Google Book Search e le politiche di digitalizzazione libraria

Gino Roncaglia (Università della Tuscia)

Versione preliminare e non rivista di un intervento per il n. 2/2009 di Digitalia (http://digitalia.sbn.it/genera.jsp?s=1&l=it). Osservazioni e commenti sono benvenuti.

Questo intervento si propone di presentare e discutere brevemente le caratteristiche, la situazione attuale e le prospettive future del progetto di digitalizzazione libraria avviato da Google (*Google Book Search*), a poco più di cinque anni dalla sua presentazione, avvenuta nell'ottobre 2004 in occasione dall'annuale fiera del libro di Francoforte.

Si tratta di un tema assai discusso[1], e rispetto al quale le posizioni sono spesso fortemente cristallizzate: sostegno entusiastico a un progetto considerato – a ragione – come il primo tentativo di digitalizzazione libraria su scala globale, o, al contrario, rifiuto altrettanto netto di scelte percepite, a seconda dei casi, come culturalmente 'imperialiste' nella selezione delle priorità, dei partner e dei testi da digitalizzare (che vedono una larga prevalenza dell'inglese), guidate da logiche commerciali e non culturali, poco attente alla salvaguardia dei diritti esistenti sulle opere acquisite (e in questo caso le resistenze vengono soprattutto dalle grandi case editrici), o al contrario fonte di ulteriori restrizioni nelle modalità di accesso ad alcune tipologie di testi, e in particolare, come vedremo, alle opere orfane.

Può essere dunque il caso di premettere subito che questo intervento non intende né proporre una difesa di ogni aspetto del progetto, né partire da una posizione aprioristicamente critica: il mio obiettivo è quello di presentare e discutere alcune fra le caratteristiche fondamentali dell'iniziativa avviata da Google, cercando di argomentarne razionalmente limiti e vantaggi (spesso, come vedremo, strettamente intrecciati). Nel farlo, mi riallaccio a due interventi precedenti sugli stessi temi[2], rispetto ai quali viene qui evidentemente proposto un quadro aggiornato e più articolato, ma partendo da premesse che sono in gran parte le stesse.

Le ragioni di un progetto

Per i nostri scopi, credo sia utile ricordare in primo luogo che quello di Google è un progetto sviluppato da parte di un'azienda che ha il suo *core business* nella ricerca e nell'indicizzazione di informazione. Come sappiamo, Google nasce come motore di ricerca, ed è a questa funzionalità di motore di ricerca, capace di

[1] Per una bibliografia piuttosto ampia (ma limitata alla lingua inglese) di articoli e interventi su Google Book Search si veda Charles W. Bailey, Google Book Search Bibliography, all'indirizzo http://www.digital-scholarship.org/gbsb/gbsb.htm. Nel momento in cui scrivo, la bibliografia è disponibile nella versione 5, aggiornata al 14 settembre 2009. Tutte le pagine web citate in questo articolo sono state visitate l'ultima volta il 9 dicembre 2009.

[2] Gino Roncaglia, *I progetti internazionali di digitalizzazione bibliotecaria: un panorama in evoluzione*, in «Digitalia» n. 1/2006, pp. 11-30, in rete alla pagina http://digitalia.sbn.it/upload/documenti/digitalia20061_RONCAGLIA.pdf; Gino Roncaglia, intervento in *Google a "Fahrenheit"*, «AIB Notizie» n.1/2005, in rete alla pagina http://www.aib.it/aib/editoria/n17/0501baldacroncaglia.htm.

conquistarsi sul campo una posizione di assoluta preminenza grazie agli eccellenti algoritmi di ordinamento (*ranking*) dei risultati in base alla loro pertinenza per l'utente, che si sono man mano collegati i servizi che garantiscono la redditività dell'azienda. In primo luogo il 'microadvertising' contestuale, che costituisce la principale fra le fonti di reddito: basti pensare che i ricavi pubblicitari costituiscono ad oggi (dati del terzo trimestre 2009[3]) il 97% delle entrate di Google (il 67% proveniente da siti Google, il 30% da siti partner) rispetto a un 3% derivante da entrate legate alla vendita di licenze o di altro tipo di servizi. E si tratta di cifre di tutto rispetto: quasi 17 miliardi di dollari nei soli primi nove mesi del 2009.

Perché, allora, un motore di ricerca dovrebbe imbarcarsi in un progetto di digitalizzazione libraria? La risposta si basa a mio avviso su tre considerazioni fondamentali: in primo luogo, perché si tratta di lavorare con una quantità di informazione assolutamente compatibile – come ordine di grandezza – con quella già presente in rete e che Google è abituato ad indicizzare[4]. In secondo luogo perché – a differenza di molta parte dell'informazione grezza presente in rete – i libri contengono per lo più informazione validata, dunque di alto valore potenziale per gli utenti di un motore di ricerca. Infine, perché in questo modo Google può anticipare e cavalcare una tendenza che vede evidentemente all'orizzonte: la transizione verso il digitale anche dei supporti per la lettura (e dunque in primo luogo dei libri), dopo quelli utilizzati per la musica, per le immagini e per il video[5].

Non è un caso, dunque, che quello di Google sia un progetto di digitalizzazione *libraria* più che strettamente un progetto di digitalizzazione *bibliotecaria*. Almeno in prima battuta, a Google in fondo non interessa – se non nell'aspetto giuridico ed economico di definizione dei relativi accordi, aspetto che ovviamente ha in pratica, come vedremo, un ruolo assolutamente essenziale – se i materiali digitalizzati provengano da una biblioteca o da una casa editrice, se si tratti di testi antichi o di pubblicazioni appena uscite. Il progetto di digitalizzazione dell'azienda di Mountain View nasce come naturalmente onnivoro, guidato in primo luogo dal criterio della potenziale appetibilità dei contenuti rispetto alle ricerche effettuate dagli utenti. Gli accordi con le biblioteche e quelli con le case editrici, pur diversi in natura, rispondono a questo obiettivo comune.

La natura onnivora del progetto si scontra però, inevitabilmente, con i diversi regimi giuridici e con le diverse caratteristiche proprie di oltre cinque secoli di produzione libraria a stampa. I libri recenti sono naturalmente i più appetibili dal punto di vista delle ricerche degli utenti, ma sono ancora sotto diritti, e i diritti sono detenuti da case editrici diverse, che possono avere – e di fatto hanno – atteggiamenti diversi rispetto all'iniziativa di Google. I libri meno recenti si trovano in biblioteca, e la loro digitalizzazione – peraltro spesso più complessa per via delle modalità di stampa e delle condizioni di conservazione degli esemplari utilizzati – richiede dunque accordi con almeno alcune fra le grandi biblioteche di riferimento. Anche in questo caso ci sono inoltre spesso problemi di diritti, sia per le opere ancora in commercio sia per la vasta classe delle cosiddette 'opere orfane', quelle per le quali risulta difficile reperire l'autore, o risalire alla sua data di morte (e quindi alla data di scadenza dei diritti), o individuare gli eredi.

---

[3] Cf. Google Inc., *Public Release of Third Quarter 2009 Financial Results*, Oct. 15 2009, in rete all'indirizzo http://investor.google.com/pdf/2009Q3_earnings_google.pdf. La quota di entrate di Google derivanti dall'advertising è stata peraltro negli ultimi anni sostanzialmente costante.

[4] Cf. Gino Roncaglia, *I progetti internazionali di digitalizzazione bibliotecaria* cit.

[5] A questo tema sarà dedicato un mio lavoro in uscita nel 2010 per i tipi della casa editrice Laterza.

Un po' di storia (e di dispute legali)

La situazione che abbiamo fin qui delineato, almeno in partenza semplice e lineare negli obiettivi ma ben presto complessa e controversa negli sviluppi organizzativi, giuridici e legali, spiega un po' la storia del progetto avviato da Google. Inizialmente, l'azienda di Mountain View aveva previsto (e avviato) un lavoro di digitalizzazione relativamente uniforme nella gestione di tipologie diverse di libri, sostenendo che la digitalizzazione di opere ancora sotto diritti avveniva al solo scopo di indicizzarne i contenuti, rendendo disponibili agli utenti solo gli indici e degli 'snippets' dell'opera (brevi frammenti o 'ritagli' del testo, la cui disponibilità Google riteneva fondata sul diritto di citazione, o *fair use*), corrispondenti alle ricerche effettuate dagli utenti stessi.

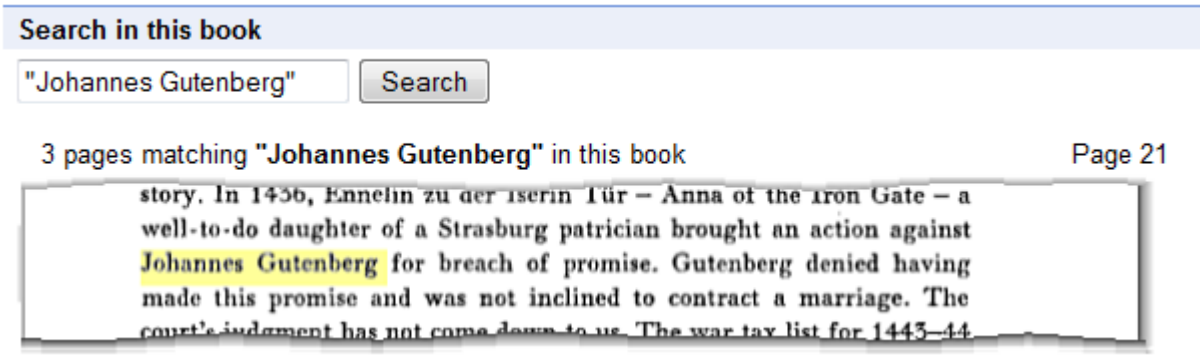

Figura 1 - Un esempio di 'snippet' restituito da Google Book Search. Il passo è tratto da Elisabeth Geck, *Johannes Gutenberg; from lead letter to the computer*, Inter nationes, Bad Godesberg 1968

Ma la situazione, si è detto, doveva rivelarsi assai più complicata. Davanti alle prime critiche, per rafforzare la tesi che il lavoro effettuato riguardasse in primo luogo la realizzazione di strumenti di indicizzazione e ricerca, Google abbandonò subito – a pochi mesi dall'avvio del progetto – il primo nome scelto, *Google Print*, a favore del più esplicito *Google Book Search*, e differenziò esplicitamente al suo interno il programma di digitalizzazione relativo ai testi in commercio (avviando al riguardo i primi contatti con gli editori) e il programma di digitalizzazione bibliotecaria (*Google Books Library Project*), che a sua volta cercava e trovava fin dal dicembre 2004 importanti partnership, fra cui le biblioteche delle Università di Harvard, di Stanford e del Michigan, la New York Public Library e la biblioteca dell'università di Oxford, che include la famosa Bodleian Library.

La tesi di fondo sostenuta da Google – e la politica 'opt out' adottata al riguardo, che prevedeva l'esclusione dal progetto di digitalizzazione bibliotecaria dei soli editori che ne facessero esplicita richiesta, anziché l'inclusione dei soli editori esplicitamente intenzionati ad aderire – restava tuttavia fortemente controversa, e veniva in particolare contestata sia dalla Authors Guild of America sia dalla Association of American Publishers, che nel 2005 intentavano a Google due cause separate, sostenendo – fra gli altri argomenti addotti – che la realizzazione degli indici si basava comunque sulla gestione di banche dati on-line contenenti il full-text delle opere digitalizzate, e che nel caso di opere ancora sotto diritti tale gestione, indipendentemente dalla disponibilità o meno del full text delle opere per gli utenti finali, costituiva di per sé una violazione massiccia e sistematica del copyright.

Gli aggiustamenti operati da Google – incluso il passaggio a politiche opt-in di esplicita e consensuale partnership con gli editori per quanto riguarda la gestione delle opere in commercio – non potevano evidentemente risolvere il problema di fondo, relativo a questo punto soprattutto al progetto di digitalizzazione bibliotecaria: con quali procedure e con quali limiti è possibile la digitalizzazione di opere sotto diritti, sia pure con il solo scopo di indicizzarle? Che tipo di risultati può restituire la ricerca, per restare all'interno del 'fair use' dell'opera? La rapida apparizione in rete di strumenti software più o meno clandestini in grado di 'rippare' (e cioè catturare e trasformare in testo riutilizzabile, in analogia con quanto fatto dai programmi in grado di estrarre dai loro supporti originali musica e video protetti) le pagine visualizzate da Google Book Search non aiutava, ovviamente, la posizione di Google.

La disputa giuridica che ne è seguita è stata fra le più discusse e commentate degli ultimi anni, e non è mia intenzione analizzarla in questa sede, anche perché ne esistono già trattazioni competenti e approfondite[6]. Basti dire che la prima proposta di accordo (*settlement agreement*) che ne è risultata è lunga, fra testo e appendici, oltre 300 pagine[7]. Fra i suoi punti qualificanti è la creazione di un Registro indipendente per i diritti su libri (*Book Rights Registry*) , che dovrebbe raccogliere in maniera aperta e trasparente informazioni sui detentori di diritti relativi a opere librarie, permettendo di utilizzare tali informazioni sia nell'ambito del progetto di digitalizzazione portato avanti da Google – in modo da compensare gli aventi diritto – sia nel caso di altre iniziative analoghe curate da altre società. In base alla proposta di accordo, Google si impegna a versare a tale registro una cifra iniziale di 34,5 milioni di dollari, e a una serie di altri adempimenti per un impegno finanziario complessivo di 125 milioni di dollari. La proposta di accordo prevede anche una divisione dei ricavi della vendita on-line di opere sotto diritti digitalizzate da Google (singolarmente o sulla base di abbonamenti cumulativi) o dei ricavi derivanti dalla pubblicità aggiunta da Google ai risultati di ricerche sul contenuto di tali opere, nella proporzione del 63% per autori ed editori detentori dei diritti, e del 37% per la stessa Google.

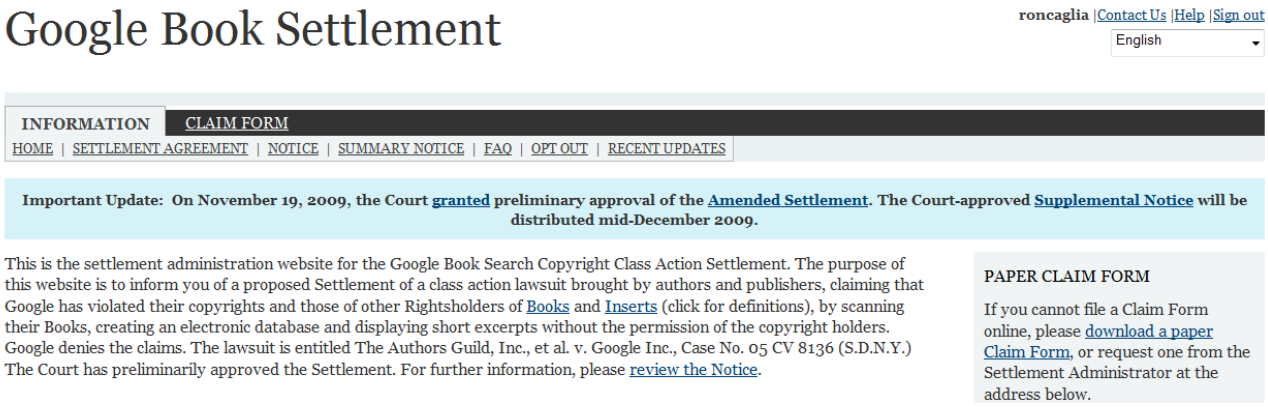

Figura 2 - Il sito del Google Book Settlement: http://www.googlebooksettlement.com/

[6] Accanto alla già citata bibliografia di Charles W. Bailey mi limito a ricordare, in ambito italiano, l'approfondito articolo di Antonella De Robbio, *La gestione dei diritti nelle digitalizzazioni di massa. Un'analisi alla luce del caso Google Book Search*, in «Bibliotime», anno XII, numero 2 (luglio 2009), disponibile anche in rete alla pagina http://www2.spbo.unibo.it/bibliotime/num-xii-2/derobbio.htm.

[7] La proposta di accordo, le sue successive modifiche e altri materiali relativi alla causa, assieme ai moduli per la rivendicazione dei diritti d'autore sulle opere comprese nel database di Google Book Search, sono disponibili alla pagina http://www.googlebooksettlement.com.

Le polemiche proseguono…

Come era facile prevedere, la proposta di accordo non ha certo messo a tacere le polemiche, e le ha anzi in diversi casi rafforzate, rimescolando però non poco i ruoli e i temi discussi. Se in precedenza, come si è visto, la disputa legale era fra Google da un lato e Authors Guild e Association of American Publishers dall'altro, dopo la diffusione del testo della proposta di accordo la controversia vede Google e le sue due controparti su posizioni assai più vicine, impegnate a rispondere alle critiche spesso assai aspre formulate nei confronti dell'accordo da almeno quattro fronti, a loro volta spesso intrecciati e variamente sovrapposti: 1) i concorrenti diretti di Google, a cominciare da Microsoft, Yahoo! e Amazon; 2) le associazioni in difesa della libera circolazione dell'informazione, a cominciare dall'Internet Archive, pronte spesso ad alleanze inattese con le grandi corporation anti-Google; 3) altre associazioni di autori ed editori e altre case editrici, soprattutto straniere, che considerano i propri diritti danneggiati da un accordo nel quale non sono parti in causa; infine, come vedremo fra breve, 4) istituzioni pubbliche (anche statunitensi) e governi esteri, preoccupati dalla definizione attraverso accordi privati di temi legati a interessi pubblici prevalenti, come quello alla libera circolazione della conoscenza, e dal rischio di creazione di posizioni dominanti o monopolistiche. Se qualcuno fra i lettori di questo articolo è un appassionato di legal thriller, la lettura anche solo di una piccola parte dell'immensa documentazione depositata presso la corte statunitense incaricata di esaminare il caso[8] basterà a soddisfarlo per mesi e mesi.

Lettere e memorandum fortemente contrari rispetto alla proposta di accordo vengono innanzitutto dagli Stati Uniti, compreso il Department of Justice, che osserva come

> "(…) the Proposed Settlement—especially the forward-looking business arrangements it seeks to create—raises significant legal concerns. As a threshold matter, the central difficulty that the Proposed Settlement seeks to overcome—the inaccessibility of many works due to the lack of clarity about copyright ownership and copyright status—is a matter of public, not merely private, concern. A global disposition of the rights to millions of copyrighted works is typically the kind of policy change implemented through legislation, not through a private judicial settlement."[9]

Fra i punti più controversi dell'accordo è il meccanismo proposto per la gestione delle opere orfane. Tali opere, rispetto alle quali erano state in passato portate avanti numerose iniziative, anche legislative, tendenti all'adozione di politiche liberali di gestione dei diritti[10] – risulterebbero invece fortemente 'chiuse'

[8] La documentazione è consultabile anche in rete, a partire dalla pagina http://news.justia.com/cases/featured/new-york/nysdce/1:2005cv08136/273913/.

[9] Cf. http://digital-scholarship.org/digitalkoans/2009/09/20/u-s-department-of-justice-files-objection-to-google-book-search-settlement/ .

[10] Così, ad esempio, negli Stati Uniti era stato proposto nel 2003 e di nuovo nel 2005 un *Public Domain Enhancement Act* che avrebbe consentito il passaggio nel pubblico dominio di un gran numero di opere orfane in caso di mancato rinnovo del relativo copyright. Tale proposta, tuttavia, non è mai stata approvata dal Congresso. Proposte di legislazione su questo tema sono comunque ancora in discussione negli Stati Uniti sia alla Camera sia al Senato, e verrebbero inevitabilmente influenzate in maniera notevole – e in senso restrittivo – dalla formulazione dell'accordo fra Google e le associazioni di autori ed editori. Politiche di apertura relativamente alle opere orfane, una volta esercitata la "due diligence" nel tentativo di individuare i detentori dei relativi diritti, anche attraverso la creazione di infrastrutture in grado di favorire lo scambio di informazioni al riguardo (come il progetto europeo ARROW, su cui cf. Piero Attanasio, *Un approccio cooperativo per la gestione dei diritti nelle biblioteche digitali: il progetto ARROW*, in «Digitalia» n. 2/2008, pp. 55-62), sono state del resto sollecitate in molte sedi. A livello europeo, il *Final Report on Digital Preservation, Orphan Works, and Out-of-Print Works* adottato nel giugno 2008 nell'ambito dell'iniziativa i2010 Digital Libraries ad opera del High Level Expert Group – Copyright Subgroup (il testo è in rete all'indirizzo

sulla base dell'accordo, che prevede un accesso per lo più a pagamento (sulla base di licenze istituzionali, individuali o – nel solo ambito bibliotecario – di accesso pubblico). Il rischio è evidentemente quello – direttamente richiamato proprio dal memorandum del Department of Justice già citato – di offrire a Google una sorta di monopolio di fatto sulla loro gestione, dal momento che per altri operatori diventerebbe assai oneroso e complesso costituire collezioni alternative di contenuti e negoziarne indipendentemente le condizioni di distribuzione.

Significativa al riguardo è la dichiarazione di Brewster Kahle, co-fondatore dell'Internet Archive e della Open Content Alliance e fra gli oppositori più decisi all'iniziativa di Google, che esprime in maniera assai decisa le proprie perplessità:

> "The settlement has been seen as solving the orphan books issue, which has served to enervate efforts to pass a bill. If Google were to abandon its attempt to grab these books for its own private gain, then technology companies and libraries could speak with a strong voice, speaking in unison, working together to get proper legislation passed. (…) Lets free the orphans, not have them pass from their legal limbo into a life controlled by Google and its proposed Books Rights Registry".[11]

Per rendersi conto dell'importanza della questione delle opere orfane, va considerato che la grande maggioranza dei libri pubblicati nel corso del secolo scorso è composta da opere o sotto copyright o orfane: dei sette milioni di libri che Google dichiara di aver digitalizzato al novembre 2008[12], circa un milione è composto da opere di pubblico dominio, circa un milione da opere in commercio, e circa cinque milioni sono orfane o sotto copyright ma fuori commercio[13].

Non meno controversa è la questione dell'applicabilità dell'accordo al di fuori degli Stati Uniti: non a caso, nella documentazione depositata troviamo lettere e memorandum di numerosissimi editori europei (compresa l'Associazione Italiana Editori) e quelli – evidentemente concordati nei contenuti – dei governi di Francia e Germania. Può essere utile, anche solo per curiosità, citare per esteso qualche passo del memorandum presentato dal governo tedesco:

> "The Federal Republic of Germany ("Germany") hereby submits this Memorandum of Law to apprise the Court of the significant adverse impact that the proposed Settlement would have on German authors, publishers and digital libraries, in particular, and more generally on authors, publishers and digital libraries in the European Union ("EU") and other countries outside the United States ("U.S."). Although this submission is directed specifically to the interest of the *amicus curiae*, it is believed that many of the issues troubling German authors, publishers and digital libraries will be of similar concern to authors, publishers and digital libraries in the United States and throughout the world.

---

http://ec.europa.eu/information_society/activities/digital_libraries/doc/hleg/reports/copyright/copyright_subgroup_final_report_26508-clean171.pdf) si basa sulla considerazione per cui "Comprehensive, large scale digitisation and online accessibility could be greatly hampered, if adequate solutions are not found to the problem of orphan works" (p. 10), raccomandando l'adozione di politiche europee comuni su questo tema e offrendo una cornice dettagliata per il concetto di "diligent search".

[11] Brewster Kahle, *Google Claims to be the Lone Defender of Orphans: Not lone, not defender*, post all'interno del blog della Open Content Alliance, http://www.opencontentalliance.org/2009/10/07/google-claims-to-be-the-lone-defender-of-orphans-not-lone-not-defender/

[12] Nel 2009, il numero complessivo di libri digitalizzati nell'ambito del progetto dovrebbe essersi avvicinato ai dieci milioni.

[13] Robert Darnton, *The Case for Books*, Public Affairs, New York 2009, p. 14.

> The 134-page proposed Settlement (not including attachments) sets forth what purports to be a comprehensive system for exploitation of copyrights in the digital age. This system, however, has not been enacted by any legislative or industry body, as would, without regard to commercial advantage or profit, typically take into account domestic and worldwide developments and trends in the industry and copyright law. Instead, to the contrary, the proposed Settlement is a privately-negotiated document that is shrouded in secrecy, formulated behind closed doors by three interested parties, the Authors Guild, the Association of American Publishers and Google, Inc. ("Google"), resulting in a commercially driven document that is contrary to established international treaties and laws. These principally include the world's oldest and most venerable multilateral copyright treaty, the Berne Convention ("Berne"), [1] as well as the more recent World Copyright Treaty (the "WCT") which attempts to regulate international copyright relations in the digital age. [2] The Settlement also runs afoul of the applicable German national laws, as well as European public initiatives to create non-commercial worldwide digital libraries."[14]

Un testo, come si vede, assai esplicito, e il lettore erudito potrà leggere con curiosità anche il seguito del documento[15], che per riaffermare il legittimo interesse del governo tedesco nella vicenda non manca di ricordare con enfasi il contributo tedesco alla cultura letteraria mondiale, il ruolo di Gutenberg come inventore della stampa, l'importanza storica delle fiere librarie tedesche.

Largamente analogo (compresi i riferimenti culturali, che in questo caso spaziano da Pascal a Descartes, da Voltaire a Diderot, da Molière a Racine, da Baudelaire a Sartre) è il documento francese, il cui incipit è identico a quello tedesco ma che prosegue sottolineando con particolare chiarezza il tema della salvaguardia della diversità culturale e linguistica europea:

> "Of particular concern to the French Government is the threat the proposed Settlement poses to cultural diversity. By placing such unchecked, concentrated power over the digitization and making available of the vast wealth of literary works created from 1923 through January 5, 2009 in a single private entity, the Settlement positions Google as the gatekeeper of public knowledge in a manner likely to stifle the rich diversity of voices heard and enjoyed by France and other nations"[16]

Fra gli altri temi controversi è la gestione delle stesse opere di pubblico dominio, che vengono messe a disposizione da Google in versioni accompagnate da licenze comunque non completamente aperte e, nel caso dei file PDF, accompagnate da una 'filigrana digitale' (*digital watermarking*) per identificarne l'origine.

Infine, controversa è anche la previsione di alcune limitazioni all'accesso e all'uso dei testi da parte delle biblioteche partner del progetto, per le quali era invece inizialmente prevista la piena disponibilità di una copia dei file di tutti i libri digitalizzati in collaborazione con Google. Per questi motivi, e per le considerazioni già ricordate legate alle modalità di gestione delle opere orfane, la biblioteca dell'università di Harvard, che era stata fra i primi partner di Google, ha minacciato recentemente di abbandonare l'iniziativa. E si tratta di una presa di posizione tanto più importante in quanto viene da una delle figure di

[14] Il documento è riportato in copia anastatica all'indirizzo http://docs.justia.com/cases/federal/district-courts/new-york/nysdce/1:2005cv08136/273913/179/, e può essere (forse più agevolmente) letto in formato testo all'indirizzo http://www.book-grab.com/germany2.txt.

[15] Come ha fatto peraltro anche Robert Darnton, nell'articolo *Google and the New Digital Future*, in «The New York Review of Books», Volume 56, Number 20, December 17, 2009; il testo dell'articolo è disponibile in rete all'indirizzo http://www.nybooks.com/articles/23518.

[16] http://www.book-grab.com/france.pdf.

riferimento nella discussione su storia e natura del libro e sul suo futuro nell'era digitale[17], il già citato Robert Darnton, che proprio della prestigiosa biblioteca di Harvard ha assunto nel 2007 (e dunque dopo l'avvio del progetto) la direzione.

Il futuro della causa legale, e della proposta di accordo fra le parti, è a questo punto tutt'altro che chiaro. Molto probabilmente, nel momento in cui leggerete questo articolo la situazione sarà già almeno in parte cambiata. Il 13 novembre 2009 è stato infatti presentato al tribunale dalle parti un accordo di patteggiamento modificato, che prevede alcuni cambiamenti di un qualche interesse; fra gli altri, l'allargamento delle licenze per l'accesso pubblico al database di opere orfane nel contesto bibliotecario (la prima proposta prevedeva che l'accesso pubblico potesse avvenire in una sola postazione per biblioteca, indipendentemente dalle dimensioni della biblioteca stessa) e una maggiore apertura, almeno sul piano formale, a iniziative di digitalizzazione alternative a quella di Google. Ma il cambiamento dal nostro punto di vista più significativo (e sul quale torneremo fra breve) è l'accordo per eliminare dalla transazione i libri non pubblicati o registrati negli Stati Uniti, in Canada, nel Regno Unito o in Australia.

Queste modifiche lasciano tuttavia sostanzialmente inalterato l'impianto generale dell'accordo, e non interrompono per il momento il procedimento legale: il tribunale dovrà decidere se accettare la nuova versione della transazione, ci sarà la possibilità di acquisire ulteriori memorie e prese di posizione, e gli avvocati delle parti coinvolte avranno indubbiamente occasione di chiedere e di riscuotere nuove parcelle. La disputa legale (e la disputa culturale) intorno al progetto di digitalizzazione libraria di Google insomma proseguirà, così come nelle decine di biblioteche coinvolte nel progetto proseguirà, parallelamente, il lavoro di scannerizzazione e acquisizione dei testi.

In questa situazione, la tentazione di disinteressarsi di una questione che sembra presentare troppi aspetti strettamente giuridico-legali e troppi aspetti strettamente tecnologici per risultare appassionante sotto il profilo culturale può essere comprensibilmente forte. Ma sottovalutare il rilievo culturale di quanto sta succedendo sarebbe un gravissimo errore. Dobbiamo sempre ricordare che dietro i cavilli legali e dietro le astrusità tecnologiche c'è in questo caso un progetto – la creazione di un'enorme biblioteca digitale globale, in linea di principio accessibile e ricercabile da parte di chiunque – che ha una portata culturale potenzialmente gigantesca, tale da far impallidire l'idea della biblioteca di Alessandria e da avvicinarsi piuttosto alla Biblioteca di Babele di Borgesiana memoria. Disgraziatamente, fare le mosse giuste in questo contesto è tanto più importante quanto più è difficile individuare *quali siano* le mosse giuste da fare.

Quale ruolo per l'Europa?

La già ricordata esclusione dall'accordo – fra gli altri – dei libri editi nell'Europa continentale è intesa evidentemente a rispondere alle proteste che come abbiamo visto erano state avanzate da editori e governi del vecchio continente, ma il risultato concreto rischia di rivelarsi un boomerang proprio per chi rivendicava l'esigenza di una maggiore apertura del progetto di Google verso lingue diverse dall'inglese e per la tradizione culturale europea.

Sostanzialmente, Google risponde a queste critiche rinunciando per il momento a rendere accessibili in maniera integrale opere orfane provenienti dal mondo non anglofono, e rilanciando la palla ai suoi critici.

[17] Si veda in particolare R. Darnton, *The Case for Books* cit.

Critici che – occorre dirlo con chiarezza – non si sono rivelati per ora capaci di avviare o di portare avanti con efficacia iniziative altrettanto ambiziose o comprensive. *Europeana*[18], che intende proporsi come il deposito digitale di riferimento per la cultura europea, aspetta una normativa europea comune per le opere orfane, che dovrebbe sfruttare il già ricordato registro europeo ARROW, ma che al momento non sembra dietro l'angolo. Ed è nel frattempo limitata a contenuti di pubblico dominio peraltro ancora lontani dall'essere pienamente rappresentativi della ricchezza culturale del vecchio continente, anche per le enormi differenze (quantitative e qualitative) nei contributi provenienti da paesi diversi.

Per farsi un'idea del problema, basti citare quanto osserva Susanne Bjørner in un recente intervento:

> "It's not just more works that must be added. More equal participation from all member libraries is required, and original-language content must be increased. As of July, France is still the biggest contributor to Europeana, providing 47% of the content, though that is down from the 52% share it had on launch. Germany is next with 15%; the Netherlands and the U.K. each have contributed 8%; Sweden, Finland, and Norway (which is not even an EU member) hover at 4%-5%; and others trail far behind with 1% or less. A more noticeable problem for many users-and an embarrassing one since Europeana prides itself on offering interfaces in the native language of all its members-is that many national treasures are not there in the original language. If you do a search on Da Vinci, for example, you will find 460 objects in French but only 140 in Italian. Shakespeare is represented with 606 items in French and only 317 in English. The Spanish writer Cervantes has 176 objects in French and just 43 in Spanish. And you might think that the Danish author Hans Christian Andersen was really German, with 387 items provided in German and only one in Danish!" [19].

In questo contesto, le dichiarazioni rilasciate a fine agosto 2009 da Viviane Reding, Commissario europeo per la società dell'informazione e i media, che hanno suscitato scandalo per l'apertura a Google proprio nel momento in cui almeno parte del vecchio continente sembrava fieramente avversarne l'iniziativa, sono in realtà comprensibili e tutt'altro che peregrine:

> The Commission takes the view that digitisation of cultural products, including books, is a Herculean task that requires close cooperation between right holders and ICT companies, as well as between the public and the private sector. The Commission therefore supports an approach that is open to private-sector initiatives and to technological innovation. Ideological answers are certainly not the best way forward for stimulating innovation, creativity and consumer-friendly offers of digital content. (…) Commercial projects alone certainly cannot cover the public interest dimension of the digitisation of cultural products. (…) However, the Commission also looks with interest to new solutions now tested between Google and right holders in the US for making orphan works (works under copyright where the author cannot be identified) better accessible to a broader public. Finding solutions to the orphan works issue is a very important part of the EU's objective to preserve cultural works. The Commission therefore recommends that stakeholders in Europe look very closely at the discussions in the U.S. to see how the experience made there could best be used for finding a European solution on orphan works. "[20]

[18] http://www.europeana.eu/portal/.

[19] Susanne Bjørner, *Europeana and Digitization: The Collaboration Is Only Beginning*, in «Information Today», September 10, 2009, in rete alla pagina http://newsbreaks.infotoday.com/NewsBreaks/Europeana-and-Digitization-The-Collaboration-Is-Only-Beginning-56079.asp.

[20] http://www.euractiv.com/29/images/Reding%20Position%20Google%20Books_tcm29-184905.doc. La dichiarazione afferma anche il sostegno per una eventuale collaborazione fra il progetto Google Books e le biblioteche nazionali francese e italiana (che dovrebbe comunque riguardare solo opere di pubblico dominio).

In questo quadro, la decisione presa dai ministri della cultura dell'Unione Europea, riuniti a Brussels il 27 novembre 2009, di creare "a bloc-wide project to digitize books, starting with the formation of a committee to work out a blueprint for the plan"[21], rappresenta certo una nobile dichiarazione di intenti, ma appare difficilmente realizzabile (e pericolosamente vicina a tante altre dichiarazioni di intenti fatte in passato, che non sempre hanno portato ai risultati concreti auspicati) senza una qualche forma di accordo che includa partner in grado di portare avanti concretamente – e in tempi ragionevoli – il lavoro.

Può valer la pena, allora considerare anche alcuni altri fattori di questa complessa equazione: fattori che l'asprezza della disputa legale può portare a sottovalutare, e che invece è sicuramente opportuno tenere ben presenti.

Google e la concorrenza

Un primo fattore importante è rappresentato dai risultati fin qui ottenuti dal progetto. Abbiamo già ricordato che al dicembre 2008 Google dichiarava il completamento della digitalizzazione di circa sette milioni di libri, e che a fine 2009 questo numero è arrivato a superare i dieci milioni (tra i due e i quattro milioni dei quali sono rappresentati da libri europei[22]). Si tratta di un risultato sicuramente notevole, che mostra sia l'enorme impegno dell'azienda di Mountain View in un progetto nel quale evidentemente crede molto (e molto ha investito), sia la rapidità con la quale – disponendo di fondi adeguati e di volontà politica adeguata – la digitalizzazione del nostro patrimonio culturale potrebbe procedere.

Si può affermare senza timore di smentita che, nonostante le controversie che abbiamo appena esaminato, e nonostante i numerosi limiti e le numerose imperfezioni anche tecniche del lavoro portato avanti da Google, su cui torneremo fra breve, un'azienda privata è riuscita ad avviare un progetto che nessuna istituzione pubblica di alcun paese al mondo ha avuto (purtroppo) la forza, le risorse e la capacità di affrontare in maniera altrettanto determinata ed efficace. Rispetto alle dimensioni del lavoro avviato da Google, i progetti nazionali e istituzionali di digitalizzazione bibliotecaria – pur se di norma assai più validi dal punto di vista scientifico – hanno dimensioni e numeri quasi amatoriali[23]. Mentre Google estrae già petrolio – pur se non raffinato e di qualità spesso decisamente scadente – i suoi concorrenti istituzionali sono impegnati in carotaggi esplorativi, certo più sofisticati ma al momento assai meno redditizi in termini di risorse informative effettivamente acquisite e messe a disposizione del pubblico.

[21] Cf. http://www.dw-world.de/dw/article/0,,4942737,00.html.

[22] Il dato è stato fornito dai rappresentanti di Google nel corso di una audizione sul Google Book Search Settlement e sulle sue conseguenze per l'editoria e gli autori europei, tenuta alla Commissione Europea il 7 settembre 2009: cf. http://www.ifrro.org/upload/documents/NOTESandANNEX_EC_Hearing_Google_Settlement_7Sept09_1.pdf.

[23] È ben vero che Europeana dichiara di raccogliere circa cinque milioni di 'oggetti digitali', ma nel conto sono inclusi anche immagini, brani sonori e video, record catalografici. Per dimensioni, il secondo progetto internazionale di digitalizzazione libraria – dopo Google Book Search – è HathiTrust (http://www.hathitrust.org/), che vede la partecipazione delle biblioteche di 12 università del Midwest (fra cui la University of Michigan, la University of Iowa e la University of Illinois) e di 11 biblioteche della University of California. La base dati di HathiTrust comprende al momento quasi 5 milioni di volumi digitalizzati, ma la grande maggioranza di essi proviene proprio dal progetto Google Book Search, di cui le biblioteche partecipanti sono partner. Si tratta dunque più di una forma di conservazione ridondante finalizzato all'uso dei testi in un contesto universitario che di un progetto alternativo: il suo vero interesse è nel proposito di predisporre strumenti sofisticati di *data mining* e analisi testuale sui *corpora* testuali acquisiti.

Fra i concorrenti di Google, un ruolo particolare – ma comunque caratterizzato da una minore capacità di produrre risultati tangibili, nonostante le dimensioni di alcuni dei partner privati inizialmente coinvolti – ha l'Open Content Alliance. Riassumiamone brevemente la storia, che come vedremo chiama in causa il principale concorrente di Google, Microsoft.

Mentre attorno al progetto di Google si sviluppavano le prime polemiche, Microsoft aveva infatti avviato un progetto rivale, denominato *Live Search Books*. Si trattava di un'iniziativa esplicitamente anti-Google, che rifiutava le posizioni dell'azienda di Mountain View in materia di *fair use*, rifiutava la digitalizzazione libraria basata su politiche *opt-out* (richiedendo invece l'esplicita adesione delle case editrici interessate), e si concentrava in primo luogo – almeno inizialmente – sulla digitalizzazione di opere di pubblico dominio. È difficile pensare che le scelte di Microsoft fossero in questo caso legate all'adesione convinta verso politiche di accesso aperto (che l'azienda di Redmond contrastava in molti altri settori): si trattava piuttosto dell'esigenza, più che comprensibile dal punto di vista delle logiche di mercato, di attaccare i punti deboli di un'iniziativa rivale, della quale Microsoft avvertiva il potenziale rilievo, ma che era stata avviata e gestita da parte del suo più pericoloso concorrente commerciale.

In questa situazione, Microsoft agiva innanzitutto per catalizzare consenso e sottrarlo al progetto di Google, stringendo un'alleanza (che non poteva non far sollevare qualche sopracciglio) con associazioni del calibro dell'Internet Archive, paladine della diffusione aperta dei contenuti, e contribuendo – insieme ad Yahoo! e ad altri partner pubblici e privati – alla costituzione nell'ottobre 2005 dell'Open Content Alliance. Iniziativa, questa, vista inizialmente con grande favore anche in Europa: la British Library aveva aderito al progetto, e attenzione era stata manifestata dalla Bibliothèque Nationale de France, all'epoca presieduta da Jean-Noël Jeanneney, uno dei più attivi oppositori di Google Book Search e dell'idea di una singola 'biblioteca universale', al cui interno Jeanneney temeva il prevalere di logiche commerciali e il predominio della produzione culturale anglofona[24].

L'avvio del progetto aveva lasciato ben sperare: la sola Microsoft aveva contribuito alla digitalizzazione di circa 750.000 opere, con una qualità media probabilmente maggiore di quella garantita da Google. Ma nello stesso periodo, come abbiamo visto, Google Book Search aveva digitalizzato diversi milioni di volumi. E Microsoft si era nel frattempo accorta che, almeno dal punto di vista strettamente commerciale, concentrarsi sulle opere fuori diritti non portava lontano: ai fini delle necessità della maggior parte degli utenti e dello sfruttamento nell'ambito di un motore di ricerca, una volta acquisiti alcune decine di migliaia di 'classici' maggiori e minori, i testi pubblicati prima del '900 – pur preziosi per gli studiosi – avevano un interesse assai più limitato di quelli pubblicati nel corso del secolo scorso e di quelli in commercio. Come abbiamo ricordato, oltre l'80% dei libri digitalizzati da Google è costituito da opere orfane o sotto diritti, e proprio queste opere costituiscono il principale vantaggio competitivo del progetto avviato dall'azienda di Mountain View.

Così nel maggio 2008, dopo un tardivo tentativo di includere anche testi sotto copyright avviato l'anno precedente, Microsoft decide un clamoroso ripensamento, e abbandona il proprio progetto.

[24] L'assai discusso pamphlet di Jeanneney *Quand Google défie l'Europe : plaidoyer pour un sursaut*, Mille et une Nuits, Paris, 2005, riassume le ragioni di questa polemica, e ha avuto una notevole influenza anche nel mondo anglofono, grazie alla traduzione inglese edita nel 2007 dalla Chicago University press sotto il titolo *Google and the Myth of Universal Knowledge*. Interessanti sono al riguardo i materiali raccolti nella primavera 2008 da Randal C. Picker per il *Tech Policy Seminar* della University of Chicago Law School: http://picker.typepad.com/picker_seminar/week_1/ .

Nel farlo, lascia in eredità all'Open Content Aliance – che rimane attiva – una ricca dote di volumi, ora accessibili attraverso la Open Library[25] e integrati con altre opere acquisite nell'ambito dell'iniziativa (e con i soli dati bibliografici di quasi 23 milioni di altri libri). Ma lascia anche un'eredità di recriminazioni e problemi per le biblioteche che si erano 'fidate' del progetto e che si sono ritrovate improvvisamente prive di quello che ne doveva essere il principale partner privato (e il principale finanziatore).

Una situazione, come è facile capire, che ha forse sottratto inizialmente a Google alcuni potenziali partner nel mondo bibliotecario, ma che ne ha alla lunga confermato il ruolo di punto di riferimento più affidabile – e di fatto al momento unico – per la digitalizzazione libraria su grande scala.

La questione dei formati e della qualità

La battaglia legale attorno al *settlement agreement* e l'enorme attenzione che si è – a ragione – concentrata sulle vicende legali del progetto Google, hanno lasciato decisamente troppo in ombra gli aspetti che rappresentano, dal punto di vista scientifico e teorico, il cuore di ogni progetto di digitalizzazione libraria: le scelte operate in sede di rappresentazione del testo, i formati utilizzati per la sua codifica, la qualità e affidabilità del lavoro svolto.

Ogni digitalizzazione parte da una specifica forma del testo – quella costituita dalla particolare edizione a stampa considerata – e porta a una forma testuale diversa ma altrettanto specifica, quella dell'oggetto digitale che viene prodotto. In questo passaggio, vengono operate scelte ben precise rispetto alla rappresentazione dei fenomeni testuali. Alcuni di questi fenomeni vengono rappresentati, spesso attraverso il ricorso a specifiche *marcature* del testo, altri possono non esserlo. E il testo elettronico risultante dal processo di digitalizzazione può avere caratteristiche diverse, anche *molto* diverse.

Al livello più basso, può trattarsi solo di una sorta di 'fotocopia digitale' dell'originale, in cui il testo non è ricercabile perché all'immagine digitalizzata non è mai stato associato il corrispondente testo elettronico. Possiamo poi integrare – e si tratta della scelta per molti versi preferibile – l'immagine del testo a stampa con un testo elettronico prodotto attraverso un procedimento di riconoscimento ottico dei caratteri, e auspicabilmente corretto attraverso una revisione manuale (gli OCR – soprattutto se applicati a edizioni non recenti – producono di norma testi assai scorretti). Possiamo infine abbandonare ogni pretesa di conservare una rappresentazione della forma originale a stampa del documento ed affidarci alla sola rappresentazione digitale del testo, offrendone una nuova reimpaginazione o consentendone la reimpaginazione dinamica in funzione del dispositivo di lettura utilizzato.

In tutti questi casi, vanno prese decisioni che incidono in maniera essenziale sulla qualità del lavoro: per fare solo qualche esempio, il formato dell'immagine, la sua risoluzione, la scelta di lavorare su immagini a colori o in scala di grigi (e l'ampiezza della tavolozza di colori o di grigi utilizzata), l'insieme di metadati (cioè di informazioni descrittive e gestionali) associata alle immagini e al testo elettronico, e così via. Se si utilizza un programma OCR, bisogna decidere quale programma utilizzare, e capire, ad esempio, se e in che misura è in grado di gestire testi multilingui (ad esempio, testi in una lingua che contengano brani o citazioni in lingue diverse): un OCR pensato per l'italiano produrrà spesso risultati disastrosi davanti a una breve citazione latina. E il testo risultante dal processo di digitalizzazione potrà essere o no marcato, ed essere

[25] http://openlibrary.org/.

marcato utilizzando insiemi di marcatori diversi: dall'insieme curato e validato rappresentato dai marcatori TEI (Text Encoding Initiative)[26] a insiemi di marcatori assai più ridotti, o addirittura fuori standard.

Su tutti questi aspetti, vitali per l'affidabilità scientifica del lavoro di digitalizzazione, il progetto Google dichiara pochissimo, e il poco che dichiara – o che è ricavabile dal lavoro svolto o dalle indicazioni fornite da chi vi ha partecipato – non sempre è soddisfacente. Ovviamente Google non si accontenta della digitalizzazione delle immagini (né potrebbe farlo, dato che come abbiamo visto il suo obiettivo principale è rendere ricercabili i testi), ma il tipo di formato e di marcatura del testo associato all'immagine della pagina restano poco chiari. Si tratta presumibilmente di una marcatura XML, dato che – grazie anche a una collaborazione con Sony, che produce dispositivi di lettura (*e-book reader*) compatibili con il formato aperto Epub[27] – Google Book Search offre da qualche mese la possibilità di scaricare circa 500.000 testi di pubblico dominio, oltre che in formato PDF, anche in formato Epub. Questo è un fattore positivo, ma l'assenza di informazioni attorno alla specifica marcatura utilizzata (e ai relativi metadati) non lo è[28].

Presumibilmente, l'assenza di informazioni è in questo caso legata anche al carattere di 'work in progress' proprio dell'iniziativa di Google: Jon Orwant, uno dei responsabili del progetto, ha dichiarato nel corso della conferenza *Tools of Change for Publishing* organizzata da O'Reilly nel febbraio 2009 che

> (…) the ultimate goal of Google Book Search is to convert images to 'original intent' XML. He explained the post-processing Google runs to continuously improve the quality of the scanned books, and to convert images to structured content. Retro-injecting structure accurately is no mean feat but when it's done, Google will be able to transform the books into a variety of formats. The content becomes mutable and transportable, in a sense it isn't yet, even though it is scanned, online and searchable."[29]

È auspicabile che questo lavoro possa avvenire attraverso l'adozione di formati di codifica aperti e documentati e di insiemi standard di metadati, scelti in collaborazione con il mondo della ricerca. Le biblioteche partner dovrebbero prestare un'attenzione assai maggiore a questo aspetto del progetto, e condizionare la loro collaborazione all'adozione di standard e politiche trasparenti al riguardo. Scelte che, del resto, sono nel miglior interesse della stessa Google, e solo grazie alle quali sarà possibile utilizzare le basi dati testuali che verranno a costituirsi attraverso il lavoro di digitalizzazione non solo per ricerche

---

[26] Sulla Text Encoding Initiative cf. http://www.tei-c.org/; in Italiano, si veda *Il manuale TEI Lite: introduzione alla codifica elettronica dei testi letterari*, edito da Fabio Ciotti, Edizioni Sylvestre Bonnard, Milano 2005. Del rapporto fra TEI e progetti di digitalizzazione libraria su larga scala si è recentemente occupato John Unsworth, nel corso del suo intervento al meeting annuale del Consorzio TEI, svoltosi ad Ann Arbor nel novembre 2009. Un abstract del suo intervento è alla pagina http://www.lib.umich.edu/spo/teimeeting09/keynotes.html e le slide sono scaricabili dall'indirizzo http://www3.isrl.illinois.edu/~unsworth/TEI.MM09.pptx.

[27] Il formato Epub – promosso dall' International Digital Publishing Forum (IDPF) – è basato su XML, nasce per la realizzazione di e-book in grado di garantire la reimpaginazione dinamica del testo in funzione della dimensione del carattere scelto dall'utente per la visualizzazione e in funzione delle dimensioni dello schermo, ed è l'erede del formato OEB-PS. È molto usato dai dispositivi di lettura per e-book (*e-book reader*) di seconda generazione, e si sta rapidamente affermando come lo standard per la realizzazione di libri elettronici. Per maggiori informazioni, il riferimento principale è il sito dell'IDPF, all'indirizzo http://www.openebook.org/.

[28] Google ha reso disponibili nel settembre 2008 le API per Google Book Search, e questo rende possibile includere in altri siti (*embedding*) l'anteprima del libro visualizzabile attraverso il sito. Tuttavia, le API distribuite non dialogano direttamente con la codifica XML del testo, che rimane nascosta.

[29] La sintesi è di James Long, nell'intervento *My tee oh see* in «the digitalist. a blog by the digital team at Pan Macmillan», in rete alla pagina http://thedigitalist.net/?p=447. L'intervento di Orwant è disponibile in video all'indirizzo http://blip.tv/file/1781212/.

occasionali (e con poche garanzie sulla completa affidabilità dei risultati) ma anche per forme di *data mining* e di analisi testuale complesse[30].

A questo aspetto è legato anche il miglioramento della qualità del lavoro di scannerizzazione e di restituzione del testo elettronico. Al momento, in molti casi sembra che il prodotto dell'applicazione di un OCR alle immagini scannerizzate sia poco o per nulla controllato da operatori umani competenti[31], e che i metadati siano spesso estratti in automatico dal sistema sulla base del lavoro di parser e algoritmi evidentemente ancora largamente imperfetti. Il risultato è quello che Geoffrey Nunberg, uno dei maggiori esperti del mondo dei libri elettronici, definisce enfaticamente ma correttamente "a disaster for scholars", osservando che "the book search's metadata are a train wreck: a mishmash wrapped in a muddle wrapped in a mess"[32] e riportando una lista impressionante di errori di ogni genere, spesso evidenti, talvolta elusivi, non di rado sistematici, e in alcuni casi ai limiti del comico.

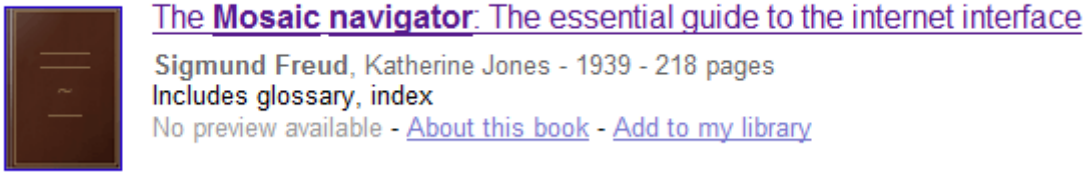

Figura 3 - Uno dei più divertenti fra gli errori di Google Books segnalati da Nunberg: avreste mai sospettato che Sigmund Freud non solo avesse previsto l'avvento di Internet, ma avesse scritto già nel 1939 una "essential guide to the internet interface"? Peccato che di questo libro non sia disponibile un'anteprima…

Le conclusioni di Nunberg mi sembrano interamente condivisibili, e perfettamente compatibili con le considerazioni proposte all'inizio di questo articolo: più che a una politica volontariamente trascurata o disattenta, questi problemi sono legati al fatto che il progetto è figlio del carattere 'onnivoro' di Google come motore di ricerca.

> I have the sense that a lot of the initial problems are due to Google's slightly clueless fumbling as it tried master a domain that turned out to be a lot more complex than the company first realized. It's clear that Google designed the system without giving much thought to the need for reliable metadata. In fact, Google's great achievement as a Web search engine was to demonstrate how easy it could be to locate

[30] Analisi di questo tipo pongono evidentemente ulteriori problemi nel caso delle opere sotto diritti. Il *settlement agreement* offre al riguardo un passo interessante, relativo alle attività eseguibili dall'utente nell'ambito delle biblioteche che aderiscono al progetto, che includono "research in which computational analysis is performed on one or more Books, research in which a researcher does not read or display substantial portions of Book to understand the intellectual content presented" (un tipo di ricerca definito *non-consumptive*).

[31] Per qualche esempio, si veda David Ludlow, *Google turns classic books into free gibberish eBooks*, in «Computer Shopper», 27 August 2009, http://www.expertreviews.co.uk/news/267379/google-turns-classic-books-into-free-gibberish-ebooks.html.

[32] Geoffrey Nunberg, *Google's Book Search: A Disaster for Scholars*, in «The Chronicle of Higher Education», August 31, 2009, in rete alla pagina http://chronicle.com/article/Googles-Book-Search-A/48245/. Si veda anche Geoffrey Nunberg, *Google Books: A Metadata Train Wreck*, nel blog «Language Log» del Linguistic Data Consortium presso l'University of Pennsylvania: http://languagelog.ldc.upenn.edu/nll/?p=1701. Quest'ultimo intervento è seguito fra gli altri da un lungo commento di Jon Orwant, che spiega le ragioni di alcuni degli errori e ribadisce l'impegno di Google nel migliorare la situazione. Molti degli errori rilevati da Nunberg sono però ancora riscontrabili – non corretti – nel dicembre 2009.

> useful information without attending to metadata or resorting to Yahoo-like schemes of classification. But books aren't simply vehicles for communicating information, and managing a vast library collection requires different skills, approaches, and data than those that enabled Google to dominate Web searching.

È non solo auspicabile ma necessario che questi errori vengano corretti attraverso un lavoro sistematico di revisione e attraverso il miglioramento degli algoritmi utilizzati: le biblioteche che partecipano al progetto possono lavorare nella prima direzione, mentre Google – che del progressivo perfezionamento degli algoritmi utilizzati ha fatto la propria arma vincente – dovrebbe dedicare un'attenzione assai maggiore alla seconda.

Conclusioni

Come si vede, la situazione è estremamente complessa ed è difficile adottare al riguardo posizioni manichee. Google non è il diavolo: al momento, soprattutto se sarà possibile intervenire sui (gravi) problemi di qualità della scannerizzazione e dei metadati che abbiamo appena ricordato, e se vi sarà una maggiore apertura e una più attenta discussione sui formati di codifica utilizzati, è il partner probabilmente più efficiente per progetti di digitalizzazione libraria su larga scala. Ma è comunque un'azienda privata con obiettivi e priorità che non coincidono necessariamente con quelle di una biblioteca o di una istituzione culturale. E alcuni aspetti del *settlment agreement* suscitano indubbiamente, anche nella sua versione modificata, più di una perplessità, anche se va detto che le perplessità manifestate dagli editori internazionali sono assai diverse, e talvolta di segno opposto, rispetto a quelle manifestate da chi vorrebbe una maggiore apertura nella gestione delle opere orfane.

In generale, quel che ci insegna l'esperienza di questi anni è che la digitalizzazione libraria è un'impresa costosa, complessa, per molti versi controversa, ma preziosa e necessaria: portare in rete il nostro patrimonio librario vuol dire garantirne vitalità e fruizione anche nell'era del digitale, allargarne la reperibilità, renderlo (per la prima volta) pienamente integrato e ricercabile. E questo anche in vista della disponibilità di dispositivi di lettura in ambiente elettronico che cominciano lentamente ad avvicinarsi alle caratteristiche di 'perfezione ergonomica'[33] proprie del libro a stampa. Si tratta dunque di una scelta corretta sia dal punto di vista culturale sia da quello economico, anche se assai impegnativa su entrambi i fronti.

Ma è un'impresa che comporta inevitabilmente un work in progress, un lavoro continuo, nel quale le stesse opere saranno probabilmente digitalizzate più volte, anche da soggetti diversi e con criteri diversi. Un'impresa alla quale contribuiranno soggetti pubblici e privati, che dovranno imparare non solo a convivere ma a collaborare. In questo lavoro, Google ha e continuerà ad avere nei prossimi anni sicuramente un ruolo importante, ma ha bisogno di partner che non si limitino a porgergli libri da digitalizzare o a chiedergli soldi e attrezzature: ha bisogno di partner che lo sappiano incalzare e che ne sappiano orientare opportunamente le scelte. Non solo le scelte giuridiche, ma anche quelle tecniche e culturali. Così come scelte importanti dovranno fare i governi, e in primo luogo i governi europei: scelte non di arroccamento, ma di adozione di standard e politiche condivise. La rete può permettersi di avere una pluralità di biblioteche digitali frutto di progetti diversi, ma solo se si tratterà di iniziative pienamente

[33] Cf. Jean-Claude Carriére e Umberto Eco, *Non sperate di liberarvi dei libri*, Bompiani, Milano, 2009.

interoperabili e condotte con criteri di sostenibilità economica di lungo periodo. Per raggiungere questi obiettivi, l'apertura alla collaborazione e l'uso dell'argomentazione razionale sono armi preferibili rispetto alle guerre legali.