all'amministrazione per capire l'impatto di una biblioteca, ma è una metrica molto grezza a fini di *ricerca* e *analisi*, poiché non entra nel dettaglio di quali tipo di libri siano stati prestati, o a quale segmento demografico.

Le biblioteche dunque rimangono una miniera d'oro inesplorata per quel che riguarda i cosiddetti *dati transazionali*, intendendo:

- i movimenti di documenti fisici e digitali registrati dalle biblioteche del sistema,
- le iscrizioni degli utenti
- le prenotazioni e le richieste online tramite il portale.

Amministrazioni virtuose come quella di Roma[6] hanno recentemente istituito un portale in cui è possibile scaricare la lista di tutti i movimenti delle biblioteche romane, coi dati anagrafici opportunamente anonimizzati.

Goldin (2018) analizza questo dataset illustrando molto bene il livello di dettaglio verso il quale un'analisi di questi dati si può spingere, costruendo un utile punto di partenza per analisi

[1] Per un approfondimento, il Centro del Libro e della Lettura produce ogni anno un report sul mondo del libro: http://www.cepell.it/wp-content/uploads/2016/04/La-produzione-e-la-lettura-di-libri-in-Italia-%E2%80%93-dati-2016.pdf

[2] ISTAT, 2016. Aspetti della vita quotidiana: Quotidiani e libri - sesso, età, titolo di studio. http://dati.istat.it/Index.aspx?QueryId=22373

[3] ISTAT, 2016. Biblioteche pubbliche statali. http://dati.istat.it/Index.aspx?QueryId=22037

[4] Nielsen: http://www.nielsen.com/it/it.html

[5] GFK: https://www.gfk.com/it/

[6] Portale Open Data delle Biblioteche di Roma: https://www.bibliotechediroma.it/it/open-data-biblioteche-di-roma

future. È opinione di chi scrive che il mondo bibliotecario, nonostante gli sforzi già profusi, potrebbe fare di ancora di più per il movimento open data (Berners-Lee 2001, 2012), rilasciando apertamente e sistematicamente anche i dati transazionali, e non soltanto i semplici dati aggregati. L'esperienza romana è un ottimo inizio.

Le piattaforme social

Per fortuna, la rivoluzione digitale ha cambiato le cose anche nel mondo del libro, allargando la filiera del libro ad altri attori e permettendo la costruzione di nuovi, interessanti *dati di lettura*. L'avvento di social network di lettura quali aNobii e Goodreads[7] ha infatti reso possibile costruire comunità di lettori, che costruiscono sia un enorme catalogo comune di libri (paragonabile in numero ai libri delle grandi biblioteche, o anche dell'indice SBN), sia reti di libri e lettori raggruppati per affinità e vicinanza. Come illustrato in precedenza da Faggiolani et al., questi dati sono una grande ricchezza per l'analisi e la ricerca riguardo l'ecosistema della lettura. In più, questi social hanno reso molto più semplice per il lettore medio compiere un'azione che prima era riservata ai lettori più motivati: *tenere traccia dei libri letti*, tenere traccia della propria biblioteca personale. Quest'operazione viene tuttora spesso fatta da alcuni lettori forti su carta, su taccuini personali oppure, per i più avanzati tecnologicamente, fogli Excel che rimangono privati.

Costruire la propria biblioteca personale su queste piattaforme ha di fatto reso possibile creare vere e proprie cataloghi e *bibliografie* che possono essere esportate e studiate, permettendo quindi analisi sì più limitate ma anche più *profonde* e verticali, e che permettono di osservare più nel dettaglio l'attività di un gruppo di lettori o anche un lettore singolo. Questo può rende possibile un aggiornamento di una pratica con una lunga tradizione: il racconto diaristico dei libri letti.

Data storytelling

L'utilizzo dei dati per fare giornalismo e raccontare una storia (l'inevitabile neologismo è *data storytelling*) è una pratica recente ma che raccoglie precedenti illustri, come per esempio i lavori del creatore della libreria grafica D3 Mike Bostock[8] o il blogger e statistico Nate Silver[9]: l'unione di strumenti di visualizzazione sempre più avanzati, dati strutturati e piattaforme di pubblicazione digitali rende sempre più facile la costruzione di reportage e saggi basati su analisi dei dati, visualizzate tramite grafici e infografiche. Il documento finale risulta spesso un ibrido fra *personal essay* e articolo scientifico, prendendo in prestito lo stile dal primo e le tecniche di analisi dal secondo genere.

Il presente articolo nasce in questo modo: una precedente versione di questo articolo, dal tono più informale, è stata pubblicata come auto-analisi personale su Medium (Zanni 2018). L'articolo è un caso di studio riguardante l'analisi statistica dell'attività di un singolo lettore, partendo dai dati memorizzati sulle piattaforme di social reading e analizzandoli tramite una serie di strumenti, che vanno dal raffinamento, all'analisi statistica e alla visualizzazione. È per definizione un'analisi che non può avere un valore statistico, dato che rappresenta un campione composto da *una sola* persona, ma è piuttosto da considerarsi una prima esplorazione di una serie storica di dati: nonostante il dataset in questione sia unico e di poca rilevanza scientifica in quanto tale, il procedimento utilizzato è *riproducibile*, con le dovute modifiche, per ogni catalogo bibliografico, sia esso un catalogo personale, il catalogo di una biblioteca scolastica, accademica o di pubblica lettura; il catalogo di una casa editrice; il catalogo dei libri letti da una classe o una scuola in un semestre; la bibliografia su un determinato argomento.

Lo scopo del saggio è dunque fornire un'esperienza pilota di analisi quantitative che possono essere proposte con un protocollo simile per cataloghi e bibliografie differenti, non trattandosi altro che analisi statistiche, raffinamenti e visualizzazioni di metadati bibliografici.

---

7 ANobii è uno dei primi social network dedicati ai libri, fondato nel 2006 a Hong Kong da Greg Sung, e che è ha avuto una travagliata storia di cambiamenti di proprietà. Attualmente è di proprietà della Mondadori. Goodreads è un social network molto simile, di proprietà di Amazon dal 2013.

8 Si veda a titolo di esempio, il saggio *Visualizing Algorithms*, 2014.

9 *FiveThirtyEight*: https://fivethirtyeight.com

 doi:10.4399/97888255181535

Nota metodologica

Il dataset utilizzato comprendeva i metadati bibliografici di tutti i libri letti dall'utente dal gennaio 2008 al dicembre 2017, e che sono stati tracciati sulla piattaforma aNobii.

I dati sono stati:

- esportati in formato CSV dalla piattaforma aNobii
- puliti e corretti con OpenRefine[10]
- arricchiti con Wikidata
- visualizzati e analizzati con Google Fogli e RAW[11].

Il dataset è liberamente consultabile[12]. Per semplificare l'analisi e la visualizzazione, sono state compiute alcune scelte binarie:

- per tutti i libri è stata usata la *data di fine lettura* o, quando non possibile, la *data di inizio lettura*;
- i libri possono essere "finiti" o "non-finiti": nella seconda categoria sono compresi libri a diversi stadi di lettura (libri solo iniziati, libri abbandonati, ecc.), mentre la prima categoria è ben definita (solo i libri davvero finiti);
- i libri possono essere o "fiction" o "non-fiction": tutto ciò che non è strettamente narrativo, romanzo o racconto è compreso nella seconda categoria;
- ci sono due modi di contare i libri: contare ogni libro singolo come uno, o utilizzare il suo numero di pagine. Entrambi i modi hanno vantaggi e svantaggi: il numero di libri è più semplice e immediato, ma la suddivisione in pagine può rendere conto meglio di quanto tempo si è dedicato a ciascun libro. Proprio per questa intrinseca diversità, sono stati usati entrambi gli approcci, quando ritenuti opportuni.

[10] OpenRefine è un software open source per la pulizia e la bonifica di dati, utilizzato da diverse comunità di professionisti dell'informazione. Scaricabile all'URL: http://openrefine.org

[11] RAW è una webapp di visualizzazione sviluppata da un team del DensityDesign del Politecnico di Milano. Permette la creazione di visualizzazioni e infografiche complesse a partire da dati tabellari. È liberamente utilizzabile all'URL: http://app.rawgraphs.io

[12] Il dataset originale è disponibile all'URL: https://docs.google.com/spreadsheets/u/1/d/1VhVqIXzMhsWy8O9FDnkpEIXXfYnWqNqVyjEKDIwyRME/edit?usp=sharing

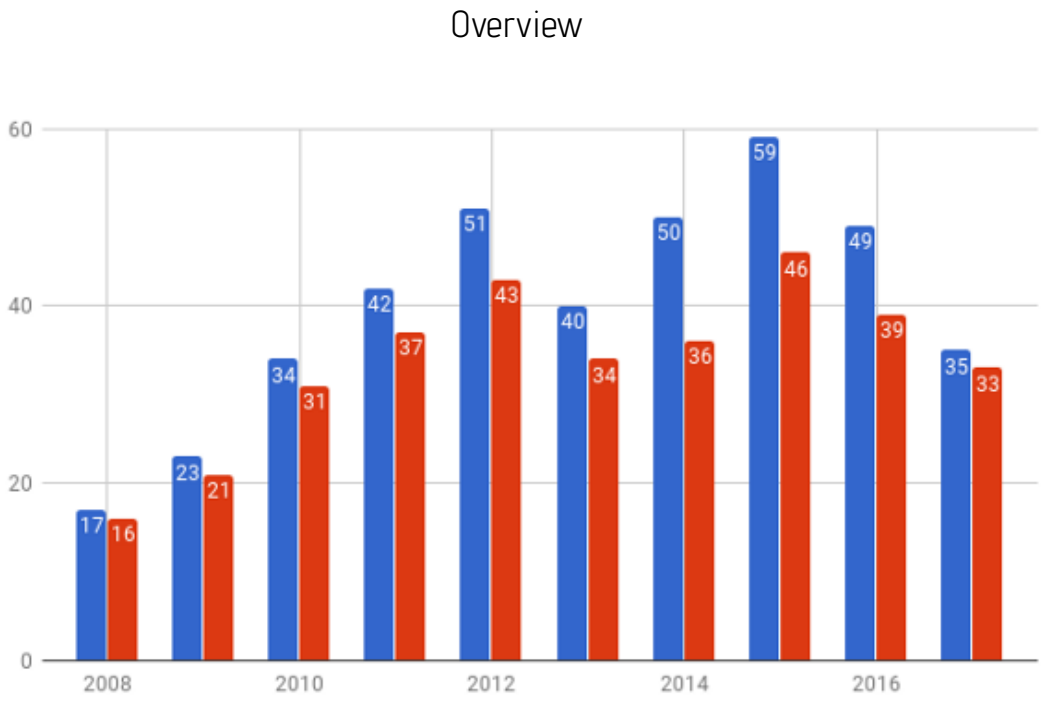

Figura 1. In blu i libri totali letti, in rosso i libri finiti.

I libri letti sono stati 467: finiti 332, mentre non-finiti 135, cioè un 28% del totale. La media dei libri finiti è 33,2 libri l'anno, cioè sono stati letti quasi tre libri al mese (anche se la distribuzione temporale di lettura è molto varia e non segue affatto la media). Secondo la suddivisione dell'ISTAT, dunque, si tratta di un "lettore forte", anche se per entrare in questa categoria basta leggere circa un solo libro al mese. Su dieci anni la divisione è 61% non-fiction, 29% fiction, mentre il dato non è stato osservato anno per anno.

Figura 2. In rosso fiction, in ciano non-fiction.

 doi:10.4399/97888255181535

Ebook

Figura 3. A sinistra gli ebook in ciano, a destra in rosso i libri di carta. F sta per fiction, N per non-fiction.

Riguardo il supporto di lettura, cioè ebook e libro cartaceo, la preponderanza della carta è netta, con un 87% (rosso nel grafico). È anche possibile osservare dai dati che l'ebook è il supporto preferito per la saggistica e i libri in inglese, con una particolare concentrazione di ebook durante il periodo estivo: portare l'ebook reader in vacanza è molto comodo. Dei 64 libri letti in ebook (blu nel grafico), quattro su cinque quindi sono infatti non-fiction (*N* nel grafico): di questi quattro quinti, metà sono in italiano metà sono in inglese. Si nota inoltre come libri in lingua inglese vengano letti solo su supporto digitale e mai su carta, a parte qualche rara eccezione.

Date di edizione

Le date di edizione dei libri tendono a non essere precise: quello a cui si fa riferimento è la data presente sul libro, intendendo l'edizione precisa che il catalogatore che l'ha inserita su aNobii aveva in mano. Questo significa che se si sta leggendo una ristampa recente (che sia su carta o digitale) di un libro dell'Ottocento, la data pubblicata sarà quella recente della pubblicazione, e non quella dell'edizione originale dell'opera. Purtroppo questo è un problema classico della biblioteconomia, e non è facile ottenere le date "originali".

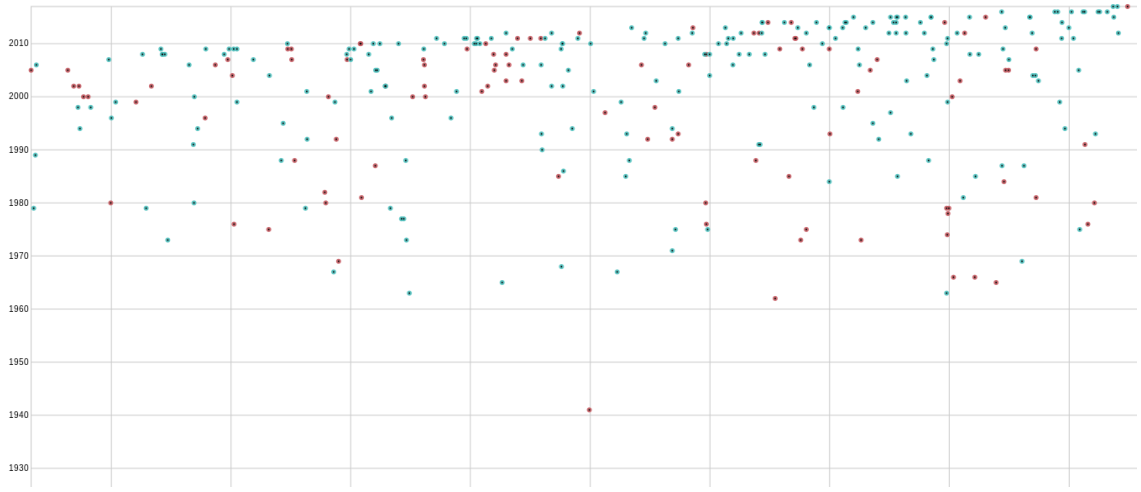

Figura 4. Distribuzione temporale dei libri letti, per data di edizione (asse verticale) e anno di lettura (asse orizzontale).

Contando tutto questo, è abbastanza evidente che vi sia una certa attenzione a libri e autori del passato: in un anno di lettura (rappresentato nel grafico dalle colonne) sono presenti libri pubblicati in vari decenni, con una prevalenza legata ai libri pubblicati dopo il 2000 o il 2010 (nel grafico, le prime due righe).

Questo aspetto temporale si può si osservare ancora meglio guardando agli autori. Tutti gli autori sono stati manualmente *riconciliati* con Wikidata: questo significa che, per ogni autore, sono stati recuperati da Wikidata alcuni dati anagrafici come sesso, nazionalità, data di nascita, eventuale data di morte. Questo ci ha permesso di vedere che la distribuzione fra autori morti e viventi è quasi simmetrica, con il 52% e 48% rispettivamente. Questo aspetto si può spiegare anche grazie al fatto che tendenzialmente il lettore analizzato compra *esclusivamente* in librerie e bancarelle dell'usato, per cui è normale che l'orizzonte letterario sia più spostato verso il passato (quello che in editoria chiamano il catalogo) piuttosto che verso le novità.

Esplorando meglio questa differenza, si può notare inoltre come sia abbastanza raro che un libro venga letto nel suo anno di pubblicazione, cioè quando è una novità in senso editoriale: succede 3–4 volte l'anno, poco più di 1 su 10. Una risposta parziale potrebbe essere che, comprando usato, ci sia ovviamente un periodo di latenza per cui un libro passa dallo scaffale delle novità allo scaffale delle occasioni, o direttamente alla bancarella dell'usato. Sarebbe interessante osservare questo tipo di comportamento in lettori diversi, che sono soliti comprare in librerie del nuovo.

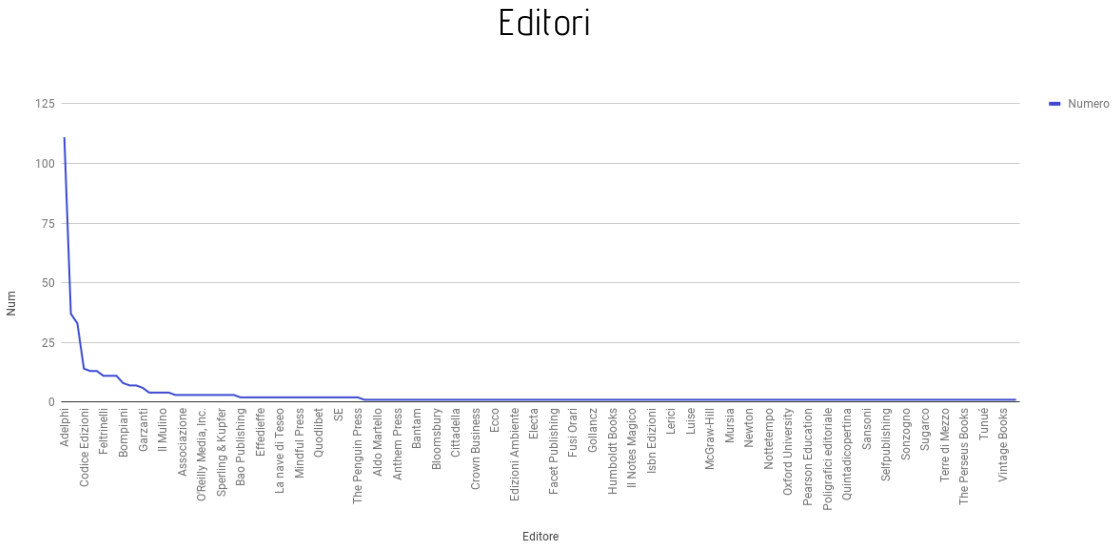

Figura 5. Editori.

Gli editori presenti sono 147: Adelphi è nettamente prima, con 112 libri, seguita in serie più omogenea da Einaudi (37), Mondadori (33), Codice (14), Franco Maria Ricci (13). I primi 7 editori equivalgono, in numero di libri, agli altri 140: una definizione quasi esatta di *legge di potenza*[13], o distribuzione paretiana, fortemente asimmetrica e caratterizzata da pochissimi elementi con moltissime occorrenze e una coda lunga[14] (la cosiddetta *long tail*) di elementi con una o poche occorrenze. Questo tipo di distribuzione, opposta alla classica *distribuzione normale* (la classica "curva a campana") si ritrova molto spesso nell'ambito dell'editoria (Greco 2014, 4). Seguono una distribuzione paretiana, infatti:

- i libri venduti in libreria o prestati in biblioteca, dove abbiamo pochissimi best seller e una lunga coda di libri venduti in pochissime copie;

- gli autori più venduti e letti: un autore di bestseller può vendere letteralmente decine di milioni di libri, mentre l'autore medio si ferma sulle poche migliaia o addirittura centinaia;

[13] Per un approfondimento, vedere: http://it.wikipedia.org/wiki/Legge_di_potenza
[14] Per un approfondimento, vedere: http://it.wikipedia.org/wiki/Coda_lunga

 doi:10.4399/97888255181535

- la prolificità degli autori: scrittori come Georges Simenon hanno pubblicato decine di libri, mentre la maggior parte degli autori si ferma ad uno solo.

È interessante osservare che questo tipo di curve a coda lunga sono presenti anche con un campione statistico di poche centinaia di libri, che fa riferimento ad un solo lettore.

Questa discrepanza fra l'editore più letto e gli altri è abbastanza stupefacente: i libri editi da Adelphi sono tre volte più numerosi del secondo editore. Allo stesso modo, è curioso vedere una così lunga coda di editori con solo uno o due libri. Da questo punto di vista, siamo di fronte ad una domanda interessante, a cui non è mai stata data (forse) risposta: come leggono i lettori italiani? Qual è il comportamento medio rispetto agli editori? È normale leggere così pochi editori tante volte, e allo stesso tempo così tanti diversi editori? Non avere una *lettore medio* come benchmark non aiuta a capire se questo tipo di lettura così caratterizzato sia "normale" o meno.

Adelphi

Dati i numeri significativi, Adelphi merita un breve approfondimento. I libri letti sono 112, per un totale di 26091 pagine.

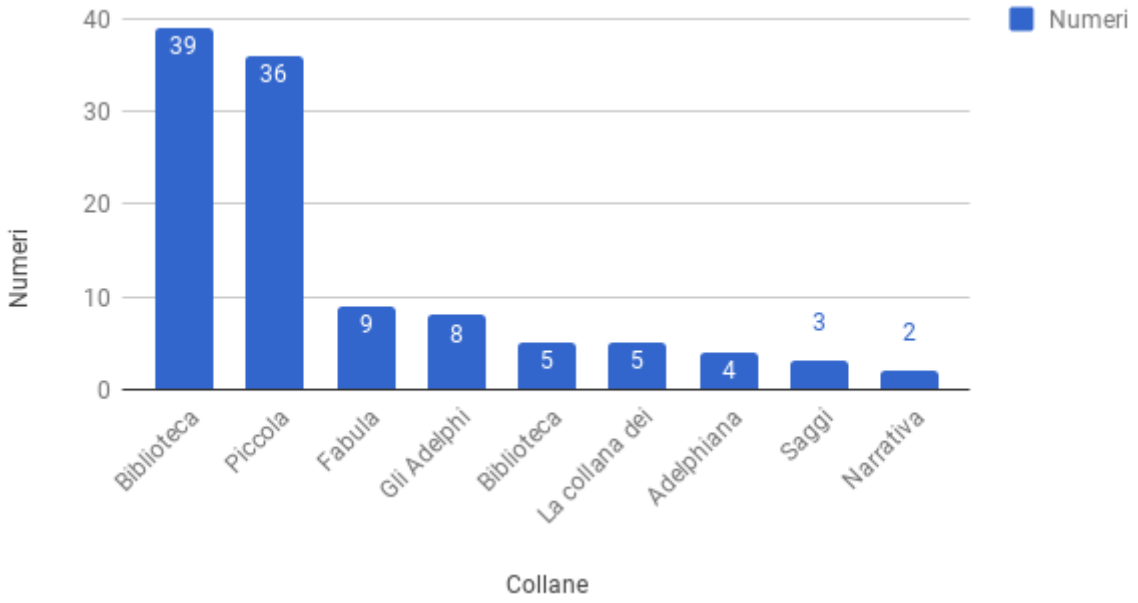

Figura 6. Collane Adelphi

Tenendo traccia anche delle singole collane, si può notare come le più presenti siano *Biblioteca* e la *Piccola Biblioteca*, rispettivamente con 39 e 36 libri. Più sotto, *Fabula* (collana di narrativa, composta di romanzi veri e propri), *Gli Adelphi* (la collana economica, fatta quasi esclusivamente di ristampe), sino ad andare a cose più specifiche come i *Saggi*, la *Narrativa Contemporanea* (collana delle degli anni '60, poi evolutasi in *Fabula*) fino ad arrivare alle *Adelphiana* e *La collana dei casi*.

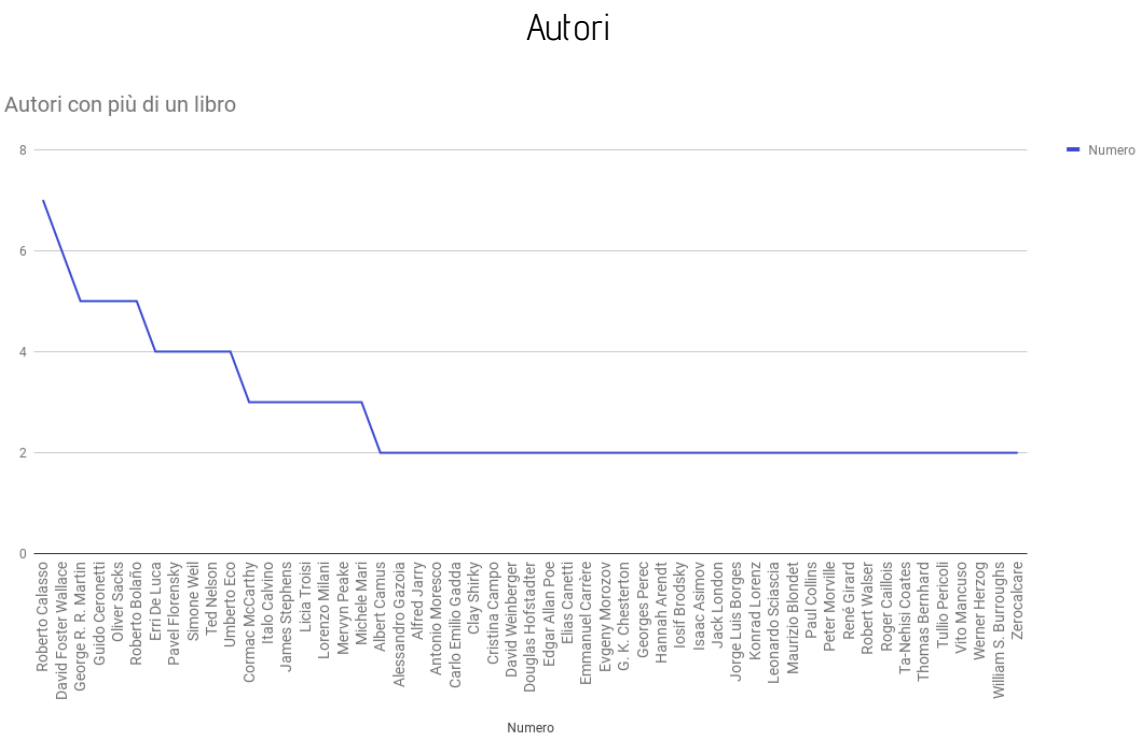

Figura 7. Autori più letti.

Gli autori sono in totale 363, su 467 libri complessivi. A “spacchettare” gli autori multipli (come capitano nelle antologie, raccolte, o anche co-autori in saggi) si arriva a 569.

L’autore più letto in questi ultimi dieci anni è Roberto Calasso (che di Adelphi è presidente), con sette libri. A seguire sei libri di David Foster Wallace poi, a quota cinque, Guido Ceronetti (con le sue traduzioni bibliche), Oliver Sacks, Roberto Bolaño e George R. R. Martin. In termini di numero di pagine (migliaia) vince chiaramente George R. R. Martin, con le *Cronache del ghiaccio e del fuoco* (il nome originale della saga *Trono di spade*), che coprono gran parte delle letture nel 2012.

Genere degli autori

Figura 8. Uomini in ciano, donne in rosso.

 doi:10.4399/97888255181535

In termini di diversità di genere, il dato è abbastanza sconfortante: solo 21 autrici donne su 363 autori totali, poco più del 5%. Si arriva ad un perfetto 10% se vengono incluse anche le donne all'interno degli autori multipli, ma la sproporzione rimane evidente.

Quella delle autrici femminili è una questione importante, che sta giustamente ricevendo un'attenzione maggiore negli ultimi anni: si può riassumere nel detto "*gli scrittori uomini scrivono per tutti; le scrittrici scrivono per le sole donne*". La questione femminile è presente ad ogni livello della nostra società, ed il mondo editoriale non fa eccezione. In altra sede è stata compiuta un'analisi parziale della distribuzione fra autori uomini e autrici donne nei cataloghi editoriali di alcuni importanti editori, e si rimanda a lì per un approfondimento. Come in altri casi, questa analisi non può che porre nuove domande, a cui si potrebbe rispondere solo con un'analisi di *benchmark* sul lettore medio italiano. A livello di singolo lettore, infatti, le informazioni non sono abbastanza: è il lettore che inconsciamente predilige autori uomini ad autrici donne? O è la produzione libraria ad essere sproporzionatamente maschile? I temi che interessano il lettore sono dominio incontrastato di uomini? Nel nostro caso particolare, quasi sicuramente, è un insieme di tutto ciò.

Nazionalità

Anche riguardo al discorso della nazionalità l'omogeneità è evidente: gli autori sono quasi esclusivamente italiani o americani, e molto più in basso nella classifica troviamo nazioni europee come Francia, Spagna, Germania. Il resto del mondo è quasi solo rumore statistico.

È importante notare come questo grafico sia incompleto e solo indicativo, poiché per molti autori non si è riusciti a ritrovare la nazionalità tramite Wikidata.

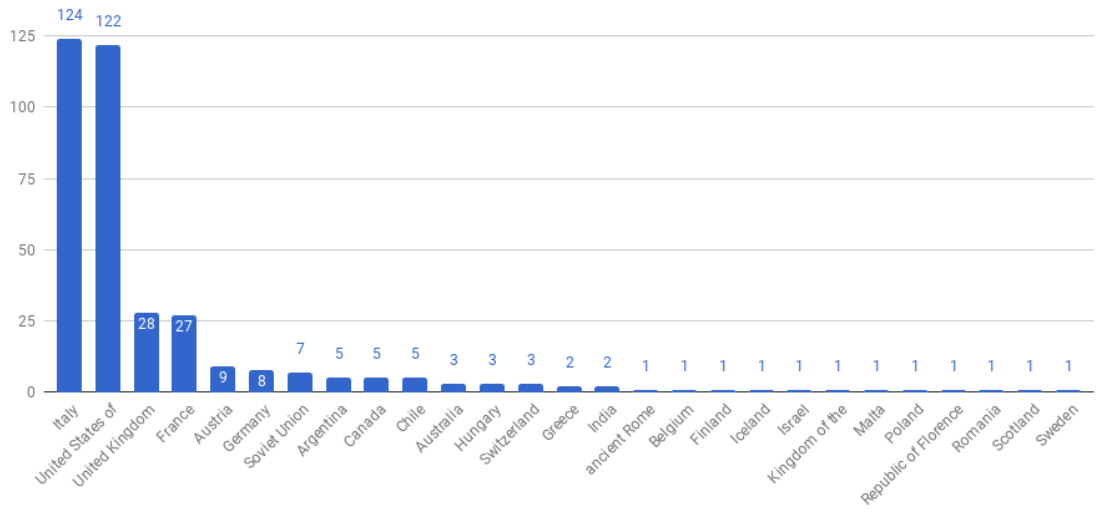

Figura 9. Paesi di cittadinanza degli autori letti.

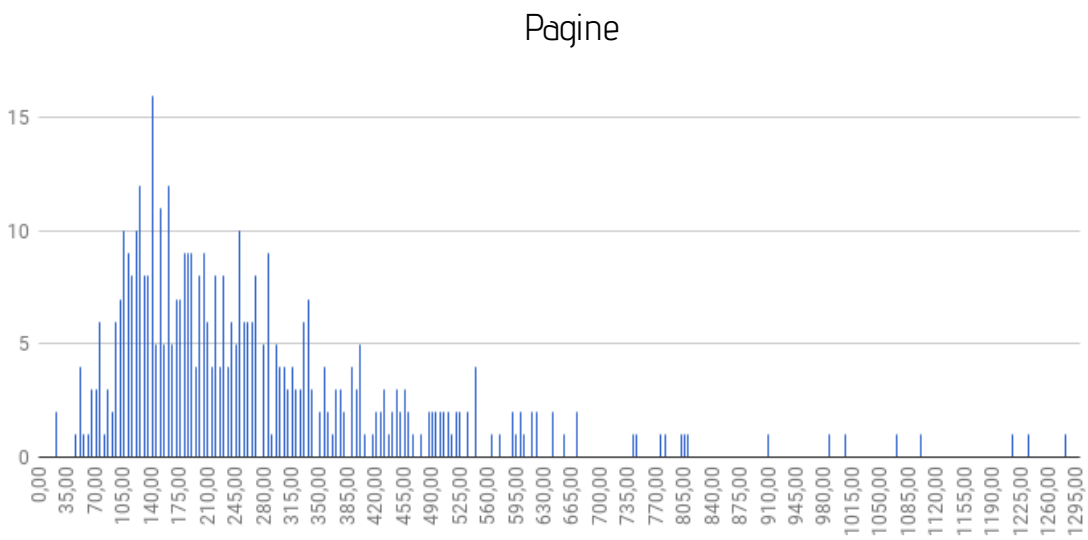

Figura 10. Distribuzione dei libri per numero di pagine – Dataset di partenza.

Con i numeri di pagine si può fare un'analisi più quantitativa: su 467 libri letti, la media di pagine è 271.38, mentre la mediana (cioè il valore che divide la distribuzione a metà) è 221 e la deviazione standard 187.25. Di fatto quasi una classica curva a campana, ma molto più spostata verso destra: una piccola ma non insignificante porzione di libroni sopra le 500 pagine. All'estrema destra, pochissimi libri davvero grossi, sopra le mille pagine, come *Infinite Jest* di David Foster Wallace e *It* di Stephen King.

In questo caso, possiamo fare una comparazione con alcuni dati presi dagli open data delle biblioteche romane citati in introduzione.

La distribuzione sembra simile:

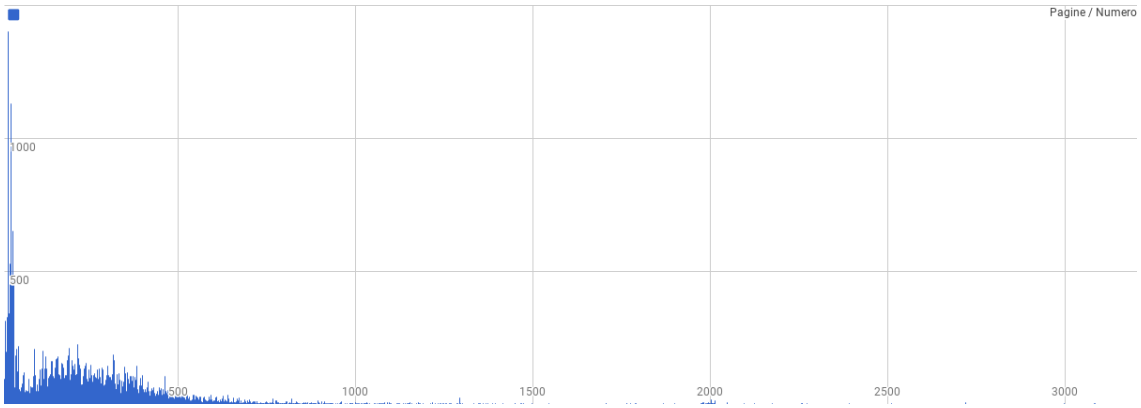

Figura 11. Distribuzione dei libri per numero di pagine – Biblioteche romane.

Il picco di libri molto corti sulla sinistra (si parla di 20–30 pagine) sono libri per bambini, che, eliminati, ci darebbero un grafico molto simile ad una distribuzione normale, e paragonabile con la nostra.

Infine, la distribuzione temporale, cioè l'analisi dei libri letti durante il nostro periodo, ci mostra una certa omogeneità nella lettura dei libri in base alla loro "dimensione". Quasi ogni anno ci sono libri sopra le 600 pagine, e sicuramente sopra le 500, ma il grosso dei libri sta tra le 100 e 250 pagine, una dimensione molto più canonica.

 doi:10.4399/97888255181535

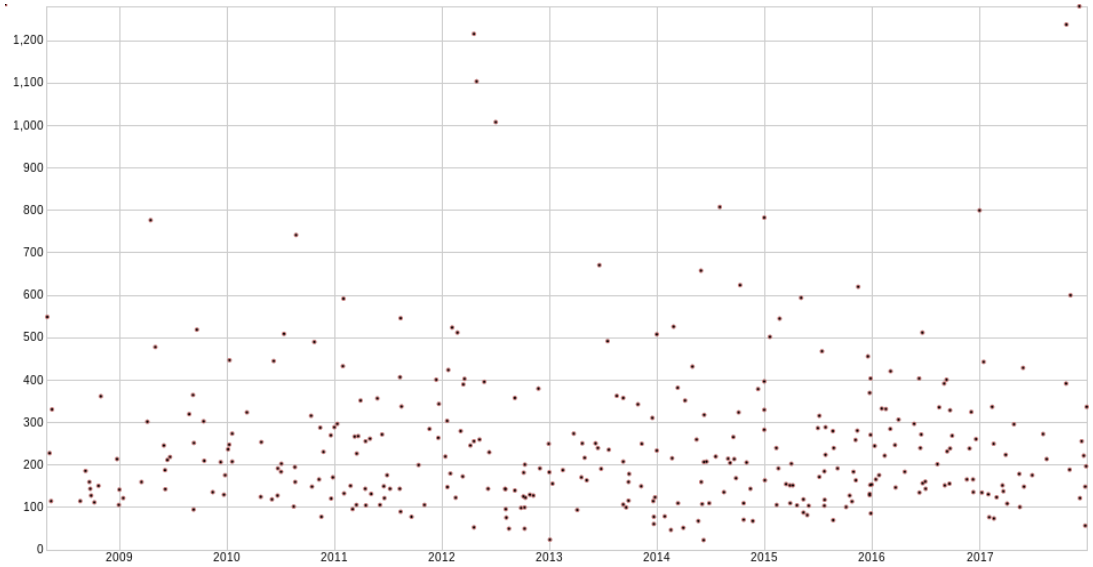

Figura 12. Distribuzione dei libri per numero di pagine – Biblioteche romane.

Metodologia

Le azioni compiute sul dataset si possono essere raggruppate in cinque fasi:

1. *Tracciamento ed esportazione dati*

   La piattaforma utilizzata per tracciare i libri letti (o abbandonati) è stata aNobii, piattaforma di social reading che permette, fra le altre cose, la costruzione di biblioteche personali. Anobii registra i metadati descrittivi del libro (autore, titolo, editore, collana, anno di pubblicazione, numero di pagine) assieme ad alcuni dati aggiuntivi e personali come tag, data di inizio e fine lettura. Non tutti i libri godono della stessa completezza o ricchezza nella descrizione bibliografica, dato che le schede sono create collettivamente dalla comunità di *anobiani* volontari. Anobii permette l'esportazione in CSV o Excel della propria biblioteca personale, che è stato dunque il dataset grezzo da cui si è partiti.

2. *Pulizia dati*

   I dati non possono essere analizzati senza essere "puliti", e per questa fondamentale dase è stato usato il software open source OpenRefine. Con "pulizia" si intende il processo, spesso molto *artigianale*, di standardizzazione e omogeneizzazione dei dati. Spesso infatti uno stesso autore o editore può avere nomi o grafie diverse (emblematico è il caso di Mondadori, che può essere espressa in una dozzina di forme diverse). Queste varie forme vanno ricondotte ad uno standard. Dato che non tutti i record erano completi, quando possibile sono stati completati manualmente attingendo i dati da cataloghi librari o bibliografici. Altri dati (fra cui la divisione in fiction/non-fiction) sono stati inseriti a mano.

3. *Riconciliazione dati*

   La *reconciliation* è la pratica di arricchire un dataset prendendo i dati da un altro database esterno. In questo studio, tutti gli autori sono stati confrontati dentro la base di dati libera Wikidata, importando nel dataset le relative nazionalità, date di nascita e di morte, sesso. Il processo viene svolto in maniera integrata e automatica da OpenRefine attraverso le API di Wikidata. Come sempre accade, anche in questo caso i dati non erano sempre disponibili, completi o puliti, per cui è stata necessaria un'altra fase di pulizia degli stessi.

4. *Esplorazione e analisi*

   La fase di pulizia è in un certo senso, anche una fase di esplorazione e conoscenza del dataset. Uno degli strumenti più usati in OpenRefine è stato "Text Facet" (analogo alla

funzione *Pivot tables* di Excel): tale funzione permette di "contare" le occorrenze in una determinata colonne, ad esempio il numero di autori e editori che si ripetono. Tali liste sono poi state spesso ordinate in maniera decrescente, permettendo di vedere le numerose distribuzioni paretiane osservate. Google Fogli è stato utilizzato per alcune formule statistiche, quali media e mediana.

5. *Visualizzazione dati*

   I grafici più semplici (es. istogrammi, torte) sono stati creati con Google Fogli, mentre le visualizzazioni più complesse (es. *bubble chart, gantt chart, scatterplot*) con l'app online RAW.

Le cinque fasi descritte non vanno intese come strettamente indipendenti o a "compartimenti stagni": la pulizia del dataset ad esempio viene fatta durante tutto il ciclo, dato che dopo l'arricchimento o la riconciliazione automatica è necessario controllare i nuovi dati aggiunti. Allo stesso modo, uno dei modi migliori per esplorare i dati è la visualizzazione degli stessi in grafici, che permettono facilmente di riconoscere pattern invisibili all'occhio umano nei meri numeri.

Conclusioni

Come anticipato, quest'analisi non può avere alcun valore statistico, e non si possono escludere errori nella pulizia del dataset e nell'arricchimento dei dati con fonti esterne quali Wikidata. Siamo però convinti che l'analisi abbia un importante valore *esplorativo*, come progetto pilota per indagare alcune distribuzioni e statistiche nell'analisi di elenchi di libri (cataloghi, bibliografie). In questo caso, si è scelto di analizzare in profondità la biblioteca di un singolo lettore. Quando si parla di statistiche di lettura, solitamente si fa sempre riferimento a numeri che ci dicono di una popolazione che legge sempre meno… senza per altro sapere *cosa* legge, *quando, come,* e *perchè*.

Si ama parlare, e giustamente, di bibliodiversità, ad indicare quell'ecosistema complesso e diversissimo di diverse case editrici, diversi autori e diversi libri. Vorremmo aggiungere a questo ecosistema anche la "specie" del *lettore*: sempre diverso nei libri che legge da altri lettori, diverso anche da sé stesso in base alle circostanze, alla stagione, al supporto di lettura, alla lingua, all'esigenza di leggere per lavoro, per scuola o per semplice divertimento.

Reiteriamo l'esigenza che altri istituti, come ISTAT, CEPELL, AIE, AIB - con più risorse e ben più preparati metodologicamente - siano in grado di porre queste domande per esempio ai dati transazionali delle biblioteche o delle librerie, per poter costruire diversi benchmark riguardo il comportamento dei lettori italiani.

 doi:10.4399/97888255181535

Bibliografia

Berners-Lee, Tim. “The Semantic Web”. *Scientific American* 284 (2001): 34-43.

Berners-Lee, Tim. “Raw data, now!”. *Wired* (2012). Consultato il 5 settembre 2018. Disponibile all’URL: https://www.wired.co.uk/article/raw-data.

Barabási, Albert-László. *Lampi*. Torino: Einaudi, 2011.

Barabási, Albert-László. *Link*. Torino: Einaudi, 2004.

Barabási, Albert-László. *Network Science*. Cambridge: Cambridge University Press, 2016.

Bostok, Mike. *Visualizing algorithms*, 2014. Consultato il 5 settembre 2018. Disponibile all’URL: https://bost.ocks.org/mike/algorithms/

Faggiolani, Chiara e Maurizio Vivarelli (a cura di). *Le reti della lettura*. Milano: Editrice bibliografica, 2016.

Faggiolani, Chiara, Lorenzo Verna e Maurizio Vivarelli. “Text mining e network science per analizzare la complessità della lettura. Principi, metodi, esperienze di applicazione.” *JLIS.it* 8.3 (2017): 115-136. doi 10.4403/jlis.it-12414

Goldin, Marco. *Open data, libri e biblioteche*, 2018. Consultato il 5 settembre 2018. Disponibile all’URL: https://medium.com/@inmediaref/open-data-libri-e-biblioteche-con-un-pizzico-di-data-science-ccba26a6b385

Greco, Albert. *The Book Publishing Industry*. New York: Taylor & Francis, 2015.

Pometti, Maria e Francesco Tissoni. *Comunicare con i dati. L'informazione tra data journalism e data visualization*. Milano: Ledizioni, 2018.

Silver, Nate. *The Signal and the Noise*. London: Penguin Books, 2012.

Tauberg, Michael. *Power law in Popular Media,* 2018. Consultato il 5 settembre 2018. Disponibile all’URL: https://medium.com/@michaeltauberg/power-law-in-popular-media-7d7efef3fb7c

Zanni, Andrea. *I libri che ho letto negli ultimi dieci anni*, 2018. Consultato il 5 settembre 2018. Disponibile all’URL: https://medium.com/@aubreymcfato/i-libri-che-ho-letto-negli-ultimi-10-anni-2008-2017-fdafca622e3