STRUMENTI DI RICERCA E METADATI NELLE BIBLIOTECHE DIGITALI: MODELLO 'GUTENBERG' E MODELLO 'GOOGLE BOOKS'

INDICE

1. INTRODUZIONE

Questo lavoro intende analizzare i pro e i contro di alcuni modelli di ricerca implementati per le Biblioteche Digitali e messi a disposizione dell'utente, muovendosi attraverso casi rappresentativi, da quelli che fanno maggior uso di metadati a quelli che tentano di farne quasi completamente a meno. Gli archivi e le biblioteche che raccolgono fonti digitali possono contenere una mole impressionante di informazione: i loro mezzi di indicizzazione e gli strumenti messi a disposizione dell'utente per la ricerca sono fondamentali. Se la biblioteconomia è una disciplina da sempre conosciuta e coltivata, la possibilità di creare un'app che operi ricerche 'full text' su documenti digitali ha messo in dubbio l'utilità di molti metadati. I metadati in qualche misura vengono usati dalla stragrande maggioranza delle BD – ad esempio, per segnare la data di pubblicazione – ma la questione aperta è quanti metadati servano per una fruizione ottimale degli archivi: se posso cercare 'full text' la parola "Intrattenimento" in una raccolta di saggi ed organizzare il risultato con i vari algoritmi di information retrieval, è ancora necessario che i servizi di una BD si avvalgano del tag "Intrattenimento" per testi che parlano prevalentemente di questo tema?

Questo lavoro non intende essere una prospettiva esaustiva degli archivi e degli strumenti finalizzati al recupero di documenti più o meno rari, né si tratta di un'analisi manualistica sulla situazione attuale delle BD e sulle vie che potrebbero imboccare: quello che ho tentato è un semplice confronto tra due diversi tipi di servizio attualmente utilizzati da BD online per rendere brevi ed efficaci le indagini degli utenti. Cercare testi in una raccolta digitale è un'attività che presenta molte somiglianze con una 'semplice' ricerca web, ma anche alcune grandi differenze: il materiale raccolto di solito non è altrettanto eterogeneo – anche le raccolte più "generaliste" di libri, ad esempio, si basano su unità culturali ben definite, quelle appunto dei libri messi a disposizione da privati e biblioteche – e l'utenza può essere più specificamente orientata.

Questa è la ragione per cui, a mio parere, una discussione sull'indicizzazione e le interfacce di ricerca utilizzate in una BD non equivale a una discussione sull'indicizzazione del web: sistemi adottati da biblioteche digitali specializzate, come il *Perseus Project*[1] o l'ETCSL[2], potrebbero sembrare improponibili per l'intero Web ma altamente competitivi all'interno del dominio di competenza. Discipline umanistiche con solide tradizioni filologiche possono richiedere un uso massiccio di metadati digitali e di categorie costruite a mano il cui superamento appare ancora lontano per la Linguistica Computazionale. L'ETCSL, una delle migliori biblioteche digitali dedicate alla letteratura sumera, si divide in aree fortemente 'strutturate' come "Composizioni mitologiche", "Composizioni storiche" e così via[3]: questi aspetti, come si vede facilmente, non possono ancora essere ricostruiti da un'analisi automatica dei testi eppure richiedono una mole di lavoro umano in alcuni casi proibitiva per progettazione, implementazione e aggiornamento. Condurre operazioni

1 http://www.perseus.tufts.edu/hopper/
2 http://etcsl.orinst.ox.ac.uk/
3 http://etcsl.orinst.ox.ac.uk/edition2/etcslbycat.php

simili per l'intero Web sarebbe semplicemente improponibile, date le dimensioni coinvolte. Questa è una delle ragioni per cui si sono cercati strumenti informatici automatici per creare categorie, tramite principi di riconoscimento condivisi, in cui inserire i documenti[4]. Un aspetto complementare è quello della visualizzazione dei risultati delle ricerche da parte di una utenza più o meno specialista. Non mi pare vi sia per ora un trend generale o accettato dalla maggioranza, benché questa potesse essere una speranza dei primi bibliotecari digitali. Di conseguenza potrò limitarmi ad esporre in modo più organico possibile i vari siti di volta in volta rappresentativi di un determinato approccio, evitando nude questioni di tecnica per concentrarmi sulla fruibilità del sistema di ricerca – più che di classificazione – da parte dell'utente. Molte di queste biblioteche hanno un sistema di ricerca solido ed esistono da un tempo relativamente lungo sul web; alcune sono diventate quasi 'istituzionali' e hanno imposto il loro *management system* ad altri progetti, ma non esiste una 'regina' delle BD. Il mondo archivistico digitale insomma è frammentario e se gli aspetti più avveniristici riguardano soprattutto l'informatica e la linguistica computazionale, la gestione delle categorie e la creazione – che essa avvenga automaticamente o manualmente – di nuovi criteri per suddividere i testi o i vari documenti ha ancora moltissime possibilità di sviluppo. Si tratta in genere di una situazione *in fieri*, dove le aree coinvolte sono in rapida evoluzione.
Anche i destinatari della biblioteca digitale naturalmente influenzano la scelta del sistema di ricerca: un filologo classico sarà capace di utilizzare sistemi di filtro più complessi di un neofita delle lettere classiche e vorrà rappresentare la conoscenza attraverso servizi più sensibili – con una serie di richieste spesso troppo complessa per un algoritmo automatico. Il problema della ricerca per genere, ad esempio, costringe ancora l'informatica a servirsi di suddivisioni manuali per risolverlo: non esistono algoritmi buoni a sufficienza da proporre una suddivisione automatica per genere su vasta scala. Gli anni forse mostreranno una crescente automazione anche in questo settore. Per ora, quando le biblioteche scelgano di agire attraverso categorie di tipo manuale devono preoccuparsi della loro comunicazione all'utente e assicurarsi che sia chiaro come possa muoversi su queste categorie. In caso contrario, la difficoltà sarà nell'ottenere risultati 'degni di bibliotecari' almeno fino ad un certo livello e a non inciampare troppo spesso nelle trappole della semantica.

2. PREMESSA: LA RICERCA 'MANUALE' E IL MODELLO 'WIKIPEDIA'

Cercare un testo raro nelle vastità del Web può essere un problema non indifferente. Testi bizzarri, rari, preziosi, semi-dimenticati e così via non sono necessariamente offerti al primo colpo dai motori di ricerca. Ancora più problematico, forse, può essere cercare un *tipo* di testo particolare, ad esempio un genere, o un argomento, o qualche altra categoria generale, senza avere già un titolo in mano. Ad esempio, quanto tempo impieghereste a trovare tramite Google due poesie sul Boeing 247?
Almeno una volta nella vita abbiamo dovuto affrontare questo genere di problema, nonostante il grande progresso dei motori di ricerca.
Due sono le strade che possiamo teoricamente percorrere per trovare i nostri aghi nel

4 Kessler et al., 'Automatic Detection of Text Genre', http://arxiv.org/pdf/cmp-lg/9707002.pdf

pagliaio: utilizzare archivi ricchi e ben indicizzati oppure rivolgerci a sistemi e motori di ricerca tali da trovare quello che vogliamo in materiale più o meno indiscriminato.

In genere, la maggioranza tende a combinare questi due modelli per raffinare la propria ricerca durante l'*iter*, muovendosi con un sistema che viene a volte definito di 'autorità e vertici' ('autorities and hubs')[5]. Un 'hub' in un grafo diretto è un nodo che ha un alto numero di collegamenti in uscita; un'autorità in questo modello corrisponde a un nodo con molti collegamenti in entrata. Si tratta di cercare l'esistenza di titoli che rispondano a particolari caratteristiche in un hub accreditato da molte autorità e i titoli stessi in un archivio "autorità", ovvero indicato da molti hubs. Ad esempio, potremmo cercare attraverso Google, un hub che non necessita introduzione, il bigramma 'digital library'. Supponiamo che la prima biblioteca indicata da Google sia anche la prima indicata da molti altri motori di ricerca presenti sul web: questo darà alla biblioteca indicata da Google un'autorità maggiore in materia, ossia saremo più propensi a credere che sia effettivamente una biblioteca importante; se d'altro canto un motore di ricerca restituisce molto spesso in prima posizione siti 'di autorità', saremo propensi a ritenerlo un buon motore di ricerca, vale a dire un buon hub. Questa è una forma di ricerca che molti di noi applicano per orientarsi tra i depositi di informazione: man mano che ci abituiamo a fare ricerche, aumentiamo la nostra fiducia in alcune autorità insieme ad alcuni 'hubs' che ce le hanno indicate. È interessante che questo meccanismo sia limpidamente formalizzato nel così detto algoritmo 'PageRank' di Google, che ha contribuito all'originale soddisfazione degli utenti per il motore di ricerca: PageRank tende proprio ad eleggere hubs e autorità principali e col passare del tempo a renderli sempre più 'unici'[6].
Questo genere di doppia ricerca per così dire 'manuale' è assolutamente naturale nell'utilizzo del Web ed è importante tenere in considerazione, quando valutiamo gli strumenti di *information retrieval* messi a disposizione da una BD online, che essi verranno tendenzialmente utilizzati all'interno di meccanismi simili dalla maggioranza delle persone.
Supponiamo che stia ancora cercando le poesie sui Boeing 247. Wikipedia offre un'intera pagina "Aircraft in fiction" in cui trovo anche il Boeing 247: non è in una poesia, bensì in una serie televisiva ('The Great Air Race'). Posso decidere che anche la serie mi interessa, definendo meglio la mia stessa indagine, e tentare di trovarla su altri hubs come le grandi raccolte di scripts online, in cui una richiesta diretta di "scripts di fiction sul Boeing 247" sarebbe stata difficilmente formulabile; oppure posso continuare la ricerca.
Wikipedia rimane uno dei migliori *hubs* per condurre indagini 'manuali' di raffinamento.
I vantaggi di Wikipedia e di simili enciclopedie sono da questo punto di vista notevoli:

1) Posso cercare un argomento 'generico' per chiarirmi le idee.
2) Posso trovare collegamenti a materiale scientifico o più specifico.
3) Poiché le pagine sono spesso organizzate con principio enciclopedico anche le

5 http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html
6 Sullivan, Danny. "What Is Google PageRank? A Guide For Searchers & Webmasters", *Search Engine Land*.

fonti sono ordinate secondo criteri cronologici o argomentativi facili da seguire, almeno in teoria, poiché relati al discorso portato avanti nell'articolo.

Per quanto riguarda argomenti di nicchia, la presenza in rete di siti altamente specifici può far pensare che la questione della ricerca di documenti si risolva in un problema di individuazione di autorità specializzate, dove le ricerche saranno agevolate dalla particolarità del tema. In sostanza potremmo immaginare un modello composto da piccole BD peculiari e da una grande 'gestione di collegamenti' che indirizzi l'utente al ridotto deposito di conoscenza che più lo può interessare, da poesie in linguaggio dei segni a tracce audio di balene, depositi che potranno anche essere organizzati con complesse categorie manuali, dal momento che conterranno una mole molto ridotta di documenti. Queste prospettive non considerano la normale caoticità del web, in cui tutto, dalle azioni ai film, si mischia senza un vero controllo; a creare solo piccoli siti organizzati come dizionari di dominio, diretti ai ricercatori del campo, si rischia di rimanere sommersi.

Utilizzare Wikipedia o siti simili come hubs per trovare autorità nella materia che ci interessa è per certi versi la ricerca manuale 'per metadati' nel senso più puro. Per trovare ad esempio articoli pionieristici sul motore del jet, cercheremo il tema *jet engine* e frugheremo tra le citazioni e le fonti del paragrafo 'Pioneer Studies'. Wikipedia è ottima per chiarire le idee e indirizzare l'utente a siti di argomento più specifico, ma *non* è un motore di ricerca: l'ideale per un motore di ricerca su documenti sarebbe poter arrivare ad una sorta di enciclopedia questionabile con queries artificiali, ossia ad una gestione automatica intelligente della conoscenza.
Wikipedia ha anche sviluppato un progetto-sorella, Wikisource[7], che è una vera e propria biblioteca digitale e le cui categorie includono indici alfabetici, cronologici e tematici, macro argomenti – es. *Studi Archeologici* – divisi in argomenti più particolari – es. A*stronomia:Satelliti Naturali* -, liste di autori, etc.. L'aspetto qualificante di Wikisource come di Wikipedia, insomma, è ancora la manualità nella gestione dei testi digitali e la creazione manuale[8], che non accenna a diminuire, di voci, categorie, elenchi, etc..
La ricerca enciclopedica può fornire panoramiche ragionate estremamente esaurienti

7 http://it.wikisource.org/wiki/Pagina_principale
8 Tranne casi di voci molto tecniche, che possono essere create a enorme velocità tramite generatori automatici di testo, come nel recente caso-scandalo di Sverker Johansson: http://online.wsj.com/articles/for-this-author-10-000-wikipedia-articles-is-a-good-days-work-1405305001

su letterature tecniche piuttosto specifiche (i testi più importanti sui motori a pistone, la letteratura critica più autorevole sul metodo socratico, etc.) e rimane il mezzo migliore per capire se stiamo cercando ciò che ci interessa.
Come trovare altrimenti il nome esatto di un'idea che abbiamo in mente o scivolare a nuove ricerche che ci intrigano di più? Naturalmente lo scopo della BD non può essere quello di chiarire le idee a un ricercatore confuso: eppure l'indicizzazione artificiale di documenti risente ancora di una mancanza di elasticità nella comprensione delle richieste degli utenti, di una necessaria attinenza alla presenza di un determinato n-gramma, o al limite di un suo stretto sinonimo, all'interno di un documento. Non esiste ancora ad esempio un algoritmo di ricerca su vasta scala che riconosca la presenza di un tema in una perifrasi.
Le BD offrono tuttavia un vantaggio ovvio all'utente: le ricerche sono indipendenti dalla classificazione del creatore della biblioteca. In una biblioteca divisa per categorie o tags, i 'contenitori' offerti sono gli unici che possiamo utilizzare – se non vengono offerti tags che rispecchino ciò che stiamo cercando, dovremo trovare una strada diversa e avvicinarci a combinazioni di tags diversi, complicando di gran lunga le interrogazioni.
A seconda del ruolo che un utente può assumere, anche le sue esigenze di ricerca varieranno. Gli utenti possono essere fornitori di contenuti digitali – digitatori di documenti, operatori addetti ad un qualche dominio, ricercatori che lavorino su un tema e comunichino la propria conoscenza, etc. - o consumatori di conoscenza, categoria in cui si trovano il fruitore occasionale e i membri di comunità designate per cui la BD svolge il proprio lavoro, ossia individui che, secondo la definizione OAIS, “sono in grado di capire un certo tipo di informazione”. Per una conservazione a lungo termine di una BD, è fondamentale che il ruolo degli utenti in questione sia chiaro ai suoi gestori.
In conclusione la ricerca manuale che si muove attraverso BD diverse è il 'modello' che molti dovrebbero avere in mente per realizzare i sistemi e gli stessi algoritmi di ricerca della conoscenza. È il motivo per cui i servizi offerti da Google, che fa uso dell'algoritmo di relazione sopra descritto, funzionano bene, ma anche per cui biblioteche che si servano di categorie non completamente banali – ad esempio manoscritti sulla filosofia naturale o libri di attività manuale – rimangono tutt'ora competitive. Inizieremo da queste ultime: le biblioteche con molti metadati.

3. METADATI E CATEGORIE: MODELLO 'GUTENBERG'

Illustrazione 1: La suddivisione in categorie concentriche è una caratteristica tipica delle biblioteche 'strutturate'

Progetto Gutenberg

Ho definito questo modello pensando al **Progetto Gutenberg,** la più vecchia biblioteca digitale esistente, fondata nel 1971 quando eBooks ed Arpanet erano strumenti per specialisti[9]. Ad oggi è una vasta BD, contenente testi prevalentemente letterari, organizzata in un ampio sistema di categorie e sotto-categorie per la gestione dei libri. La collezione, oltre a dividere i testi per lingua e a permettere una flessibile ricerca 'full text' su tutto il materiale, è organizzata per argomenti e scaffali, ovvero sovra-argomenti; inoltre dove possibile gli argomenti sono divisi in sotto-argomenti. Esistono anche tags 'Classe' che specificano meglio l'identità del testo. I libri possono appartenere a tutti gli argomenti e le classi necessari a definirlo in modo soddisfacente e quando la classificazione sia effettuata secondo criteri appropriati questa organizzazione permette, attraverso un sistema di metadati, di visualizzare insiemi interessanti come 'Poesia per Bambini', 'Precursori della Fantascienza', etc.. E' interessante notare come l'uso di metadati manualmente aggiunti sia perseguito nonostante le dimensioni della biblioteca non siano piccole[10] e, soprattutto, non esista un tema portante. Il Progetto Gutenberg permette di utilizzare espressioni regolari per filtrare i risultati sulla base dei metadati: ad esempio, cercando 'n.1102' troviamo il libro numero 1102, mentre con 'cat. juvenile l.german' cerchiamo tutti i testi che rientrino nella categoria della letteratura giovanile e siano in lingua tedesca. Il Progetto permette anche di usare operatori sui metadati: 'lovecraft(l.ita|l.ing)' cercherà Lovecraft in italiano o inglese; 'horror story ! ghost' cercherà tutti i testi di tipo 'horror story' che non abbiano il tag 'ghost'.

Poetry Foundation

La **Poetry Foundation Library[11],** una costola dell'organizzzione 'Poetry Foundation', è una delle più grandi raccolte di poesia in lingua inglese disponibili online. A differenza del Progetto Gutenberg, si tratta dunque di una raccolta per genere: se cerco la parola chiave 'airplane' in questa BD so già di ottenere *poesie* sul tema degli aeroplani. Se cerco 'sea abyss', sono poesie con questo bigramma che ottengo. Il sito offre l'opzione di raffinare la ricerca per tipo (poemi, poeti etc.) e all'interno del tipo per argomento. Questo significa che anche dopo aver cercato qualcosa di relativamente specifico come 'sea abyss' tra le poesie, posso chiedere di trovare solo i risultati con il label 'Nature', rafforzando così la probabilità di individuare una poesia che parli di abissi marini non usanti solamente come metafora.
Gli argomenti contengono sotto-argomenti che è possibile specificare: ad es. all'interno di 'Nature' si potrebbe specificare la categoria 'Stars, Planets, Heavens'. Si può notare che se cerco *abyss* e specifico il pattern 'Poem'>'Nature'>'Sea' ottengo due risultati (*The Witness*[12] di Longfellow e *Ode to a Large Tuna in the Market* di Neruda) differenti dai due risultati che rispondono ad *abyss* nel pattern 'Poem'>'Nature'>'Stars, Planets, Heavens' (*Lithium Dreams* di Beeder e *Compound Hibernation* di Alexander). Questo è un ottimo esempio delle potenzialità di una

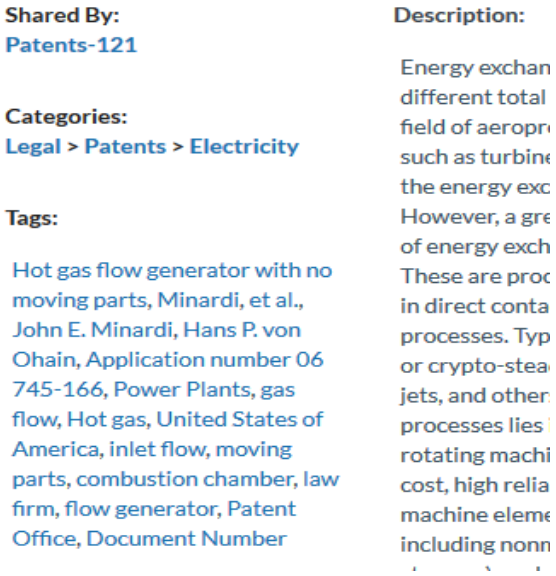
9 Hart, Michael S., "Gutenberg Mission Statement by Michael Hart". Project Gutenberg. 15 August 2007.
10 Circa 45.000 documenti, http://www.gutenberg.org/ebooks/search/
11 http://www.poetryfoundation.org/
12 http://www.poetryfoundation.org/poem/173919

estesa indicizzazione 'manuale' tramite metadati: per chiedere un simile livello di raffinatezza a un sistema che faccia poco uso di metainformazioni dovremmo inventare complicate combinazioni di parole chiave e i risultati rimarrebbero probabilmente molto rumorosi.

Docstoc e le raccolte di documenti tecnici

L'uso che alcune biblioteche fanno dei metadati rende chiaro come, in contesti meno 'liberi' di una risorsa generica, essi siano estremamente importanti. Ad esempio, nelle raccolte di documenti specifici come gli archivi di brevetti online, di articoli scientifici su un determinato tema eccetera, una ricerca che fornisca all'utente l'elenco dei tags collegati a un documento non pone solamente l'enfasi sul tipo di risorsa visualizzata ma potrebbe avere un ruolo nell'estensione dell'indagine stessa dell'utente, che può utilizzare l'informazione rappresentata tramite metadati per muoversi verso altri documenti interessanti.

Docstoc[13]**,** un sito che acquista e archivia documenti legali e finanziari di vario genere, ha un tipico approccio strutturato su questo principio: la ricerca viene operata in gran parte per parole chiave e ci si può muovere attraverso il sistema di numerazione dei brevetti per 'scivolare' in documenti simili a quello cercato.



Purtroppo l'accesso gratuito a Docstoc è limitato: gli utenti che vogliano scaricare i documenti cercati hanno a disposizione varie tariffe per navigare la biblioteca in diversi lassi di tempo, con o senza pubblicità.

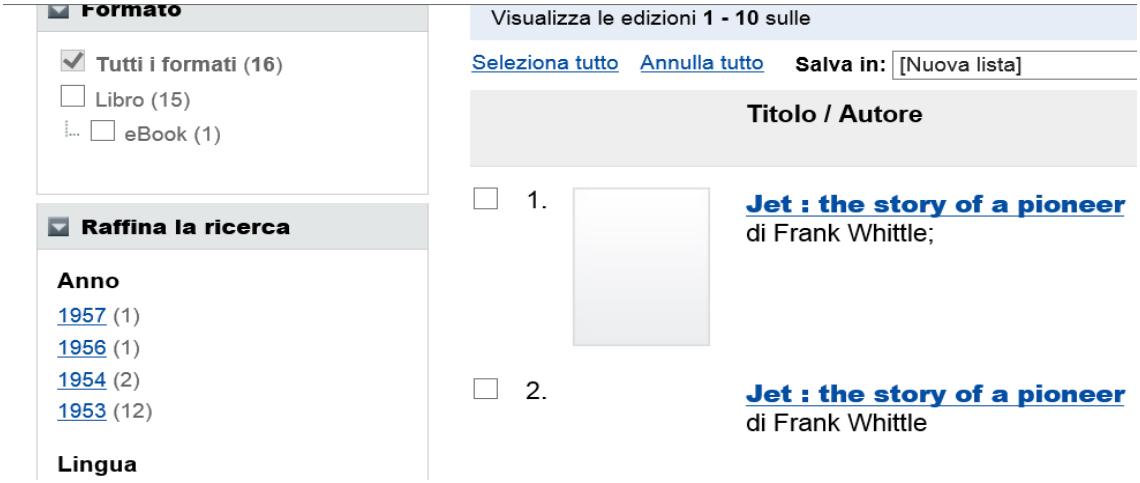
Una raccolta di brevetti anche organizzata tematicamente, seppure in modo un po' meno strutturato, è ***Freepatentsonline***[14]: i brevetti sono organizzati per argomento, benché non sia possibile poi muoversi per tags o per temi simili.

L'uso estensivo di tags e metadati per migliorare le indagini degli utenti è piuttosto diffuso in raccolte di articoli scientifici, documenti e così via. Non è altrettanto frequente, ed è forse naturale che non lo sia, la presenza di siti fortemente tematici o tematizzati in generi diversi da quello saggistico-divulgativo. In alcuni casi,

13 http://www.docstoc.com/docs/

14 http://www.freepatentsonline.com/

un'organizzazione tematica in testi narrativi o poetici potrebbe diventare difficoltosa. Si tratterebbe di documentare il contesto, il 'tema portante', elemento che può essere molto suscettibile all'interpretazione del creatore della BD (quale tag riassume il 'tema portante' del *Don Chisciotte* di Cervantes?), e così via. Questo non significa che non esistano numerose BD letterarie o di contenuto non-tecnico che facciano uso di tags anche piuttosto specifici. Ad esempio, in una BD di fantascienza anni Cinquanta non sarebbe troppo sfuggente individuare ed assegnare il tag tematico 'satellite naturale'. Allo stesso tempo, un algoritmo di classificazione automatica applicato alla BD potrebbe avere difficoltà a reperire gli stessi testi: un racconto in cui tutte le azioni siano svolte, ad esempio, su un satellite immaginario che viene nominato quasi solo attraverso un nome d'invenzione, sono destinati a rimanere fuori dalla via di ricerca dell'algoritmo.
Una via di mezzo è rappresentata ad esempio dalle raccolte in cui è possibile 'raffinare la ricerca' per anno, tipo di documento (libro, ebook) e lingua.
Un buon esempio potrebbe essere WorldCat, anche se a rigore non è una vera e propria BD: si tratta di un catalogo internazionale unificato di oltre 72.000 biblioteche, iniziato nel 1971 ed ora online, contenente 300 milioni di voci[15]. WorldCat presenta la possibilità di cercare i libri, oltre che per titolo ed autore, per anno, lingua[16] e poco più. Se le possibilità di ricerca di WorldCat possono apparire limitate, il catalogo offre connessioni ai siti internet delle singole biblioteche, in modo che l'utente possa controllare la disponibilità e le altre informazioni del libro direttamente 'in loco'. Il sistema tuttavia non è dei più comodi, dal momento che è impossibile filtrare una ricerca sulla base della disponibilità, stato di conservazione etc. del nostro libro e bisogna controllare a mano biblioteca per biblioteca.



Quest'uso 'ridotto' dei metadati è probabilmente uno dei più diffusi e rappresenta un livello di categorizzazione già in parte raggiunto dagli algoritmi di elaborazione del linguaggio naturale, capaci di individuare con relativo successo, all'interno di servizi informatici, la lingua prevalente di un documento. Una combinazione di questi metadati 'essenziali', come il luogo di pubblicazione, e di ricerche 'full text' fornisce già risultati che soddisfano una grande parte delle ricerche. Se esiste da qualche parte

15 "A global library resource". Online Computer Library Center. Gennaio 24, 2014
16 Le lingue rappresentate in WorldCat sono ben 450.

un testo incentrato sui gatti pubblicato nel 1861, è tutto sommato probabile che riusciremo già a trovarlo così.
Non è necessario pensare sempre l'uso di metadati come un sistema di ricerca pesantemente strutturato. Tuttavia, di solito, un metodo di classificazione e ricerca fortemente basato sulle metainformazioni richiede a sua volta svariate strutture di supporto alla categorizzazione della conoscenza: tassonomia di settore, vocabolari di linguaggio tecnico, thesauri di relazione semantica tra termini dello stesso linguaggio, authority files di gruppi che collaborano alla BD, gazetteers geografici, ontologie di dominio. Inoltre i metadati possono essere prodotti in logiche descrittive efficaci e apprezzate ma piuttosto complesse, ideate per gestire non tanto i documenti quanto le informazioni su di essi.

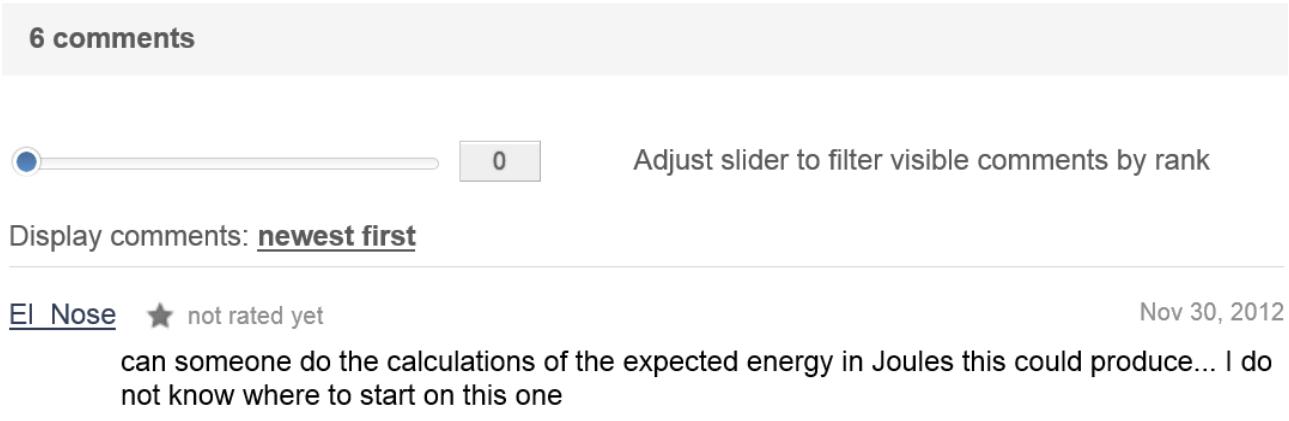


*Illustrazione 2: Non è necessario pensare l'uso di metadati come un sistema di ricerca pesantemente strutturato. I commenti di **PhysOrg,** una BD di articoli scientifici, si filtrano attraverso un solo metadato, il rank ricevuto. Il sistema è semplice e rapido.*

Toronto Library

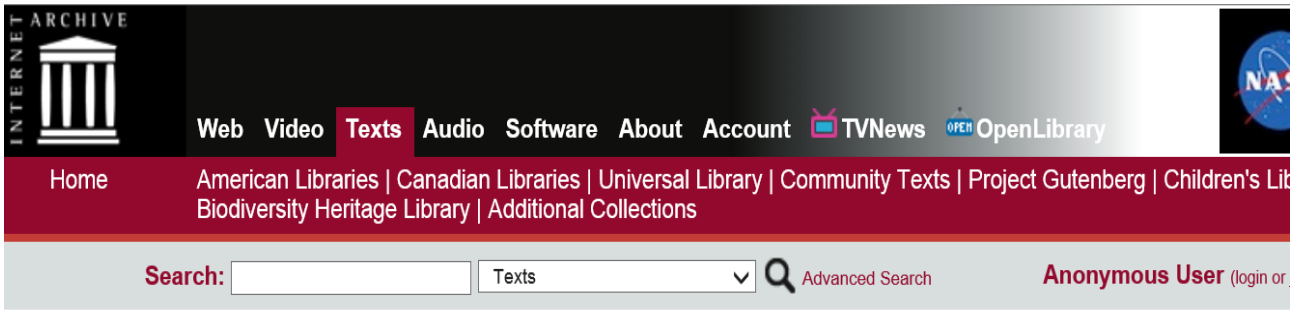
Moltissime biblioteche digitali ad oggi presentano interfacce che fanno uso di svariate e ingegnose categorie in cui dividere i testi: non solo elementi come 'Lingua Inglese' o 'Scritto da', ma il *periodo storico* nel quale il testo è stato prodotto - come 'Illuminismo' o 'Controriforma' – tipo d'opra e così via. Chi conosce le esigenze di alcuni tipi di ricerca storica o letteraria può immaginare quanto uno solo di questi principi più generali di classificazione possa essere benefico, quando queste categorie o le ontologie che sottostanno la loro formazione siano costruite con cura. In alcuni casi, la mancanza di flessibilità delle BD a scompartimenti fissi può essere perfino bilanciata dalla scoperta di una categoria che non avremmo neanche pensato potesse esistere[17].
È difficile ad esempio sottovalutare la potenzialità di un sistema di ricerca fortemente strutturato in taluni argomenti: vedere la differenza tra un archivio indicizzato solamente per autore come il così detto 'Canpoetry' della BD Toronto Library[18] sui

17 La creatività nella categorizzazione dei documenti è uno degli aspetti più preziosi – e ad ora meno "automatizzabili" - della gestione manuale delle BD. Alcune BD comprendono categorie come 'Banned Books Online' http://onlinebooks.library.upenn.edu/banned-books.html o 'Aboutness' http://arizona.openrepository.com/arizona/handle/10150/105066/browse?type=subject

18 http://www.library.utoronto.ca/canpoetry/pratt/

poeti canadesi, con l'interfaccia generale della stessa BD che permette invece di ricercare una poesia per rima, forma, critica e perfino movimento letterario di appartenenza[19], elementi questi ultimi ancora lontani dalle reti degli algoritmi che si muovono senza metadati.

Internet Archive



Una delle BD più interessanti del Web, per qualità e quantità dei documenti, è probabilmente il progetto Internet Archive[20].
Si tratta di un'ambiziosa BD fondata nel 2001 che raccoglie gratuitamente materiale di diverso genere – libri, documenti di ogni tipo, filmati, registrazioni etc. - da molte grandi collezioni, da utenti singoli e da crawlers automatici che esplorano il Web cercando di 'salvare' quanto più materiale possibile. La biblioteca comprende oltre 4.4 milioni di libri, un numero inferiore solo a GoogleBooks ma con la differenza che tutti i documenti sono interamente accessibili.
Internet Archive mantiene i propri documenti divisi per categorie e, soprattutto, non permette ricerca *full text* ma solo attraverso metadati. Questo, nonostante la straordinaria dimensione dell'archivio, non è sorprendente: l'Archive comprende uno dei maggiori progetti di digitalizzazione al mondo e buona parte del suo staff (oltre 200 persone) lavora allo scanning del materiale cartaceo. La gestione di testo digitale non facilmente processabile come il testo scannerizzato invalida tutti i sistemi di ricerca che non usino metadati: poiché il segno non è più 'visibile' alla macchina, tutto il significato deve essere raccolto in informazioni di contorno, parole chiave, categorie, etc.. Internet Archive deve una parte del proprio valore come BD alla gestione di categorie piuttosto raffinate.
Da questo punto di vista gli sviluppi dell'OCR, disciplina che non è ancora perfetta ma che ha visto una serie di grandi miglioramenti negli ultimi decenni, aiuta a muoversi verso BD dotate di testi sempre più processabili.

Europeana

Europeana[21] è un raro caso di biblioteca digitale pianificata "dall'alto": l'idea fu

19 http://rpo.library.utoronto.ca/
20 https://archive.org/
21 http://www.europeana.eu/

lanciata da Chirac nel 2005 e raccolta dai Presidenti di Germania, Spagna, Italia, Polonia e Ungheria. Europeana intende raccogliere quanto più possibile materiale culturale europeo – testi, immagini, files musicali, etc. - per creare una biblioteca digitale comunitaria. Europeana rappresenta un caso 'estremo' di uso di metadati: si tratta infatti di un autentico archivio di metadati, dal momento tutti gli elementi che raccoglie sono rappresentati solo attraverso un insieme di descrizioni che l'utente può ricercare. Qualora l'utente voglia vedere l'oggetto vero e proprio – ad esempio leggere il testo trovato – può accedere al sito dell'istituzione che ne ha fornito la descrizione. Di conseguenza, i metadati sono sottoposti a stretta regolamentazione: gli uploaders delle varie istituzioni sono tenuti a seguire lo standard Europeana Semantic Elements – ben presto Europeana Data Model - per informare le proprie descrizioni.

Europeana rappresenta un tipo di biblioteca digitale impegnato nella conservazione di materiale culturale ritenuto altamente qualitativo o storicamente significativo. In questi casi, in cui gli elementi in sé non sono 'di nicchia' e gli utenti non sono specialisti, il tema della cercabilità sconfina facilmente nel tema delle interfacce e della loro usabilità: un ottimo sistema di ricerca che manchi di una chiara introduzione anche grafica per l'utente è un sistema praticamente inutile; *escamotages* grafici per permettere una comprensione migliore dei risultati di una ricerca su documenti digitali possono fare un'enorme differenza. Vedremo più avanti alcuni strumenti del web che, data un'interfaccia semplice e brillante, riescono a favorire la comprensione e la gestione stessa delle conclusioni delle ricerche tanto da diventare in effetti strumenti di retrieval più validi di altri[22].

4. NLP E FULL-TEXT: MODELLO 'GOOGLE BOOKS'

Google e gli altri motori generici hanno dovuto tentare nella propria implementazione di fare il minor uso possibile di metadati manualmente aggiunti, dal momento che la "biblioteca" su cui andavano ad agire, ossia il Web, è troppo vasta e caotica per essere pre-indicizzata. Questi motori di ricerca hanno utilizzato complessi algoritmi di ricerca testuale basati sulla frequenza e la forza di correlazione delle parole cercate nei documenti, su riconoscitori di sinonimi, espressioni regolari e così via. Nonostante tutto, molto materiale nel Web rimane ignorato dai motori: è il così detto "deep Web", dove le tecnologie fin qui escogitate non arrivano.
Il sistema di ricerca di **Google**, la "search engine" più celebre mai costruita, si muove naturalmente attraverso una serie complessa di algoritmi di aiuto per ottenere, dal tentativo di ricerca di un utente, il risultato più soddisfacente[23].

Da una prospettiva di ricerca per BD, tuttavia, è importante sottolineare come Google

22 Alcuni semplici accorgimenti grafici per la visualizzazione delle categorie possono rendere molto più semplice da utilizzare un sistema di metadati, come dimostra la Wolrd Digital Library: http://www.wdl.org/en/

23 Con buone capacità di interpretazione delle categorie ricercate: se digitiamo 'jet disaster 1954' il primo risultato è *BOAC Flight 781*, ossia quello che sarebbe il più logico "occupante" di una simile categoria: https://www.google.it/search?q=jet+disaster+1954&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:it:official&client=firefox-a&channel=sb&gfe_rd=cr&ei=rzL7U4ScLYOI8QfrqoHoBA.

tenti di fare meno uso possibile di metadati e più uso possibile dei collegamenti inseriti negli ipertesti per associare siti diversi e calcolare la loro importanza in modo che il risultato appaia più naturale. Insomma se assimilassimo i motori di ricerca del web a quelli delle biblioteche digitali noteremmo come questi ultimi siano costretti a fare a meno di annotazioni manuali e di ontologie pre-ordinate; questo limite ha tuttavia incentivato l'invenzione di sistemi che possono tornare molto utili anche a una biblioteca, sia per ordinare materiale in rapida evoluzione – articoli in lavorazione, bozze etc. - sia soprattutto per sfuggire alle rigide costrizioni dei metadati e delle categorie.

Google Books

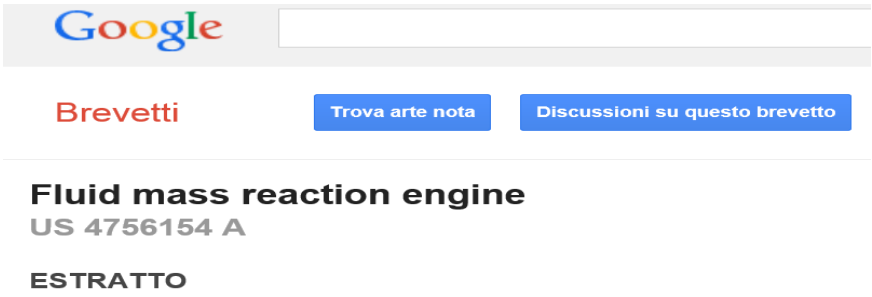
Google Books è una BD ideata da Google e una delle più fornite presenti sul Web. Il principale problema di Google Books è la presenza di moltissimi testi sottoposti a copyright e non visualizzabili o visualizzabili solo parzialmente. Date le dimensioni inoltre gli errori risultanti all'OCR non vengono corretti. In compenso, la mole del materiale contenuto supera di gran lunga i 5 milioni di documenti processabili.
Il meccanismo di ricerca base di Google Books non è dissimile per certi aspetti da quello di Google, ma deve rinunciare a un elemento fondamentale, il così detto 'PageRank' di cui nell'Introduzione[24]. Google Books permette all'utente di muoversi attraverso un numero minimo di categorie esplicite: in generale le parole richieste sono cercate nei documenti con algoritmi per determinare il valore del risultato; descrizioni ed eventuali parole chiave sono trattati come parti del documento stesso.
Questo sistema ha avuto enorme successo nel web ma può diventare scomodo quando, in una ricerca su BD quale è Google Books, abbiamo richieste particolari: ad esempio non è possibile nella ricerca base discriminare molto facilmente un anno di pubblicazione da un anno contenuto in un testo. Supponiamo che io voglia documenti riguardanti lo Heinkel He 178 pubblicati nel 1938: poiché il 1938 è l'anno della sua invenzione, questa data ricorrerà spesso in documenti centrati sull'Heinkel He 178, confondendo del tutto la ricerca[25].
Questo genere di problema ha indotto Google ad inserire strumenti di 'raffinamento' su metadati che permettono di filtrare per lingua, periodo, tipo di documento e Paese. Ma molte possibilità rimangono ancora fuori della portata di questo approccio: ad esempio, se cerchiamo un argomento solo in un genere letterario. Saggisti, poetici, documentari, narrativi, i risultati dei motori generici sono confusi come quelli di archivi non indicizzati. Da questo punto di vista, i “grandi archivi culturali” si offrono relativamente disordinati: Google Books restituirà alle nostre richieste documenti di ogni tipo. Se cerchiamo 'poetry' chiedendo di visualizzare solo testi otteniamo certamente risultati omogenei nella nostra vaga ricerca; e possiamo ordinarli per anno, Paese, etc.. Ma che succede se vogliamo un testo di finzione sugli abissi del mare? Ricerche come 'fiction abyss' non riportano solamente finzione e tentativi

24 Un meccanismo che invece si può applicare, tramite metadati, a BD di tipo diverso, come ad esempio le raccolte di articoli scientifici.

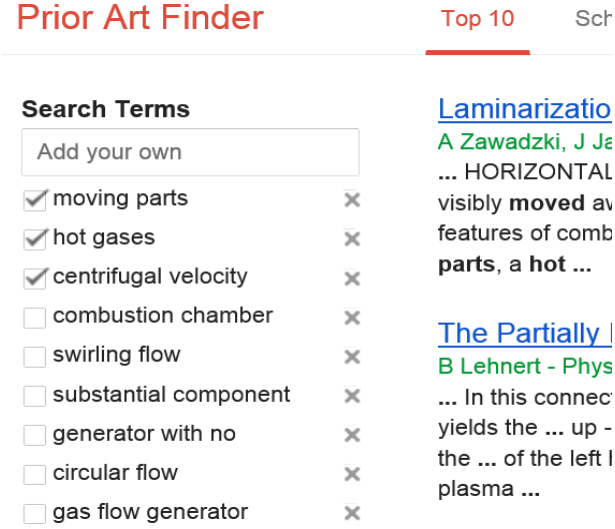
25 Qui il link ai risultati della ricerca: https://www.google.it/search?q=Heinkel+He+178+1938&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:it:official&client=firefox-a&channel=sb&gfe_rd=cr&ei=zS_7U6mpOIuI8QfGzoHoBQ#channel=sb&q=Heinkel+He+178+1938&rls=org.mozilla:it:official&tbm=bks.

come 'poetry fiction abyss' non danno risultati molto soddisfacenti: una grande distanza dalla precisione delle categorie ad incastro di Poetry Foundation!

Google Brevetti e 'Trova Arte Nota'



Dominando oltre il 71% del mercato, Google tenta con un cangiante gruppo di strumenti di ovviare alle limitazioni sopra descritte. **Google Prior Art Finder**[26] ad esempio è dotata della possibilità di creare una serie di 'tag di ricerca' personalizzati, che vengono configurati come tags tradizionali da combinare e condividere



Tuttavia il sistema è piuttosto rudimentale e non è molto diverso da una semplice ricerca 'full text' e tramite links, se non per la resa grafica. Da questo punto di vista i tentativi di superare gli ostacoli dell'indicizzazione automatica sono ancora piuttosto frustranti.

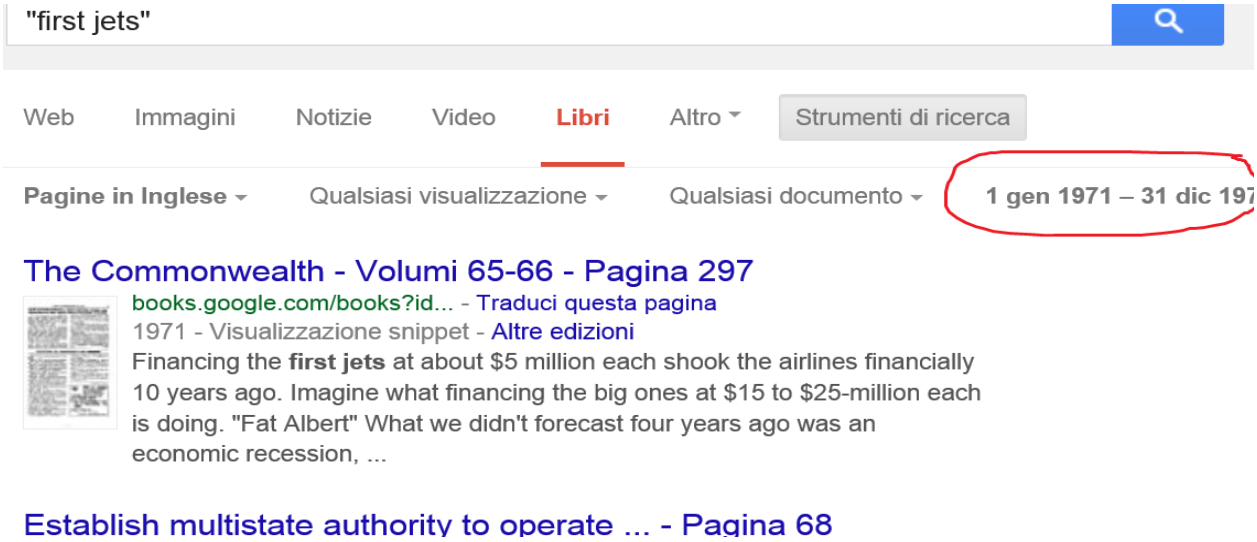
Ngram Viewer

Ngram Viewer[27] non è una BD a sé ma un sistema di ricerca e visualizzazione delle informazioni che si muove sulla BD Google Books. Ciò nonostante, esso merita una trattazione a parte per via della sua peculiare efficacia nel gestire le ricerche. Ngram Viewer permette di fare ricerche all'interno di una piccola serie di corpora, quasi tutti caratterizzati dalla lingua in cui sono scritti i testi, tranne la categoria 'English Fiction'. I risultati non vengono visualizzati normalmente ma formano un grafico

26 http://www.google.com/patents/
27 https://books.google.com/ngrams

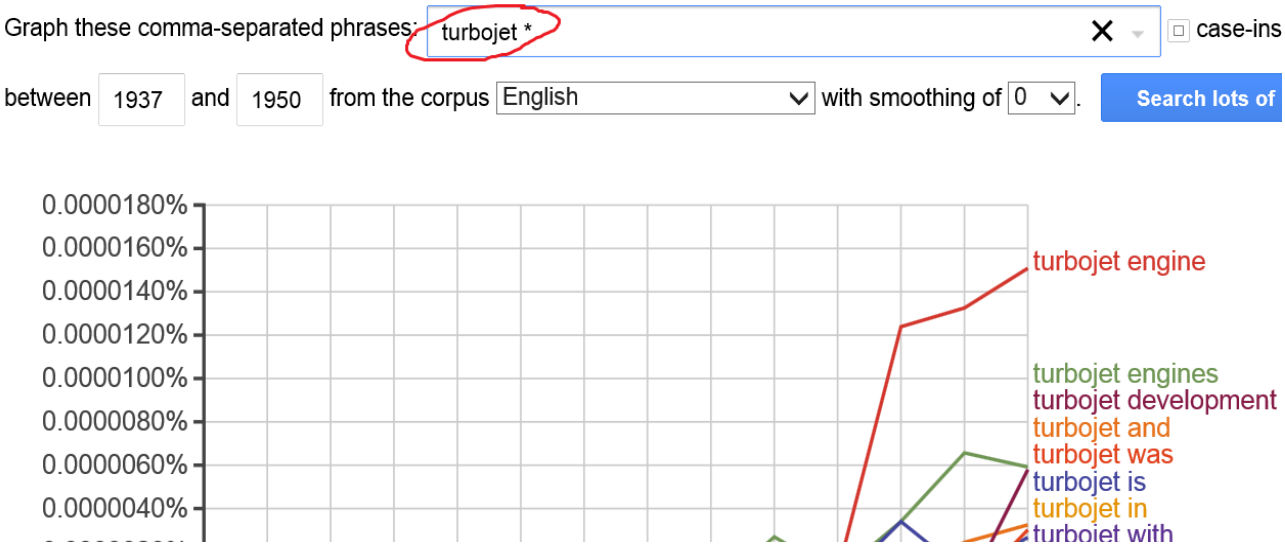
pesato che permette di vedere la frequenza degli 'n-grammi' ricercati all'interno di una finestra temporale decisa dall'utente.

L'interfaccia grafica e il semplice sistema di NgramViewer che fa uso di un solo metadato, quello della data di pubblicazione, permettono di estrarre una straordinaria quantità di informazioni sul tema che stiamo ricercando. Possiamo in un solo sguardo vedere la gloria e il declino di modi di dire, oggetti, tecnologie, con la possibilità di visualizzare i documenti che, per scaglioni di date, hanno giustificato il risultato grafico.



Vorrei soffermarmi su Ngram perché credo che sia un ottimo esempio di come strumenti di ricerca che dispongono di pochissimi metadati – data di pubblicazione, lingua e poco altro – possano configurare risultati anche più fluidi e interessanti di quelli ottenuti tramite l'uso intensivo di metadati. Da una parte, si tratta di una semplice ed efficace interfaccia: se l'idea del grafico cronologico non è nuova, la sua applicazione così diretta in servizi di grandi BD è piuttosto rara. Dall'altra parte la possibilità di visualizzare i *trends* di n-grammi in una finestra cronologica consente di comprendere molto sulla conoscenza che stiamo cercando. Non si tratta in sostanza solo di un modo migliore di visualizzare il risultato del lavoro dei servizi di ricerca. Il grafico di Ngrams consente di raffinare la nostra stessa ricerca: possiamo vedere

quando la frequenza d'uso delle parole si sia ribaltata, o stabilizzata, o quando sia diminuita fino a scomparire[28].
Il problema del mancato trattamento della conoscenza formalmente rappresentata viene inoltre contrastato tramite quello strumento semplice e potente che sono le espressioni regolari, questa volta applicate non ai metadati come nel Progetto Gutenberg, ma alla ricerca 'full text'. La possibilità di usare wildcards in questo modo apre svariati scenari applicativi: possiamo capire tra le nostre risorse quali combinazioni di parole supportino il maggior numero di risultati; quali lingue in quale periodo abbiano più testi con un certo nome proprio; etc..



L'uso delle wildcards ha reso poi Ngram Viewer una risorsa ancora più potente per la ricerca lessicografica, storica, letteraria, e così via. Il modello di Ngram Viewer sarebbe forse un migliore strumento di ricerca per la letteratura scientifica rispetto alle combinazioni di tags, ma è necessario saperlo usare con attenzione.
Poter scegliere la lingua ad esempio mi permette di reperire la stessa parola internazionale, come JumboJet, in letteratura inglese, francese, cinese, tedesca, e così via. Il sistema mi permette anche di capire il successo delle varianti grafiche, e quale di esse abbia avuto maggiore diffusione: *jumbo-jet*, *jumbojet, jumbo jet,* etc..
Cercare 'full text' con un numero minimo di metadati offre insomma una interessante gamma di ricerca, ma molte divisioni rimangono al di fuori della sua portata: non permette ad esempio di selezionare il genere del testo ricercato. Basti vedere il genere dei primi risultati della stessa ricerca, 'jumbo jet', su Ngram Viewer in Inglese (narrativa), Francese (memorialistica), Tedesco (grottesco) e Spagnolo (divulgativo).

28 La possibilità di confrontare i trends di diverse parole ci permette anche di comprendere subito il successo di una rispetto all'altra: https://books.google.com/ngrams/graph?content=jet%2Cturbojet&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Cjet%3B%2Cc0%3B.t1%3B%2Cturbojet%3B%2Cc0

Illustrazione 4: Google è in grado di riportare automaticamente una ricerca come 'nazi flettner fl propaganda' a quella che sarebbe una delle sue "categorie di appartenenza" più legittime, le Wunderwaffen, armi avveniristiche sbandierate dalla propaganda nazista di cui il Flettner Fl è un esempio.

Illustrazione 5: Naturalmente Google Books utilizza semplici stratagemmi come quello di dare maggior peso al titolo del documento; ma fa poco uso di metadati.

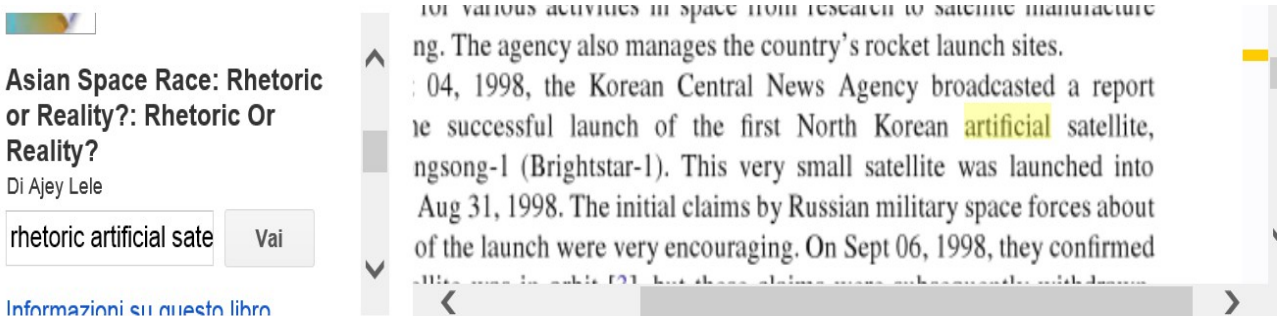


Illustrazione 6: Il motore di ricerca di Google Books si muove tramite ricerca full text.

Gli ordinamenti cronologici costruiti manualmente sono certamente più affidabili, ragionati e strutturati[29]: ma sono per necessità molto più piccoli e molto più rigidi.
Se Ngram non può raggiungere la precisione di siti individuali, le ricerche sull'uso linguistico diventano infinitamente più interessanti; le progressioni sono quantificate e possono essere messe in correlazione con altre ricerche. D'altro lato se cerco in Ngram *Leonardo da Vinci* so che il risultato mescolerà libri scritti da Leonardo, su Leonardo, libri che contengono frasi riprese da Leonardo e così via.

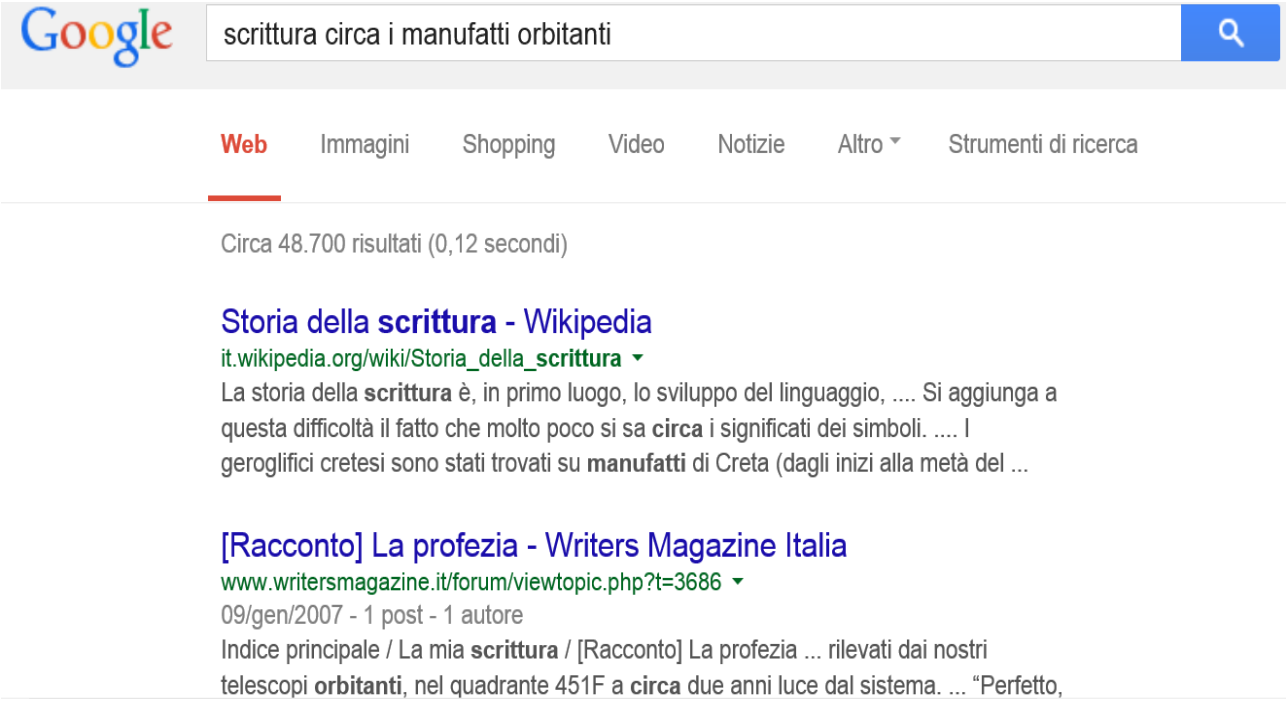


Illustrazione 7: Gli algoritmi di NLP non riescono ancora a comprendere tutto: ricerche espresse in modo bizzarro mandano tutt'ora i motori fuori strada

In conclusione, sistemi che non facciano uso di descrizioni permettono molta creatività nella ricerca dell'utente: se cerco testi circa la proposta, l'annuncio di realizzazione e il successo del telescopio spaziale Hubble in una BD attraverso il filtro di metadati devo sperare che per qualche ragione i gestori della biblioteca digitale abbiano scelto proprio questo elemento come descrizione rilevante. Con la 'ricerca semplice' di Ngram Viewer, pericolosamente libera da qualsiasi descrizione, posso a livello ideale cercare di intravedere questo peculiare fenomeno all'interno della mia BD[30]: posso supporre che nello scaglione con indice minore, 1989-1990, si trovino i testi della "proposta" e del lancio, ossia la primissima ricezione dell'elemento, e in quelli più tardi il crescente e 'popolarizzato' uso del termine e dunque la celebrità dell'oggetto. Posso virtualmente avere risultati più raffinati di quanto qualsiasi insieme di metadati mi permetterebbe, ma sono esposto anche a rischi molto maggiori.

29 Vedere ad esempio l'ottima BD cronologica dedicata a Leonardo da Vinci: http://leopardi.letteraturaoperaomnia.org/index.html

30 Per i risultati di una simile ricerca: https://books.google.com/ngrams/graph?content=hubble+telescope&case_insensitive=on&year_start=1988&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t4%3B%2Chubble%20telescope%3B%2Cc0%3B%2Cs0%3B%3BHubble%20telescope%3B%2Cc0%3B%3BHubble%20Telescope%3B%2Cc0%3B%3BHUBBLE%20TELESCOPE%3B%2Cc0

5. CASI PARTICOLARI

Open Library

Esistono moltissimi casi che potremmo definire “intermedi” rispetto ai modelli che abbiamo visualizzato e che meriterebbero una trattazione a parte. **Open Library**[31] gestisce i propri contenuti con una suddivisione per soggetti e categorie estremamente dettagliata.

Tuttavia il suo scopo dichiarato, creare “una pagina web per ogni libro sul pianeta”, non ne fa una BD e neanche una raccolta di collegamenti come Europeana, ma un insieme di pagine che confermano l'esistenza di un libro con determinate caratteristiche.
Si tratta insomma di un'altra raccolta di metadati puri, ma gestiti attraverso vere e proprie categorie costruite a mano: in breve, metadati gestiti tramite metadati.

In generale tuttavia la maggioranza delle raccolte e delle BD sul web si muove con pochi metadati e una ricerca full-text non troppo sofisticata.

31 https://openlibrary.org/

WrittenSound

Esistono poi raccolte di materiale estremamente particolare, che si trovano ai confini del concetto di BD. WrittenSound[32] ad esempio aspira a raccogliere tutte le onomatopee immaginabili della lingua inglese. Questo peculiare compito permette di gestire un sistema di metadati particolarmente fluido ed efficace che consente di cercare non solo il nome dell'elemento di cui vogliamo trovare l'onomatopea (ad esempio, 'elicottero' per trovare elementi come 'whop whop'), ma anche le caratteristiche del suono cercato: *pitch, tonality, loudness*, e così via.

6. SVILUPPI FUTURI

Di fronte a una indagine complessa, tentiamo di restringere il campo calibrando le nostre ricerche e chiarendoci le idee durante l'esplorazione per raggiungere i siti specifici in cui tutto si fa più semplice. Se, per riprendere l'esempio iniziale, cerchiamo testi di narrativa in qualunque forma che contengano un Boeing 247, cercheremo di incappare in siti come "impdb.org", che è diviso per elementi che ricorrono in sceneggiature e presenta una intera pagina sul Boeing 247, sfoggiando ben sei titoli[33]. Useremo questa pagina come un ulteriore piccolo hub per cercare i titoli in appositi archivi contenenti migliaia di copioni più o meno ordinati e indicizzati. Questa è per così dire una forma di indagine ancora molto empirica e manuale che il web consente di svolgere con rapidità. Gli strumenti di ricerca

32 http://www.writtensound.com/
33 http://www.impdb.org/index.php?title=Category:Boeing_247

possono fare di più? Molto probabilmente sì, sia da un punto di vista della Linguistica Computazionale sia da un punto di vista dell'implementazione dei metadati.

Si è ipotizzato al principio di questo lavoro che il web potrebbe strutturarsi in un insieme di piccolissime BD ultra-specializzate[34] gestite da un grande motore di ricerca privo di metadati che funzioni con un meccanismo alla 'hubs and authorities'. Se le micro-BD rischiano però di rimanere soffocate nelle grandezze del Web, il sistema 'hubs and authorities' provoca anche una grave *cristallizzazione* dei propri risultati, quando autorità ed hubs diventino troppo forti e continuino a rafforzarsi a vicenda: in molti casi questo può andare perfino a scapito di una ricerca generica.

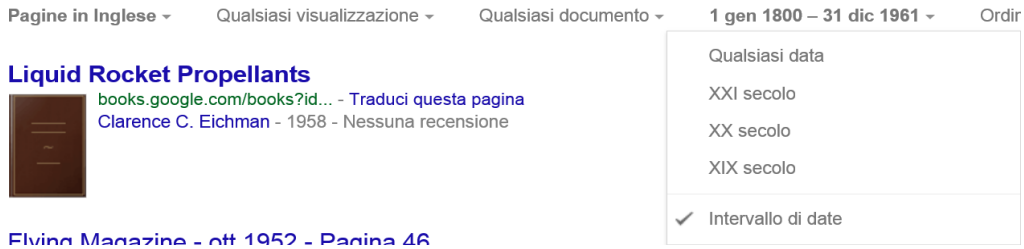
Illustrazione 8: Se cerco "Pratt & Whitney tutti i brevetti possibili" su Google, benché esistano raccolte di brevetti online di questi autori non ottengo come primo risultato un archivio o una BD, ma pagine di Wikipedia: è il fenomeno della cristallizzazione delle autorità.

Dagli strumenti che fanno uso di questo sistema dobbiamo attenderci una forte tendenza 'conservatrice': a lungo andare vinceranno sempre le stesse applicazioni del web, gli stessi libri, lo stesso linguaggio, gli stessi profili. Nell'introduzione ho esposto il meccanismo di ricerca manuale reiterata su cui sono basati strumenti digitali automatici come l'algoritmo *PageRank*. Data la tendenza a rafforzarsi a vicenda di autorità ed hubs, le conclusioni del processo tenderanno sempre più a premiare gli stessi risultati.

34 Che non avrebbero il problema dell'indicizzazione manuale, avendo a che fare con pochissimi elementi. Vedere ad esempio le raccolte di articoli tecnici nei siti dedicati a singoli autori: http://www.prattandwhitney.com/Content/Technical_Papers.asp

Per linguisti e lessicologi gli strumenti a ricerca 'full text' si stanno ormai moltiplicando e costituiscono la risorsa migliore: da Google Correlate che consente di analizzare le ultime ricerche degli utenti agli archivi digitalizzati dei giornali che permettono di cercare anche molte decadi nel passato.

Da questo punto di vista i modelli alla Ngram Viewer, fondati sulla ricerca full-text in intervalli di date, hanno un brillante futuro.

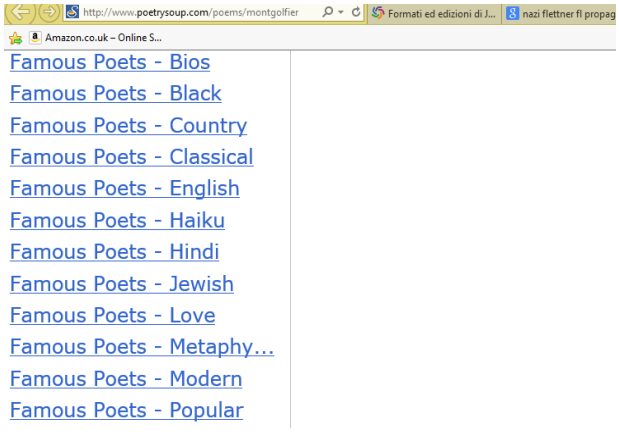


Allo stesso tempo è credibile che le BD specializzate, che già sono moltissime sul Web[35], siano destinate ad aumentare, con suddivisioni e metadati manuali curati dagli utenti. I due sistemi sembrano insomma destinati a competere ancora per un po'.

7. CONCLUSIONI

Se si cerca un tipo di documento, un genere, un argomento non particolarmente tecnico, le BD antologizzate manualmente e fornite di molti metadati sono la migliore risorsa che si possa avere. In questo campo la varietà è grande: si possono trovare suddivisioni per autori nominati, per qualità del testo, per nazionalità, e così via.

35 Per farsi un'idea approssimativa, esistono lunghi elenchi di links a biblioteche digitali esistenti: http://www.digital-librarian.com/literature.html, http://lang.nagoya-u.ac.jp/~matsuoka/EngLit.html



Modelli di BD simil-enciclopedici come *Wikisource*, che facciano ricorso a suddivisioni estremamente dettagliate e ragionate degli argomenti, stanno avendo recentemente un più che discreto successo.

- **Introduction**
- **Background History**
  - East vs West: The Ultimate Love-Hate Relationship
  - How did it all start?
  - How was the Cold War fought?
- **First Tensions**
  - A Divided Europe
  - Yalta
  - Potsdam
  - The Truman Doctrine

Il caso di Wikisource è emblematico perché ha instaurato connessioni stabili con il suo modello ideale Wikipedia, che fornisce a propria volta articoli analitici multilingue su una varietà di argomenti, offrendo potremmo dire 'descrizioni molto estese' dei testi che vengono infine collegati alle voci.

Illustrazione 9: Un'enciclopedia come Wikipedia permette di cercare informazioni molto dettagliate e grandi quantità di links a testi e siti d'autorità nelle più diverse lingue del mondo

Sull'altro versante, i sistemi che tentano di sostituire il ricorso a metadati con tecniche di information retrieval più elaborate rischiano di funzionare solo 'per antologie', ossia per testi che abbiano già in qualche misura in sé quel metatesto che manca alla ricerca.
Per altri aspetti, una indagine quasi del tutto libera da categorie prestabilite, non basata su un meccanismo di metadescrizioni, garantisce come abbiamo visto una fluidità e continuità di risultati fondamentale per molti tipi di ricerca.

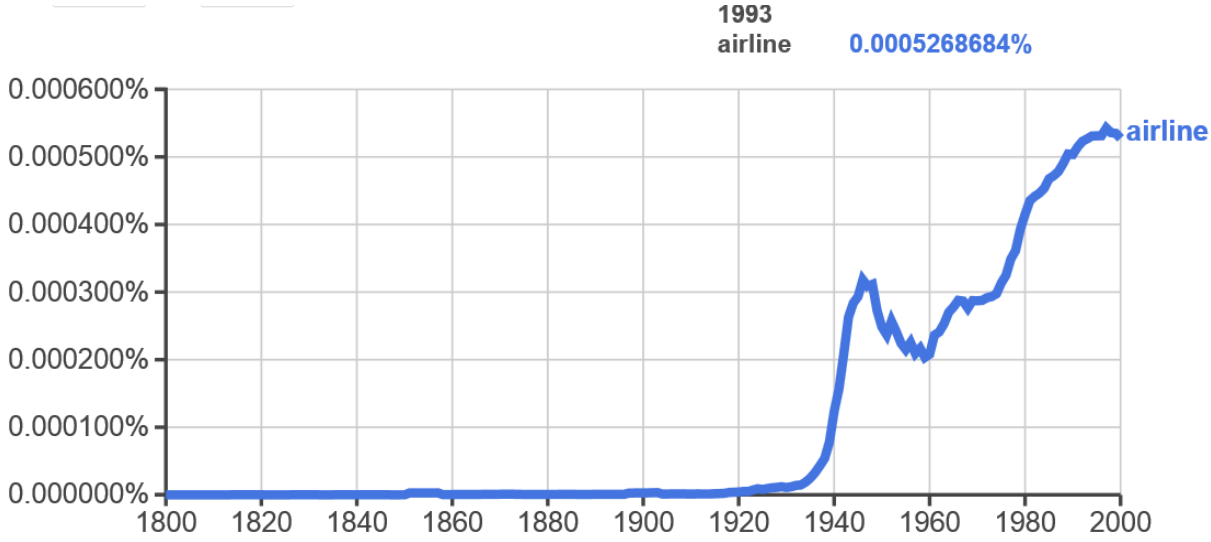
Mega Cities: The European Space Agency's Contribution to...

books.google.it/books?isbn... - Traduci questa pagina
Lothar Beckel - 2001 - Anteprima

International Cooperation in Space: The Example of the ...

books.google.it/books?isbn... - Traduci questa pagina
Roger-M. Bonnet, Vittorio Manno - 1994 - Anteprima - Altre edizioni
This text describing the the European Space Agency shows how such a co-operative enterprise has worked over the past 30 years and how

Trattare la conoscenza formalizzata di una libreria, come descrizioni e ontologie, alla stregua di una parte stessa della conoscenza da indagare consente a volte rappresentazioni più efficaci di quelle che utilizzano le descrizioni come metadato.
Cercare discussioni della parte mediana degli anni Quaranta sui voli di linea è fattibile solo attraverso una rappresentazione piuttosto complessa del contenuto se la ricerca deve essere fatta via metadati, ma è molto semplice per ricerche full-text.



Non bisogna però dimenticare che i motori di ricerca 'full text', per quanto ben fatti ed accurati,cadono ancora spesso in fraintendimenti. Un caso classico è quello in cui il motore dia importanza a una parola polisemica all'interno della ricerca, prendendo lucciole per lanterne.

Se questi approcci sono straordinariamente utili per i linguisti e i lessicologi e consentono una dinamicità e internazionalità delle ricerche che categorie e metadati non potrebbero fornire in alcun modo, e l'aggiunta ingegnosa di sistemi a 'wildcards' permette di raffinare ulteriormente le proprie ricerche, non bisogna dimenticare che essi non sono ancora in grado di fornire risultati ragionati o 'intelligenti', tanto meno di distinguere il genere o il tipo di documento cercato.

Illustrazione 11: Le wildcards permettono ad esempio di visionare i risultati di ricerche su diversi corpora in un solo grafico

In breve, perfino Ngram è meno comodo di Wikipedia per ricerche specifiche come: testi di letteratura *pionieristica* sui principi del Turbojet; o: critica letteraria dell'Ottocento al genere gotico. In Ngram Viewer posso al massimo cliccare sul primo scaglione offerto e vedere cosa contenga. Con Wikipedia posso fruire di pagine come *Storia del jet* che indirizzano ai teorizzatori pionieristici della tecnologia, e tramite le note bibliografiche raggiungere links ad articoli d'avanguardia sulla possibilità di motori jet[36]. Il rischio che comporta, e non è affatto indifferente, è quello dell'errore umano, volontario e involontario.

Non esiste fino ad ora un modo vincente per rappresentare la conoscenza né per ovviare alla deviazione di ricerche fraintese verso risultati non inerenti. Lo sviluppo di strumenti e algoritmi di linguistica computazionale avanzata, pur ognuno nel proprio ambito specifico, permetterà probabilmente rappresentazioni automatiche del contenuto di un testo abbastanza efficaci da superare una buona parte dell'attuale dicotomia “metadati- non metadati”.
L'introduzione delle biblioteche digitali è avvenuta certamente attraverso un maggior uso di categorie e tags di quanti vengano ora utilizzati nel web e nelle BD, ma data la loro importanza i metadati sono ancora irrinunciabili, a meno di perdere, in conclusione, grandi possibilità di ricerca.
Se lo scopo di una BD è di rendere accessibile la conoscenza, in generale entrambi i sistemi, con diverse strade e risorse, sono avviati su una strada di successo, uno con la duttilità delle ricerche su materiale grezzo, l'altro con la precisione delle categorie.
I due sistemi sono sufficientemente importanti da dover essere, per ora, integrati per ottenere il miglior uso possibile della conoscenza da parte dell'utenza e le BD che saranno in grado di trarre il meglio da entrambi gli approcci e di farli convivere coerentemente probabilmente saranno le biblioteche vittoriose del prossimo futuro.

8. BIBLIOGRAFIA

BIBLIOGRAFIA

Anonimo. "A global library resource". *Online Computer Library Center*. Gennaio 24, 2014

Chowdhury. “Introduction to digital libraries”. *Facet Publishing*, 2002.

Hart, Michael S.. "Gutenberg Mission Statement by Michael Hart". *Project Gutenberg,* 15 August 2007.

Kessler et al.. 'Automatic Detection of Text Genre'. *Arxiv*, 2013.

Marshall et al.. 'Annotation: from paper books to digital library'. *Arxiv*, 2009.

36 Tramite questa semplice sequenza di passi si arriva davvero a BD specializzate con articoli coerenti alla ricerca: http://www.flightglobal.com/pdfarchive/view/1941/1941%20-%202221.html

Sullivan, Danny. "What Is Google PageRank? A Guide For Searchers & Webmasters", *Search Engine Land,* 2013.

SITOGRAFIA

http://www.perseus.tufts.edu/hopper/
http://etcsl.orinst.ox.ac.uk/
http://etcsl.orinst.ox.ac.uk/edition2/etcslbycat.php
http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html
http://it.wikisource.org/wiki/Pagina_principale
http://online.wsj.com/articles/for-this-author-10-000-wikipedia-articles-is-a-good-days-work-1405305001
http://www.poetryfoundation.org/
http://www.docstoc.com/docs/
http://www.freepatentsonline.com/
http://onlinebooks.library.upenn.edu/banned-books.html
http://arizona.openrepository.com/arizona/handle/10150/105066/browse?type=subject
http://www.library.utoronto.ca/canpoetry/pratt/
http://rpo.library.utoronto.ca/
https://archive.org/
http://www.europeana.eu/
http://www.wdl.org/en/
http://www.google.com/patents/
https://books.google.com/ngrams
http://leopardi.letteraturaoperaomnia.org/
http://www.writtensound.com/
https://openlibrary.org/
http://www.impdb.org/
http://www.prattandwhitney.com/
http://www.digital-librarian.com/
http://www.flightglobal.com/