DigItalia

Numero 1 - **2006**

Rivista del digitale nei beni culturali

ICCU-ROMA

I progetti internazionali di digitalizzazione bibliotecaria: un panorama in evoluzione

Gino Roncaglia
Università della Tuscia

L'articolo offre un panorama della storia e delle caratteristiche dei principali progetti di digitalizzazione libraria, in particolare in ambito bibliotecario. Per l'autore, la digitalizzazione bibliotecaria è di particolare importanza per i motori di ricerca, dato che permette di estendere la loro copertura a contenuti validati, di alto valore informativo e potenzialmente anche commerciale. Inoltre, nonostante le sue caratteristiche e le sue esigenze specifiche, la complessità di una indicizzazione full-text *dei testi prodotti da progetti di digitalizzazione bibliotecaria è paragonabile a quella dell'indicizzazione del Web, rendendo questo settore un campo ideale per l'espansione dei motori di ricerca (e per la vera e propria battaglia commerciale che li vede contrapposti).*

Questa situazione può offrire alle biblioteche e ai ricercatori una straordinaria opportunità per ridurre il gap esistente fra cultura del libro e culture di rete, fra contenuti a stampa e contenuti digitali. Tuttavia, è necessaria una particolare attenzione per evitare che il rilievo e la natura anche commerciale di molti progetti di digitalizzazione bibliotecaria possano contribuire in qualche modo a limitarne il valore scientifico. Una cattiva qualità nella digitalizzazione, la mancanza o l'incompletezza di metadati adeguati, la scelta di formati proprietari, una limitata interoperabilità, politiche troppo restrittive nella tutela del copyright, sono tutti fattori che possono limitare in maniera anche seria l'utilità del lavoro svolto. L'articolo discute in dettaglio i progetti di digitalizzazione collegati a Amazon Book Search, a Google Book Search (già Google Print), alla Open Content Alliance e al progetto di Biblioteca digitale europea. Attenzione è riservata anche alle controversie suscitate dalle scelte di Google Book Search relative alla digitalizzazione di contenuti sotto copyright anche senza un esplicito consenso dei detentori dei diritti.

Le prime fasi: volontariato telematico e mondo della ricerca

Negli oltre trent'anni compresi fra il 1971 – anno in cui Michael Hart avviò il progetto Gutenberg[1] – e il 2002, il tema della digitalizzazione del patrimonio

[1] Primo, pionieristico e tuttora attivo progetto di biblioteca digitale, il progetto Gutenberg si basa sulla digitalizzazione in formato solo testo di opere fuori diritti in lingua inglese (i testi digitalizzati erano circa 18.000 nell'aprile 2006). Nel corso del tempo, iniziative analoghe sono state avviate anche per altre lingue europee (per l'Italia il progetto Manuzio, nato nel 1993: http://www.liberliber.it/biblioteca/index.htm). Per una storia dei primi anni del progetto

testuale dell'umanità, attraverso la creazione di vere e proprie *biblioteche digitali*[2], si è progressivamente imposto come uno dei più rilevanti per quanto riguarda l'applicazione in ambito culturale delle nuove tecnologie digitali. Ma fino al 2003, nonostante la progressiva crescita di competenze specializzate anche attraverso l'avvio di servizi commerciali come Ebrary[3], Questia[4], NetLibrary[5], il tema della digitalizzazione libraria è rimasto legato in primo luogo al mondo della ricerca e a quello della gestione dei beni culturali, con un contributo iniziale da parte del volontariato telematico. A questa prima fase appartengono un gran numero di iniziative pilota e di progetti locali e nazionali[6], alcuni dei quali – come il francese Gallica[7] – particolarmente impegnativi.

Un incontro fra mondo della ricerca e volontariato telematico era alla base anche della più ambiziosa fra le prime iniziative collaborative di digitalizzazione bibliotecaria, il Million Book Project, nato nel 2001 dalla collaborazione fra le Carnegie Mellon University Libraries e l'Internet Archive. L'aspetto più innovativo del progetto è sicuramente rappresentato dalla collaborazione con istituzioni universitarie e di ricerca in India e in Cina, con l'obiettivo di allargare la digitalizzazione anche a testi in lingue non occidentali; anche la maggior parte dei centri di digitalizzazione è fisicamente collocata in India o in Cina. Il progetto si proponeva la digitalizzazione di un milione di libri entro il 2005, una scadenza che si è però rivelata troppo ottimistica: attualmente, il raggiungimento di questo obiettivo è stato spostato alla fine del 2007. Lo stato effettivo di completamento del lavoro è comunque un po' nebuloso[8]: sono stati infatti scannerizzati – a fine 2005 – oltre 600.000 volumi (dei quali solo 135.000 in inglese), ma a fine aprile 2006 solo

Gutenberg si veda Michael Hart, *History and Philosophy of Project Gutenberg*, August 1992, http://www.gutenberg.org/about/history; per gli sviluppi più recenti si vedano i materiali disponibili sul sito del progetto, http://www.gutenberg.org/. Queste e tutte le altre risorse in rete citate sono state visitate l'ultima volta nell'aprile 2006.

2 Non entreremo qui nei problemi legati alla definizione e alla natura delle biblioteche digitali; per una introduzione all'argomento si vedano, fra i testi più recenti, Michael M. Lesk, *Understanding Digital Libraries*, San Francisco-Amsterdam: Morgan Kaufmann (Elsevier), 2005 (2nd. ed.); Lucy A. Tedd – Andrew Large, *Digital Libraries*, London: Bowker-Saur, 2004, e – in ambito italiano e con nutriti riferimenti bibliografici – Riccardo Ridi, *La biblioteca digitale: definizioni, ingredienti e problematiche*, «Bollettino AIB», 44 (2004), n. 3, p. 273-344. Per notizie aggiornate sull'evoluzione del settore si può consultare, fra le molte risorse disponibili in rete, «D-Lib Magazine», http://www.dlib.org/.

3 http://www.ebrary.com/.

4 http://www.questia.com/.

5 http://www.netlibrary.com/.

6 Un quadro di alcune fra le principali iniziative in quest'ambito, accompagnato da un elenco ricercabile di circa 25.000 testi in inglese liberamente disponibili in rete, è offerto dalla The Online Books Page della University of Pennsylvania: http://digital.library.upenn.edu/books/.

7 http://gallica.bnf.fr/. Il progetto, avviato nel 1997, ha portato finora alla digitalizzazione di oltre 80.000 testi, la maggior parte dei quali in formato immagine (i testi digitalizzati in formato testuale sono a fine 2005 circa 1.250).

8 Per i dati, si vedano le *Frequently Asked Questions About the Million Book Project*, sul sito delle Carnegie Mellon University Libraries: http://www.library.cmu.edu/Libraries/MBP_FAQ.html. Per un

10.561 di essi sono effettivamente accessibili in rete, attraverso il sito dell'Internet Archive[9], in formato DjVu (un formato aperto e fortemente compresso utilizzato per la distribuzione di documenti scannerizzati[10]). L'impressione – almeno quella che si ricava dai materiali fin qui disponibili – è che la qualità di alcune delle scannerizzazioni lasci un po' a desiderare, e che in generale l'avvio dei progetti di digitalizzazione con partner commerciali "forti" abbia determinato una certa perdita di iniziativa e forse anche di entusiasmo per il progetto. È a mio avviso probabile che esso finirà per collegarsi con l'Open Content Alliance, di cui ci occuperemo in seguito, e che il suo risultato più rilevante riguardi comunque la digitalizzazione dei testi non in inglese.

Il valore informativo e commerciale della digitalizzazione bibliotecaria

Le iniziative fin qui considerate non prevedevano, almeno inizialmente, un sostegno rilevante da parte di partner commerciali "forti". Sono bastati però poco più di due anni, fra l'ottobre del 2003[11] e la fine del 2005, a cambiare radicalmente la situazione. In maniera apparentemente improvvisa, la digitalizzazione del patrimonio librario è diventata uno dei terreni di battaglia più combattuti del mondo della *net economy*, con l'intervento massiccio dei colossi del settore, a cominciare da nomi del calibro di Microsoft, Google, Yahoo!, Amazon.

Quali sono i fattori che hanno determinato questa evoluzione, che potrebbe sembrare per certi versi sorprendente?

Per individuarli, occorre allargare la prospettiva. La battaglia in corso attorno alla digitalizzazione libraria, infatti, non è che il capitolo più recente – almeno per ora – di una storia già lunga e di enorme rilievo: quella dei tentativi di integrare l'informazione disponibile in rete e l'informazione disponibile fuori dalla rete.

Per capire l'importanza (ma anche la difficoltà) di questi tentativi, bisogna tener presente che – nonostante l'indubbia centralità assunta da Internet negli ultimi anni – l'informazione disponibile in rete non è che una frazione dell'informazione circolante nella nostra società. La School of Information Management & Systems della University of California a Berkeley ha diffuso nel 2003 la più recente versione di *How Much Information*[12], uno studio non privo di aspetti discutibili, ma che ha il

commento sullo stato di avanzamento del progetto, si veda Walt Crawford, *Discovering Books: The OCA/GBS Saga Continues*, «Cites & Insights», 6 (2006), in rete all'indirizzo http://cites.boisestate.edu/civ6i6.pdf (in formato PDF) e http://cites.boisestate.edu/v6i6a.htm (in formato HTML).

[9] All'indirizzo http://www.archive.org/details/millionbooks.

[10] Per informazioni si veda http://www.djvuzone.org/.

[11] Data in cui è stata introdotta da Amazon la funzionalità Search Inside the Book, sulla quale torneremo in seguito.

[12] Peter Lyman – Hal R. Varian, *How Much Information 2003*, disponibile alla pagina http://www.sims.berkeley.edu/how-much-info-2003.

merito di far percepire le dimensioni del problema: secondo tale ricerca, nel 2002 sono stati prodotti complessivamente 5 exabyte[13] di informazione conservata su un qualche supporto (vi è naturalmente anche informazione "di flusso" che non viene conservata, come le telefonate[14]). Ebbene, di questi 5 exabyte "solo" 170 terabyte – cioè meno di un decimillesimo – fanno parte del cosiddetto *surface Web*, ovvero dell'informazione disponibile su Web e raggiungibile da un motore di ricerca.
Dove si trova tutto il resto dell'informazione che produciamo? La maggior parte (circa il 90%) è sparsa in milioni e milioni di dischi rigidi, ed è fatta di audio, filmati, documenti d'ufficio, lettere (la posta elettronica genera a livello mondiale circa 400.000 terabyte di nuova informazione l'anno, e i sistemi di *instant messaging* generano nello stesso periodo quasi 300.000 terabyte di informazione). Non stupisce dunque che una delle tecnologie "calde" degli ultimissimi anni sia stata quella dei cosiddetti strumenti di *desktop search*, che consentono di integrare ricerche in rete e ricerche sui documenti presenti nel nostro disco rigido. Così come non stupisce che anche in questo campo sia cominciata una battaglia commerciale che – non a caso – vede impegnati molti fra gli stessi protagonisti della "guerra della digitalizzazione libraria": quasi in contemporanea, fra l'ottobre e il dicembre del 2004, sono infatti apparsi Google Desktop[15], Yahoo! Desktop Search[16] e Windows Desktop Search[17], quest'ultimo di casa Microsoft. Un altro settore naturale di possibile allargamento delle capacità dei motori di ricerca è rappresentato dal cosiddetto *deep Web*, che comprende l'intero insieme dei database accessibili attraverso la rete, ai quali si deve – fra l'altro – la realizzazione *on demand* delle pagine Web dinamiche. Si tratta di circa 90.000 terabyte di informazione, e gli sforzi fatti per rendere almeno i più importanti fra questi database direttamente interrogabili attraverso le stesse interfacce utilizzate per la ricerca su Web rappresentano una fetta cospicua, anche se non sempre evidente per il grande pubblico, degli investimenti economici dei grandi motori di ricerca.
Ma arriviamo finalmente alla carta, e alle biblioteche. Secondo la valutazione degli studiosi californiani, una scannerizzazione di buona qualità dell'informazione prodotta nel 2002 e destinata a una fruizione primaria o secondaria su carta richiederebbe circa 1.600 terabyte. Di questi, tuttavia, la grande maggioranza è costituita

[13] L'exabyte è una unità di misura della memoria. Un exabyte equivale a 1024 petabyte, un petabyte equivale 1024 terabyte, un terabyte equivale a 1024 gigabyte.
[14] Sempre secondo *How Much Information 2003*, le telefonate del 2002 – fisse e mobili, su scala mondiale – occuperebbero, se digitalizzate, ben 17,2 exabyte di memoria: più di 30 volte la quantità di informazione in circolazione in rete nello stesso periodo (è facile prevedere che l'esplosione delle connessioni a banda larga cambierà sensibilmente questo rapporto nei prossimi anni). Come è ovvio, queste valutazioni dipendono in maniera essenziale dal tipo di digitalizzazione presa in considerazione: lo studio illustra i criteri metodologici adottati per la valutazione, criteri che sono in linea di massima uniformi per ciascuno dei media esaminati.
[15] http://desktop.google.com/; la prima versione è dell'ottobre 2004.
[16] http://desktop.yahoo.com/; il primo annuncio ufficiale è del dicembre 2004.
[17] http://desktop.msn.com/; anche in questo caso, la prima versione beta compare nel dicembre 2004.

da documenti d'ufficio; in una valutazione puramente quantitativa, alle biblioteche sembrano restare le briciole: circa 39 terabyte di libri, e circa 200 terabyte di quotidiani, riviste e periodici di vario genere. Ci sono tuttavia quattro motivazioni importanti, che rendono queste "briciole" particolarmente appetibili. Da un lato, c'è un'ovvia considerazione qualitativa: pur essendo una parte relativamente piccola dell'informazione che produciamo complessivamente, libri e riviste contengono molta parte dell'informazione "autorevole", quella che ha più valore, e dunque che è più importante poter reperire. In secondo luogo, a differenza di quanto avviene per molta parte dell'informazione cartacea "non pubblicata", ma analogamente a quanto avviene nel caso dell'informazione disponibile su Web, il contenuto di libri e riviste nasce per essere diffuso nel modo più ampio possibile. C'è in questi contenuti, prima ancora di ogni considerazione economica e di gestione dei diritti, una sorta di naturale aspirazione alla reperibilità: potremmo dire che "cercano lettori". In terzo luogo, non sfuggirà che le dimensioni quantitative di questa fetta così preziosa della nostra produzione informativa sono sostanzialmente dello stesso ordine di grandezza del *surface Web* con il quale i motori di ricerca sono già abituati ad avere a che fare. Il compito di digitalizzare milioni e milioni di libri è certo ciclopico, ma chi intende affrontarlo sa già – almeno a livello puramente quantitativo – di disporre di strumenti capaci di lavorare con il volume di dati prodotto anche dal più ambizioso dei progetti di digitalizzazione[18]. Infine, quello dell'editoria su carta stampata è un mercato economicamente importante, e il suo incontro con le tecnologie digitali – già avviato da tempo e destinato ad assumere in futuro un rilievo sempre maggiore – è almeno potenzialmente in grado di generare utili abbastanza appetibili, anche se l'individuazione di strategie in grado di rivelarsi davvero funzionali e redditizie è, come vedremo, tutt'altro che banale.

L'avvio dei progetti commerciali: Amazon

Le considerazioni fin qui svolte[19] aiutano a capire perché il settore della digitalizzazione libraria sia diventato così rilevante anche dal punto di vista dei grandi operatori nel settore dell'informazione in rete. Ma come è avvenuto il loro intervento, e quali problemi sta suscitando, sia dal punto di vista dei rapporti con il mondo delle biblioteche, sia da quello dei rapporti con il mercato editoriale, sia, soprattutto, per quanto riguarda l'evoluzione delle forme della testualità e della lettura?
Come si è già accennato, una prima mossa in questo settore è quella che, nell'ot-

[18] Tutt'altra questione è naturalmente quella dell'adeguatezza qualitativa degli strumenti di ricerca e dei formati di codifica rispetto al tipo di contenuti proprio di libri e riviste: si tratta di un problema centrale, sul quale avrò occasione di tornare brevemente in seguito.
[19] Alcune delle quali riprendono e integrano un mio primo, breve intervento sul progetto di digitalizzazione bibliotecaria avviato da Google, apparso su «AIB Notizie», 2005, n.1, http://www.aib.it/aib/editoria/n17/0501baldacroncaglia.htm.

tobre 2003, ha visto il lancio della funzionalità Search Inside the Book[20] da parte della più importante libreria in rete, Amazon.com. Per offrire questa funzionalità, Amazon ha realizzato una base dati che comprende il testo digitalizzato di oltre 100.000 libri[21]. La ricerca avviene dall'interno del sito di Amazon o attraverso il sito o la *toolbar* del motore di ricerca del gruppo, A9. I testi sono tutti scannerizzati e digitalizzati direttamente da Amazon, che chiede ai propri partner una copia fisica (e non elettronica) del libro. La ricerca può avvenire sia all'interno dell'intero database (che si integra con il catalogo di Amazon), sia all'interno del singolo libro. Nel primo caso, i risultati della ricerca offrono sia le occorrenze del termine nel catalogo (tipicamente, nel titolo dei libri), di norma considerate più rilevanti, sia quelle all'interno dei libri digitalizzati. Nel secondo caso, la visualizzazione dei risultati della ricerca offre qualche riga del contesto (o dei primi contesti) in cui compare il termine cercato. In ognuno dei due casi, è possibile visualizzare a schermo la pagina del libro, che appare con la stessa formattazione del testo a stampa. È prevista la possibilità di muoversi avanti o indietro di una pagina alla volta all'interno della versione digitalizzata del libro, ma solo per un numero limitato di pagine (di norma, è complessivamente visualizzabile un massimo di tre o di cinque pagine).

Nell'agosto del 2005, la funzionalità Search Inside the Book è stata estesa anche alla versione francese di Amazon, con il nome Chercher au Cœur e una base di circa 5.000 libri digitalizzati.

Nella sua prima incarnazione, la funzionalità offerta da Amazon aveva lo scopo primario di permettere al lettore di identificare libri per lui rilevanti e di "saggiarne" il contenuto, con l'ovvio obiettivo di promuoverne l'acquisto. Ma dietro a questo obiettivo primario ve ne erano altri, non meno importanti: il miglioramento della posizione competitiva del motore di ricerca di Amazon rispetto a quelli della concorrenza, attraverso il rilevante valore aggiunto offerto dalla possibilità della ricerca anche all'interno dei libri; la sperimentazione – ritenuta strategicamente importante – di tecnologie di digitalizzazione, visualizzazione e ricerca all'interno di prodotti editoriali nati nel mondo tradizionale dell'editoria cartacea; l'acquisizione di una rilevante base dati testuale utilizzabile, in prospettiva, anche per la vendita di contenuti in formato digitale.

[20] La funzionalità Search Inside the Book sostituisce la precedente Look Inside, che permetteva la sola visualizzazione delle prime pagine del libro, in modalità immagine.

[21] All'avvio, il progetto dichiarava la digitalizzazione di oltre 120.000 libri, per un totale di circa 33 milioni di pagine e con la collaborazione di 190 case editrici. Non sono disponibili dati esatti sul numero di testi aggiunti negli ultimi due anni: il motore di ricerca di casa Amazon, A9 (http://a9.com/), continua a dichiarare a tutt'oggi l'indicizzazione di «oltre 100.000 libri», ma nel novembre 2005, in occasione del lancio dell'iniziativa Amazon Pages di cui parleremo fra breve, Amazon ha affermato (in maniera probabilmente un po' ottimistica) che la copertura della funzionalità Search Inside è pari a circa la metà dei testi disponibili per la vendita: in tal caso, il numero di libri digitalizzati dovrebbe ormai superare di gran lunga i 120mila iniziali.

Non stupisce, dunque, il recente annuncio da parte di Amazon di due nuove iniziative: Amazon Pages e Amazon Upgrade. La prima funzionalità permetterà al lettore l'accesso a pagamento a un numero maggiore di pagine del libro digitalizzato, fino a coprire interi capitoli o, eventualmente, il libro nella sua interezza. La seconda permetterà all'acquirente del libro su carta di acquistare, con un sovrapprezzo, anche l'accesso permanente alla versione digitale dello stesso testo. Un'idea dei costi per l'utente è data dall'annuncio – indipendente ma contemporaneo a quello di Amazon – di un programma di vendita *pay per page* da parte di una delle maggiori case editrici al mondo, Random House: indicativamente, 99 centesimi per ogni 20 pagine per quanto riguarda la vendita di pagine o gruppi di pagine, con un compenso per l'editore pari a circa 4 centesimi per pagina. Per quanto riguarda il programma Amazon Upgrade, le prime indiscrezioni parlano di un sovrapprezzo pari a circa il 10% del prezzo del libro a stampa (o, in altri casi, di una cifra forfetaria di 1,99 dollari) per l'acquisto della versione digitalizzata di un testo in contemporanea con quello della relativa versione a stampa.

Scende in campo Google: la sfida dei grandi numeri e i testi sotto copyright

Se Amazon è stata fra le prime aziende ad avviare programmi intensivi di digitalizzazione libraria con finalità commerciali, Google ha sicuramente la palma del progetto più ambizioso (e più discusso) in quest'ambito. Nato, un po' in sordina, poco dopo il lancio di Search Inside the Book da parte di Amazon, il progetto Google Print ha cominciato a far parlare davvero di sé nel dicembre 2004, con l'annuncio dell'intenzione di procedere alla digitalizzazione di buona parte del patrimonio bibliotecario di cinque grandi biblioteche (quelle delle università di Harvard, Oxford, Stanford e Michigan, assieme alla New York Public Library).

L'annuncio era rivoluzionario anche per il ritmo di crescita previsto: fra i dieci e i quindicimila libri la settimana, con l'obiettivo di digitalizzare complessivamente fra i 10 e i 15 milioni di libri[22]. Ma il progetto si è rivelato assai più complesso del previsto. I problemi incontrati sembrano essere fondamentalmente di tre tipi: 1) quelli, ben noti, legati alle iniziative legali attraverso le quali la Authors Guild e la Association of American Publishers (oltre ad alcuni autori individuali) stanno cercando di bloccare la digitalizzazione dei testi sotto diritti; 2) quelli, purtroppo assai meno noti, legati alle difficoltà tecniche del progetto e alle caratteristiche degli

[22] Le cinque biblioteche che hanno aderito al progetto Google Print possiedono complessivamente – escludendo le sovrapposizioni – circa 10,5 milioni di libri, sui circa 32 milioni di libri registrati complessivamente dal WorldCat nel gennaio 2005; per questo e altri dati rilevanti sul progetto si veda Brian Lavoie – Lynn Sillipigni Connaway – Lorcan Dempsey, *Anatomy of aggregate collections: the example of Google Print for Libraries*, «D-Lib Magazine», 11:9 (September 2005), www.dlib.org/dlib/september05/lavoie/09lavoie.html.

strumenti software e dei formati di codifica utilizzati; 3) quelli, ovviamente influenzati anche dalle prime due tipologie, legati alla focalizzazione dell'esatta fisionomia del progetto stesso.

La prima categoria di difficoltà è quella che ha suscitato il dibattito maggiore[23]. Secondo Google, la semplice digitalizzazione di una collezione bibliotecaria (inclusi i testi ancora sotto diritti) a scopi di indicizzazione rientra nel *fair use* previsto dalle normative americane e da molte normative internazionali in materia di diritto d'autore, e non richiede un consenso preventivo da parte dei detentori dei diritti. Una violazione del copyright vi sarebbe solo se i testi così digitalizzati venissero inseriti in rete con modalità tali da consentirne la lettura da parte degli utenti. Ma, afferma Google, il suo progetto non ha né questo scopo né questo effetto. Non ha questo scopo, perché l'obiettivo principale è rendere *ricercabile* la base testuale via via costruita, non di rendere i libri *leggibili* gratuitamente via Web. E non ha questo effetto perché le tecnologie usate permettono all'utente di visualizzare solo il contesto del termine ricercato, e non l'intero libro. Non a caso, Google adotta politiche assai restrittive sul numero di "pagine di contesto" che l'utente può visualizzare: di norma cinque o tre, come nel caso di Amazon, per gli editori partner, ma talvolta due, o addirittura meno di una pagina[24], nel caso degli editori con cui non esistono accordi diretti.

Sempre secondo Google il servizio offerto, che prevede link a siti terzi[25] per l'acquisto on line del libro (nella sua edizione cartacea ma anche, quando disponibile, in quella elettronica), è in realtà vantaggioso per gli stessi autori ed editori, promuovendo la circolazione e la vendita dell'opera.

Google offre comunque ai detentori di diritti l'opzione *opt-out*: gli autori e gli editori che desiderassero non veder digitalizzati e indicizzati i libri dei quali detengono i diritti, possono comunicarlo, scegliendo di restare esclusi dal progetto.

Dal punto di vista di Google, dunque, la digitalizzazione a scopo di indicizzazione dei libri sotto diritti è considerata analoga, rispetto ai problemi di copyright, alla memorizzazione a scopo di indicizzazione del contenuto delle pagine e dei siti Web, e cioè a quel che costituisce storicamente la ragion d'essere stessa dell'azienda. Anche in quest'ultimo caso, infatti, si tratta di indicizzare materiali il cui copyright non appartiene a Google, e anche in questo caso il meccanismo è ovviamente *opt-out* e non *opt-in* (un motore di ricerca che fun-

[23] Per una prima – ma già assai ampia – bibliografia al riguardo si veda DigitalKoans, Scholarly Electronic Publishing Weblog, *The Google Print Controversy: a Bibliography*, http://www.escholarlypub.com/digitalkoans/2005/10/25/the-google-print-controversy-a-bibliography/. Alcune indicazioni sull'evoluzione successiva del dibattito sono in Walt Crawford, *Discovering Books: The OCA/GBS Saga Continues* cit. Nel nostro paese, una discussione su questi temi si è svolta nel maggio 2006, nell'ambito del Salone del libro di Torino, in un incontro dal titolo *Google ricerca libri: motore di ricerca o motore di lettura?*

[24] Google parla al riguardo dell'offerta di *snippets*, ovvero di piccole porzioni del testo.

[25] Fra i quali – a temperare in parte la concorrenza fra le due iniziative – compare anche Amazon.

zionasse con meccanismi *opt-in* avrebbe ben scarse possibilità di successo…)[26].
La scelta di Google è stata difesa anche dal punto di vista delle funzioni di conservazione e ricerca proprie delle istituzioni bibliotecarie. Nel febbraio 2006, in un appassionato intervento alla Professional/Scholarly Publishing Division della Association of American Publishers, Mary Sue Coleman, presidente della University of Michigan, ha ad esempio ribadito con forza la partecipazione e il sostegno all'iniziativa da parte della biblioteca dell'università, dichiarando che l'inclusione di materiali sotto copyright – dei quali peraltro si garantisce una gestione separata e rispettosa delle normative – è essenziale per soddisfare due fra le funzioni fondamentali di una biblioteca: quella di conservazione, e quella di favorire la ricerca e il reperimento di informazioni[27].
Le associazioni che hanno impugnato il progetto contestano radicalmente le tesi di Google[28], rifiutando sia l'idea che la digitalizzazione e indicizzazione di un libro possano rientrare nel *fair use*, sia l'idea che la scelta di aderire o meno al progetto da parte di autori ed editori debba essere *opt-out*. Dal loro punto di vista, la digitalizzazione di un testo sotto diritti equivale comunque a una violazione del copyright, e in ogni caso l'eventuale scelta di adesione al progetto stesso da parte dei detentori dei diritti dovrebbe essere esplicita e preventiva: *opt-in* anziché *opt-out*. Inoltre, nell'ipotesi considerata, la "sicurezza" del testo digitalizzato e la sua protezione da una diffusione non autorizzata – che lo porterebbe rapidamente nei circuiti *peer-to-peer*, già ricchi delle versioni elettroniche "pirata" di moltissimi li-

[26] Non stupisce che, in maniera coerente ma paradossale, alcuni fra i commentatori che hanno avanzato i maggiori dubbi sulla legalità dell'iniziativa di Google sostengano che anche la semplice funzione di indicizzazione di siti Web, se effettuata (come avviene di norma) attraverso meccanismi *opt-out* e non *opt-in* sia in contrasto con la normativa sul copyright. Si vedano al riguardo le dichiarazioni di Sally Morris, citate in Walt Crawford, *OCA and GLP 2: Steps on the Digitization Road*, «Cites & Insights», 5, Number 14, December 2005, http://cites.boisestate.edu/v5i14d.htm. L'intervento di Crawford contiene uno dei panorami più ampi e articolati ad oggi disponibili su commenti e prese di posizione relativi alle implicazioni legali della iniziativa di Google. Fra gli interventi usciti nei mesi seguenti, oltre a quelli che avremo occasione di menzionare in seguito, merita di essere ricordato quello, lungo e articolato, che Corel Doctorow ha pubblicato il 14 febbraio 2006 nel suo Weblog Boingboing, sotto il significativo titolo *Why Publishing Should Send Fruit-Baskets to Google* (http://www.boingboing.net/2006/02/14/why_publishing_shoul.html).

[27] Mary Sue Coleman, *Google, the Khmer Rouge and the Public Good*, in rete all'indirizzo http://www.umich.edu/pres/speeches/060206google.html. La Coleman dichiara fra l'altro: «We are allowing Google to scan all of our books – those in the public domain and those still in copyright – and they provide our library with a digital copy. We insisted on this for one very important reason: our library must be able to do what great research libraries do – make it possible to discover knowledge. The archive copy achieves that. This copy is entirely, and only, for preservation and research». L'intervento ha suscitato un vivace dibattito in ambito bibliotecario: cfr. ad es. gli interventi di Siva Vaidhyanathan, http://www.nyu.edu/classes/siva/archives/002762.html, Michael Madison,http://madisonian.net/archives/2006/02/06/michigan-prez-on-google-book-search/, e Walt Crawford, *Discovering Books: The OCA/GBS Saga Continues* cit.

[28] I testi dei due *complaints* legali sollevati contro Google nel settembre e nell'ottobre 2005 sono disponibili in rete agli indirizzi http://www.groklaw.net/pdf/Google_Complaint.pdf e http://www.publishers.org/press/pdf/40%20McGraw-Hill%20v.%20Google.pdf.

bri – dipenderebbe interamente da Google. Nonostante l'indubbia affidabilità dell'azienda americana, come escludere eventuali problemi, anche involontari, legati all'esistenza sui server di Google di testi digitalizzati assai facili da distribuire in rete in maniera incontrollata[29]?

A meno di accordi extragiudiziali – che al momento non sembrano essere in vista ma che non sono da escludersi, visto che da un lato gli editori "rischiano" molto, in termini di visibilità, da un'esclusione dalle basi dati di quello che è attualmente il più importante strumento di ricerca su Web, e che dall'altro Google non ha certo interesse a favorire una fuga in massa di autori ed editori verso progetti concorrenti – il braccio di ferro finirà in tribunale, e le relative cause giudiziarie avranno conseguenze di enorme portata per la normativa sulla tutela dei diritti d'autore e di copia nell'era digitale[30]. Dal canto suo, dopo un'iniziale stop all'iniziativa per permettere a chi lo desiderasse di utilizzare l'opzione *opt-out*, Google ha dichiarato di voler andare avanti per la sua strada.

Come si accennava, la seconda tipologia di problemi legati al progetto di Google è invece di natura tecnica. Per evitare la possibilità che gli utenti sviluppino strumenti software capaci di "catturare" il testo delle pagine visualizzate, Google (come già Amazon) basa la sua indicizzazione sul collegamento fra due oggetti di natura diversa: l'immagine di una pagina, visualizzata con qualità relativamente bassa[31], e il relativo testo digitalizzato, sul quale avviene la ricerca ma che non viene in alcun modo reso disponibile all'utente. Ovviamente, il fatto che per l'utente il testo elettronico e l'immagine della pagina a stampa (e dunque il supporto per la ricerca e quello per la lettura) appaiano come due oggetti separati costituisce un grave limite, soprattutto quando l'opera è fuori diritti e dunque in linea di principio non ci sarebbero problemi nell'evitare questa "scissione" e nel fornire il testo elettronico anche in forma autosufficiente. Inoltre, per velocizzare l'acquisizione dei testi, la qualità della digitalizzazione sembra essere stata assai sacrificata[32].

[29] Questo problema, va detto, riguarda tutte le iniziative di digitalizzazione, anche a scopi commerciali.

[30] Per rendersene conto può essere utile consultare due interessanti interventi sulla portata giuridica della controversia: Robin Jeweler, *The Google Book Search Project: is Online Indexing a Fair Use under Copyright Law?*, in: *CRS Report for Congress, December 28, 2005*, disponibile all'indirizzo http://fpc.state.gov/documents/organization/59028.pdf, e Jonathan Band, *The Google Library Project: The copyright debate*, American Library Association, Office for Information Technology Policy, January 2006, disponibile all'indirizzo http://www.ala.org/ala/washoff/oitp/googlepaprfnl.pdf.

[31] Nel dibattito sviluppatosi su questo tema, si parla spesso a questo riguardo della scelta esplicita – sia da parte di Amazon, sia da parte di Google – di fornire una qualità di accesso digitale *non optimal*.

[32] Si vedano al riguardo i commenti di Michael Hart in VatulBlog, 14 giugno 2005, http://vatul.net/blog/581/, e quelli di James Jacobs nel suo blog Digilet, 16 febbraio 2006, http://gort.ucsd.edu/mtdocs/archives/diglet/007170.html. In particolare, Jacobs riporta alcune dichiarazioni di Daniel Clancy, direttore del progetto Google Book Search, secondo cui «Google was NOT going for archival quality (indeed COULD not) in their scans and were ok with skipped pages, missing content and less than perfect OCR» . Walt Crawford, *Discovering Books: The OCA/GBS*

Il motore di ricerca di Google, che usa tecniche *fuzzy* per includere i risultati rilevanti anche in presenza di alcuni tipi di errori tipografici, può in parte rimediare a questo problema, ma la realtà è che la qualità degli indici costruiti è difficilmente verificabile, e sembra comunque essere lontana dagli standard accettabili nel mondo accademico e in quello del *reference*.

Va aggiunto anche che poco si sa degli standard di codifica adottati da Google: in una situazione in cui il mondo accademico e della ricerca sta attivamente lavorando a progetti di digitalizzazione testuale sulla base di uno standard XML aperto e sviluppato in maniera collaborativa, lo standard TEI (Text Encoding Initiative)[33], la scelta di utilizzare standard diversi rende in linea di principio poco utilizzabili per scopi di ricerca scientifica o di analisi testuale testi digitalizzati utilizzando standard diversi, a meno di lavori sempre piuttosto onerosi di conversione[34].

Quanto alla terza tipologia di difficoltà, è indubbio che – anche a causa delle vicende legali sopra ricordate, nonché delle mosse dei suoi principali concorrenti – il progetto di Google ha in parte modificato la sua fisionomia col passare del tempo. Progressivamente, si sono venuti a differenziare all'interno dell'iniziativa due obiettivi in parte diversi.

Il primo obiettivo è legato alla digitalizzazione (e all'inclusione nel database del motore di ricerca) di libri in commercio. In questo caso le finalità sono, almeno inizialmente, assai simili a quelle di Amazon, a cominciare dalla identificazione di libri pertinenti e dalla promozione della vendita del libro (su carta o in versione digitale). Rispetto ad Amazon e agli altri modelli concorrenti, Google sembra tuttavia più attenta a favorire forme di circolazione del testo – anche sotto diritti – relativamente più libere, con la previsione di modalità di "prestito a pagamento", e cioè di accesso a pagamento per periodi di tempo predeterminati alla versione digitale del libro[35]. Va ricordato però, prima di entusiasmarsi per questo modello, che l'accesso avverrebbe comunque solo alla "versione immagine" protetta del libro, e sempre – programmaticamente – attraverso forme di "leggibilità non ottimale". Vi sarebbero dunque seri limiti a un uso attivo del libro ottenuto "in prestito". Inoltre, la sostanziale trasformazione del prestito in noleggio prefigura comunque

Saga Continues cit., osserva giustamente che questo rende abbastanza difficile considerare il progetto di Google anche come un valido strumento di conservazione. E si può aggiungere che un'analoga preoccupazione riguarda a questo punto le sue funzionalità di strumento di ricerca davvero affidabile.

[33] Per una introduzione allo standard TEI si veda Lou Burnard – C.M. Sperberg-McQueen, *Il manuale TEI LITE: introduzione alla codifica elettronica dei testi letterari*, a cura di F. Ciotti, Milano: Silvestre Bonnard, 2005.

[34] Anche se gli strumenti di codifica testuale utilizzati da Google sono, come è assai probabile, basati su XML, una conversione automatica sarebbe possibile solo sulla base delle caratteristiche del testo che Google ha effettivamente scelto di marcare; una scelta sulla quale la comunità della ricerca non ha evidentemente avuto voce in capitolo, e che al momento non è comunque stata esplicitata.

[35] L'ipotesi avanzata, in questo caso, è quella del pagamento del 10% del prezzo del libro per un accesso della durata di una settimana.

un rapporto del lettore con il testo nel quale la rapidità della fruizione si traduce in costi minori: in tal modo, il rapporto con il libro rischia di avvicinarsi pericolosamente a quello che abbiamo con la videocassetta noleggiata da Blockbuster. Infine, sarebbe ingenuo ritenere che la prefigurazione di questo tipo di meccanismi di noleggio sia alternativa rispetto all'esplorazione, anche da parte di Google, di forme di vendita *pay per page* o *pay per view*, qualora queste si rivelassero rispondenti a una effettiva richiesta del mercato, ed economicamente redditizie.
In ogni caso, la parte del progetto legata alla promozione attiva (e non alla semplice reperibilità e alla promozione passiva) di testi in commercio sembra essere sviluppata da Google soprattutto attraverso gli accordi di partnership con le case editrici e addirittura con i singoli autori[36], e dunque con modalità *opt-in*. E va sottolineato che in Italia – così come nei diversi altri paesi europei non anglofoni in cui sono state avviate edizioni "nazionali" del progetto – è proprio questo tipo di digitalizzazione "commerciale" che sembra essere attualmente privilegiato rispetto a quella più strettamente "bibliotecaria"[37].
Il secondo obiettivo individuabile nell'iniziativa di Google è quello più direttamente collegato al progetto avviato nel dicembre 2004, che prevede la digitalizzazione di un vasto patrimonio bibliotecario e la collaborazione diretta con le biblioteche (anziché con le case editrici). In questo caso, lo scopo primario sembra essere quello di rafforzare la leadership di Google come strumento di ricerca, e in particolare come strumento per il reperimento di informazioni qualificate e "validate". Un rafforzamento che ha però fra le sue conseguenze anche una certa "chiusura" della base dati testuale acquisita da Google, base dati che non sarebbe aperta all'indicizzazione da parte di altri motori di ricerca concorrenti.
Le due prospettive hanno evidentemente una componente di sovrapposizione, nei casi – ovviamente assai frequenti – in cui la digitalizzazione riguardi testi presenti nel patrimonio delle biblioteche coinvolte dall'iniziativa ma disponibili anche commercialmente. Ciononostante, restano comunque diverse, e l'interesse di Google – anche davanti alla battaglia legale che si appresta ad affrontare – sembra essere quello di differenziarle, e comunque di accentuare il carattere di strumento di ricerca più che quello di strumento di lettura attribuito complessivamente al proprio progetto di digitalizzazione. Non a caso, dunque, l'azienda statunitense ha prima battezzato Google Library la componente del progetto legata alla digitalizzazione bibliotecaria (differenziandola in tal modo dal progetto proposto agli editori), e poi, nel novembre 2005, ha ribattezzato Google Book Search l'intera iniziativa Google Print. Contemporaneamente, Google ha anche annunciato una partnership

[36] Si veda al riguardo la presentazione disponibile alla pagina http://www.google.com/services/print_tour/.
[37] La casa editrice italiana più attivamente coinvolta dal progetto di digitalizzazione di Google sembra essere al momento Feltrinelli. Finora Google ha lanciato progetti nazionali, oltre che in Italia, in Austria, Belgio, Francia, Germania, Paesi Bassi, Spagna, Svizzera.

con la Library of Congress in un progetto denominato World Digital Library, che – sotto il patrocinio dell'UNESCO – prevede la digitalizzazione di testi e materiali (non necessariamente a stampa) rappresentativi di tutte le culture del mondo[38]. La World Digital Library dovrebbe svilupparsi sul modello del ben noto progetto American Memory[39], ricco attualmente di oltre 10 milioni di documenti digitalizzati, allargandone l'ambito di copertura alla storia e alle culture di altri paesi.
La World Digital Library rappresenta evidentemente una iniziativa distinta, per natura e finalità, da Google Library, e la partecipazione finanziaria di Google (con una donazione di 3 milioni di dollari) sembra confermare come al momento, in una situazione in rapida evoluzione e che vede il continuo riposizionamento dei maggiori operatori del settore, l'obiettivo principale di Google sia quello di allargare al massimo le proprie "teste di ponte", in attesa di individuare quali iniziative e quali politiche possano rivelarsi vincenti. In particolare, si può pensare (o sperare) che Google consideri la World Digital Library come un possibile – e autorevole – laboratorio di studio dei formati di codifica, un aspetto del processo di digitalizzazione sul quale l'impegno di Google sembra fin qui carente, almeno a giudicare dalle informazioni disponibili all'esterno.

L'Open Content Alliance

La terza iniziativa alla quale occorre far cenno è quella che vede l'alleanza di colossi del calibro di Microsoft, Yahoo!, HP e Adobe, riuniti in questo caso in un progetto che almeno nelle sue finalità dichiarate è non-commerciale e molto *politically correct* sia nella scelta del nome (Open Content Alliance[40]) sia nella partnership con una fra le più interessanti realtà no-profit della rete, l'Internet Archive[41] fondato nel 1996 da Brewster Kahle, nonché con numerosi archivi e biblioteche (inclusa la British Library e molte biblioteche universitarie). Al progetto aderiscono anche organizzazioni professionali e consorzi, come il Research Libraries Group, che collaborerà in particolare nel settore dei metadati. L'obiettivo dichiarato è quello di portare avanti una vasta iniziativa di digitalizzazione libraria, centrata soprattutto sulla acquisizione di testi fuori diritti. L'attenzione primaria è dunque verso il mondo bibliotecario – al quale viene proposta una iniziativa priva di particolari rischi legali – e non verso quello editoriale. Inoltre, a differenza di quanto accade nel caso del progetto di Google, la base testuale acquisita sarebbe aperta anche a indicizzazioni esterne, e dunque più estesamente fruibile.
I testi digitalizzati, conservati dall'Internet Archive, entreranno a far parte delle collezioni delle biblioteche partecipanti ma dovrebbero essere anche raccolti in bi-

[38] Il comunicato stampa congiunto distribuito da Google e dalla Library of Congress è consultabile all'indirizzo http://www.loc.gov/today/pr/2005/05-250.html.
[39] http://www.loc.gov/memory.
[40] http://www.opencontentalliance.org/.
[41] http://www.archive.org/.

blioteche digitali autonome, quale quella già inaugurata, con fine sperimentale, dallo stesso Internet Archive, e denominata Open Library[42].

Per ovvi motivi, legati alla natura non direttamente commerciale dell'iniziativa, l'Open Content Alliance ha fornito molte più informazioni sui dettagli tecnici del progetto di quanto non abbiano fatto Amazon o Google: la digitalizzazione dei libri avviene utilizzando un sistema denominato Scribe in grado di garantire a costi assai contenuti[43] una qualità elevatissima, attraverso l'uso di immagini a colori e con una risoluzione di 500 dpi (basti pensare che l'immagine non compressa di ogni singola pagina pesa attorno ai 20 Mb, e l'archiviazione in formato immagine di un singolo libro di circa 300 pagine richiede quindi circa 6 Gb)[44]. Quanto al formato di archiviazione dei testi, la presenza di Adobe nell'alleanza, e gli esempi fin qui disponibili, suggeriscono una particolare attenzione verso l'uso di PDF (un formato che nelle sue versioni più recenti è comunque integrabile con l'uso di XML nella gestione dei metadati). Un servizio di *print on demand*, fornito da Lulu.com (uno dei più attivi operatori nel campo del *print on demand* via rete), dovrebbe permettere di ordinare versioni a stampa dei libri fuori diritti scannerizzati dal progetto, per il prezzo indicativo di circa 8 dollari a libro[45]. L'Internet Archive conserverà due copie indipendenti di tutti i materiali digitalizzati, ad Amsterdam e in Egitto, e garantirà la migrazione dei supporti approssimativamente ogni tre anni.

Fra gli aspetti più interessanti dell'iniziativa OCA sono le linee programmatiche relative alla gestione dei metadati: «Metadata for all content in the OCA will be freely exposed to the public through formats such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and RSS»[46].

La scelta del formato OAI per l'esposizione dei metadati ne permetterà la raccolta (*harvesting*) da parte di servizi esterni, e costituisce indubbiamente una scelta felice nella prospettiva di una integrazione, almeno a livello di metadati, con iniziative diverse, mentre l'uso di RSS garantisce una buona visibilità alle "nuove accessioni", anche da parte dei singoli utenti.

[42] http://www.openlibrary.org. Il sito dell'Open Library offre già una curiosa interfaccia Web per "sfogliare" i primi libri elettronici acquisiti dal progetto.

[43] Nel comunicato stampa relativo alla sua adesione all'iniziativa, la California Digital Library dichiara – con una valutazione forse un po' ottimistica – un costo di appena 10 centesimi a pagina. Cfr. http://www.cdlib.org/news/press_releases/oca_release_final_20050930.doc.

[44] Per queste informazioni, si veda http://www.openlibrary.org/details/openlibrary. Una interessante descrizione del funzionamento di Scribe e dello svolgimento pratico del lavoro di digitalizzazione è in Jeffrey Young, *Scribes of the Digital Era*, «The Chronicle of Higher Education», 2006, January 27, in rete all'indirizzo http://chronicle.com/temp/reprint.php?id=ksh9jjbtz9yf5ttwl678ngdr80sx01kk. Dall'articolo si ricava fra l'altro l'informazione che una postazione Scribe è in grado di digitalizzare circa 500 pagine l'ora.

[45] L'interesse della Open Content Alliance per la stampa *on demand* non stupisce: Brewster Kahle, che come si è accennato è il "padre fondatore" dell'Internet Archive, ha lanciato anche la cosiddetta *Internet Bookmobile*, una postazione mobile di *print on demand* a bassissimo costo. Cfr. http://www.archive.org/texts/bookmobile.php.

[46] Dalle *FAQ* della Open Content Alliance: http://www.opencontentalliance.org/faq.html.

Nel complesso, come si è visto, il progetto di digitalizzazione portato avanti dalla Open Content Alliance ha aspetti sicuramente innovativi, in particolare nell'attenzione per l'accesso aperto ai contenuti fuori diritti e ai metadati, nonché per quella verso i cosiddetti *orphan books*, opere teoricamente sotto diritti ma per le quali – a causa della scomparsa delle relative case editrici, o dell'irreperibilità degli autori – non sia possibile individuare a costi accettabili il detentore dei diritti stessi[47].

Inoltre, il numero di istituzioni bibliotecarie che aderiscono al progetto (34 nel gennaio 2006) è molto più alto di quelle coinvolte dall'iniziativa di Google, e questo caratterizza indubbiamente l'iniziativa come uno dei più ampi ed interessanti progetti di collaborazione in ambito bibliotecario legato ad attività di digitalizzazione[48]. La sua forza di attrazione è del resto testimoniata dal fatto che anche nuove iniziative di digitalizzazione bibliotecaria, come l'Alouette Canada Project – creato nel novembre 2005 e che vede l'adesione di 27 biblioteche di ricerca canadesi, con l'obiettivo di digitalizzare nei prossimi anni fra i 3 e i 4 milioni di volumi – hanno immediatamente dichiarato la volontà di collaborare con l'Open Content Alliance[49].

Resta da vedere quanto l'indubbia funzione anti-Google con la quale i principali partner aziendali hanno aderito all'iniziativa ne condizionerà lo sviluppo, anche tenendo presente che Google potrebbe sempre, con una mossa a sorpresa, decidere di partecipare anch'essa in qualche forma al progetto[50]. D'altro canto, anche i partecipanti "commerciali" alla Open Content Alliance possono avere (e di fatto hanno) progetti di digitalizzazione autonomi relativi a contenuti sotto diritti; così, ad esempio, la Microsoft, con il progetto Microsoft Book Search, si prepara indubbiamente a indicizzare testi fuori diritti in collaborazione con la Open Content Alliance, ma si prepara anche a digitalizzare testi sotto diritti (pur se, in contrasto con Google, con una politica rigorosamente *opt-in*) e ad offrire forme di accesso a pagamento a tali collezioni, con meccanismi che possiamo ipotizzare basati sugli stessi strumenti (*pay per page*, *pay per book*, *pay per access time*…) utilizzati dai suoi concorrenti. È possibile che nei prossimi mesi una funzionalità Book Search compaia all'interno delle opzioni di ricerca offerte da Windows Live, il nuovo servizio unificato di ricerca di casa Microsoft. Del resto, a partire dall'aprile 2006 all'interno di Windows Live è già incluso Live

[47] L'ufficio copyright della Library of Congress ha avviato un'iniziativa specifica sul tema degli *orphan works*: si veda al riguardo http://www.copyright.gov/orphan/, e, per un commento, http://weblog.ipcentral.info/archives/2005/10/orphan_books.html.

[48] Questo aspetto è sottolineato in Jeffrey Young, *Scribes of the Digital Era* cit.

[49] Il sito Web dell'iniziativa, all'indirizzo http://www.alouettecanada.ca, riporta ulteriori informazioni sul progetto. Cfr. anche Walt Crawford, *Discovering Books* cit.

[50] Personalmente, ritengo che Google sia trattenuta dal compiere questa mossa – che rimescolerebbe ulteriormente le carte di una partita senz'altro complicata – solo dalla minore qualità che probabilmente caratterizza le sue digitalizzazioni.

Academic, uno strumento dalle caratteristiche piuttosto interessanti che permette la ricerca su contenuti e dati bibliografici validati provenienti dal mondo accademico e della ricerca scientifica[51].

E l'Europa? La risposta francese e la biblioteca digitale europea

I progetti di digitalizzazione bibliotecaria avviati negli ultimissimi anni da parte dei colossi della ricerca e gestione dell'informazione in rete, e che abbiamo fin qui considerato, hanno indubbiamente messo un po' fra parentesi i progetti istituzionali sorti sotto responsabilità governative o di enti di ricerca. Negli Stati Uniti, come abbiamo visto, molti fra tali progetti sono ormai in un modo o nell'altro collegati alle grandi iniziative a finanziamento prevalentemente privato. Una reazione contro questa tendenza si è invece manifestata in Europa, attraverso la proposta francese della creazione di una biblioteca digitale europea[52]. La sua prima formulazione risale al gennaio 2005, quando Jean-Noël Jeanneney, Presidente della Bibliothèque Nationale de France (BNF), ha pubblicato su «Le Monde» un articolo dal titolo significativo: *Quand Google défie l'Europe*[53]. Secondo Jeanneney, l'idea che i progetti di digitalizzazione bibliotecaria si sviluppino su scala globale sotto la guida di Google rischia di portare di fatto a un ulteriore rafforzamento del predominio linguistico e culturale dell'inglese e degli Stati Uniti nella "immagine del mondo" fornita dalla rete. La proposta di una iniziativa autonoma europea nel campo delle biblioteche digitali è stata autorevolmente ripresa nel marzo dello stesso anno dal presidente Chirac, che ha dichiarato: «Un vaste mouvement de numérisation des savoirs est engagé à travers le monde. Riches d'un exceptionnel patrimoine culturel, la France et l'Europe doivent y prendre une part déterminante. Il s'agit d'un enjeu fondamental pour la diffusion des connaissances et la valorisation de la diversité culturelle»[54] .

[51] Microsoft Live Academy è in evidente concorrenza con Google Scholar. In un intervento sul sito SearchEngineWatch (*Microsoft Launches Windows Live Academic Search*, 12 aprile 2006, http://searchenginewatch.com/searchday/article.php/3598376), Chris Sherman osserva però che «unlike Google Scholar, which crawls the web for academic content, Windows Live Academic Search works closely with publishers and uses structured feeds to build its index. As such, all content accessed through the service comes directly from a trusted source – namely, the publisher of a scholarly journal». Va detto, peraltro, che anche Google Scholar integra contenuti strutturati e non si basa unicamente su risultati ricavati dal *surface Web*.

[52] Una risorsa preziosa per ricostruire la storia della proposta di biblioteca digitale europea e per mantenersi aggiornati sulla sua evoluzione è la pagina Dossier Bibliothèque numérique, offerta dal sito francese Formats-ouvert all'indirizzo http://formats-ouverts.org/blog/2005/09/15/536-dossier-bibliotheque-numerique-europeenne.

[53] «Le Monde», 22 janvier 2005; l'articolo è disponibile in rete all'indirizzo http://www.bnf.fr/pages/dernmin/pdf/articles/lemonde_2401.pdf. Per una esposizione più ampia della stessa tesi si veda Jean-Noël Jeanneney, *Quand Google défie l'Europe, plaidoyer pour un sursaut*, Paris: Éditions Fayard, 2005.

[54] Dichiarazione del 16 marzo 2005, disponibile in rete all'indirizzo http://elysee.fr/elysee/francais/salle_de_presse/communiques_de_la_presidence/2005/mars/communique_sur_l_accessibilite_des_bibliotheques_de_france_et_d_europe_sur_internet.28865.html.

A fine aprile, su iniziativa della BNF, diciannove biblioteche europee hanno firmato una mozione comune con lo scopo di «appuyer une initiative commune des dirigeants de l'Europe visant à une numérisation large et organisée des oeuvres appartenant au patrimoine de notre continent»[55]. Tra i firmatari ci sono anche i rappresentanti italiani, assieme a quelli francesi, tedeschi, francesi, spagnoli… ma la British Library, pur dichiarando di appoggiare l'iniziativa, non ha firmato la mozione: come abbiamo visto, pochi mesi dopo la troveremo fra i partner di Microsoft e Yahoo! all'interno della Open Content Alliance[56].

Indubbiamente la scelta della British Library – esplicitamente criticata da Jeanneney[57] – rappresenta un duro colpo al tentativo di realizzare in Europa progetti pubblici di digitalizzazione completamente autonomi rispetto a quelli promossi dai grandi protagonisti della battaglia per il controllo dell'informazione in rete. La proposta di una biblioteca digitale europea comunque va avanti: nel settembre 2005 ha avuto l'appoggio della Commissione europea[58], e nel marzo 2006 la stessa Commissione europea ha delineato le tappe per la sua costituzione. La biblioteca digitale europea sarà basata su un piano di collaborazione fra le biblioteche nazionali dei paesi dell'Unione, le cui caratteristiche dovranno essere definite entro fine 2006. Gli obiettivi – piuttosto ambiziosi – sono quelli di digitalizzare e rendere disponibili via rete 2 milioni di documenti («libri, film, fotografie, manoscritti e altre opere») entro il 2008, e 6 milioni di documenti entro il 2010[59]. Il comitato di venti esperti chiamato a seguire lo svolgimento dell'iniziativa comprende, per l'Italia, Paolo Galluzzi e Marco Ricolfi[60].

La European Digital Library sarà strettamente collegata al preesistente progetto TEL (The European Library), che aveva lo scopo di dare accesso alle risorse combinate delle biblioteche nazionali dei paesi europei e che disponeva già di un proprio portale con un motore di ricerca unificato[61]. Il primo passo per la sua costituzione consisterà probabilmente in una più stretta integrazione dei progetti di digitalizzazione

[55] http://www.lemonde.fr/web/article/0,1-0,36-643536,0.html.

[56] La British Library collabora comunque anche con Google: dal marzo 2006, il motore di ricerca Google Scholar – specializzato in contenuti validati provenienti dal mondo accademico e della ricerca scientifica – si è infatti esteso ai contenuti indicizzati dal servizio di *desktop document delivery* della British Library, denominato British Library Direct: cfr. il comunicato stampa della British Library disponibile all'indirizzo http://www.bl.uk/news/2006/pressrelease20060302.html.

[57] Il 4 novembre 2005 il direttore della BNF ha rilasciato una dichiarazione in cui, pur riconoscendo come la nascita di iniziative concorrenti rispetto a quella di Google riduca il rischio di scelte monopoliste, si deplora la scelta della British Library «de faire affaire en solo, d'un bord à l'autre de l'Atlantique, en solidarité anglo-saxonne, avec une grande entreprise commerciale américaine». Cfr. www.bnf.fr/pages/presse/communiques/british_library.pdf.

[58] http://europa.eu.int/rapid/pressReleasesAction.do?reference=IP/05/1202&format=HTML&aged=0&language=FR.

[59] La versione italiana del comunicato è disponibile all'indirizzo http://www.europa.eu.int/rapid/pressReleasesAction.do?reference=IP/06/253&format=HTML&aged=0&language=IT&guiLanguage=it.

[60] La lista completa degli esperti è disponibile all'indirizzo http://europa.eu.int/information_society/activities/digital_libraries/high_level_expert_group/index_en.htm.

[61] http://www.theeuropeanlibrary.org.

già avviati su scala nazionale (a cominciare da Gallica) e delle relative collezioni. La newsletter di TEL dell'aprile 2006[62] annuncia l'intenzione di realizzare un accesso unificato a tutte queste collezioni entro 18 mesi, nell'ambito di un progetto finanziato dall'iniziativa europea eContentPlus. Una integrazione parziale è del resto già offerta dalla *virtual collection* «On line books, images, maps, music…», disponibile attraverso TEL come sottoinsieme dell'European Library Harvest. È interessante notare che la raccolta e l'integrazione dei metadati è realizzata attraverso l'uso del protocollo OAI, che come abbiamo già ricordato è utilizzato anche dalla Open Content Alliance: la standardizzazione dei formati di esposizione e di raccolta dei metadati permette in prospettiva l'integrazione dei metadati di progetti diversi.

Conclusioni

Davanti al moltiplicarsi delle iniziative di digitalizzazione bibliotecaria, del quale abbiamo fin qui cercato di proporre un pur incompleto panorama, e davanti al continuo riposizionamento dei soggetti coinvolti, alla ricerca delle posizioni più vantaggiose e delle alleanze più prestigiose, ogni bilancio o valutazione rischia di essere prematuro. Tuttavia, qualche considerazione – anche se provvisoria – si impone.

Innanzitutto, non vi è dubbio che negli ultimi due anni l'idea stessa di digitalizzazione bibliotecaria abbia cambiato fisionomia, passando da una fase di sperimentazione pionieristica (frammentaria e in prevalenza *research-driven*) a una fase che potremmo forse azzardarci a definire di "accumulazione originaria", nella quale grandi quantità di "capitale testuale" (sia fuori diritti sia sotto diritti) vengono raccolte nell'ambito di iniziative che – nonostante l'apporto e la collaborazione di prestigiose istituzioni pubbliche – sono guidate e ispirate, o comunque finanziate, in primo luogo dai colossi della *net economy*. L'intervento privato ha indubbiamente impresso una enorme accelerazione al processo di digitalizzazione del nostro patrimonio testuale, ma sarebbe ingenuo pensare che tale accelerazione non comporti, accanto agli indubbi benefici, anche i suoi costi.

Il costo principale è rappresentato dalla mancanza (o dalla presenza solo parziale e limitata) di standard aperti e di scelte ragionate e discusse pubblicamente nel campo dei formati di codifica, in particolare per quanto riguarda la "componente testuale" (ma in certi casi anche per la "componente immagine") del processo di digitalizzazione. È ragionevole ritenere, alla luce delle informazioni fin qui disponibili, che almeno in una prima fase le basi di dati testuali che si stanno raccogliendo non saranno né uniformi dal punto di vista delle politiche di crescita, di codifica, di gestione, di accesso e di conservazione, né – di conseguenza – capaci di garantire una effettiva interoperabilità, al di là di funzioni elementari di metaricerca. E questa non è certo una buona notizia per gli utenti, in particolare nel caso dei

[62] «TheEuropeanLibrary.org Newsletter», April 2006, http://libraries.theeuropeanlibrary.org/newsletter/tel_newsletter_april_2006.pdf.

testi fuori diritti, che potrebbero (e dovrebbero) essere accessibili in forma assai più piena di quella al momento prefigurata. Per ora, comunque, l'attenzione verso le esigenze dell'accesso aperto sembra molto maggiore nel caso della Open Content Alliance che in quello di Google Book Search (anche per quanto riguarda le opere fuori diritti incluse in quest'ultimo progetto).
Occorrerà anche vigilare contro il rischio che la realizzazione guidata da capitali privati di grandi banche dati di testi fuori diritti porti, in qualche stadio, a una rivendicazione di nuovi "diritti secondari" sui testi stessi, che possano in qualche modo limitarne l'accessibilità.
Sul versante positivo, si può dire che la sperimentazione di modelli diversi di "accumulazione testuale" può essere comunque utile. In particolare, l'iniziativa di Google e la sua rivendicazione del *fair use* legato alla digitalizzazione e alla ricerca testuale su libri sotto dritti ha – anche al di là della questione della sua fondatezza legale – l'indubbio merito di offrire una nuova prospettiva alle discussioni sulla gestione dei diritti nell'era del digitale. Al di là del caso specifico e delle sue implicazioni giuridiche, infatti, è difficile non riconoscere che la digitalizzazione su larga scala del nostro patrimonio testuale e la possibilità di accedervi attraverso la rete pongono in termini nuovi il problema della natura e dei limiti – ma anche del riconoscimento e della tutela – del diritto collettivo all'accesso organizzato al sapere. La biblioteca è da tempo uno strumento fondamentale per l'esercizio di questo diritto, e l'impegno diretto delle biblioteche è essenziale per garantire che esso sia tutelato anche nella delicata fase di transizione verso il digitale. Tanto più che – anche se non sempre autori e case editrici sembrano rendersene conto – l'allargamento delle modalità di accesso alla conoscenza, pur entrando in alcuni casi in conflitto con forme troppo rigide di tutela dei diritti d'autore e di copia, costituisce comunque la miglior forma di promozione del libro e della lettura e di crescita del mercato editoriale, sia nel mondo dei libri fisici sia in quello dei libri digitali. Da questo punto di vista, non credo affatto che lo sviluppo dei progetti di digitalizzazione bibliotecaria costituisca un pericolo per autori ed editori: al contrario, ritengo offra loro nuove opportunità. Indipendentemente dall'esistenza o meno di un suo fondamento legale, l'iniziativa giudiziaria portata avanti contro Google dalla Authors Guild e dalla Association of American Publishers è dunque a mio avviso non solo culturalmente ma anche commercialmente miope.
Per un commento più approfondito sia sugli aspetti legali sia (soprattutto) su quelli culturali della battaglia che sembra essersi accesa attorno al futuro della digitalizzazione bibliotecaria, e probabilmente per modificare o rivedere alcune delle osservazioni fin qui svolte, occorrerà comunque aspettare che i primi risultati dei vari progetti di digitalizzazione siano effettivamente disponibili, in una forma più completa e meglio organizzata di quanto non avvenga oggi. Probabilmente, non dovremo aspettare troppo: i prossimi due o tre anni si annunciano, in questo settore, altrettanto ricchi di novità di quelli appena trascorsi.

The paper provides a survey of the history and of the features of book digitization projects. According to the author, library digitization is of special relevance for search engines, since it allows them to extend their scope to highly valuable, validated and potentially marketable contents. Furthermore, despite its specific features, the complexity of large-scale full-text book indexing within library digitization projects is comparable to that of Web indexing, making it an ideal field for the expansion of search engines (and for the next search engines war).
This situation offers to libraries and researches a good opportunity for bridging the gap between the book culture and the digital culture, between printed contents and digital contents. However, special care is needed in order to avoid a risk: the commercial relevance and scope of many book digitization projects might hinder the scientific validity of the results of this huge (and expensive!) effort. Poor digitization quality, lack of proper metadata, proprietary format choices, limited interoperability, too strict copyright constraints, are all factors that may severely hinder the usefulness of the work done.
The paper discusses in some detail the digitization projects connected to Amazon Book Search, to Google Book Search (formerly Google Print), to the Open Content Alliance, and to the European digital library. Special attention is devoted to the copyright debate raised by the Google Book Search opt-out approach to the digitization of copyrighted content.

L'article propose une vue d'ensemble sur l'histoire et les caractéristiques des principaux projets de numérisation des livres avec un regard particulier sur le domaine bibliothécaire. Selon l'auteur, la numérisation bibliothécaire doit son importance aux moteurs de recherche car elle permet de couvrir un champ plus vaste de contenus validés de grande valeur informative et commerciale. De plus, malgré ses caractéristiques et ses exigences spécifiques, la complexité de l'indexation full-text *des textes produits par les projets de numérisation bibliothécaire est comparable à celle de l'indexation du Web, ce qui fait de ce secteur un terrain idéal à l'expansion des moteurs de recherche (ainsi qu'à leur bataille commerciale).*
Cette situation peut offrir aux bibliothèques et aux chercheurs l'extraordinaire opportunité de réduire la distance entre la culture du livre et celle du réseau, entre les contenus imprimés et les contenus numériques. Il faut cependant faire très attention à ce que la nature commerciale de nombreux projets de numérisation bibliothécaire ne limite leur valeur scientifique. Une mauvaise qualité de la numérisation, des métadonnées incomplètes ou absentes, le choix de format propriétaires, une interopérabilité limitée, des politiques de tutelle des copyrights trop restrictives, constituent en effet des facteurs qui pourraient limiter sérieusement l'utilité du travail effectué.
L'article discute en détail des projets de numérisation reliés à Amazon Book Search, à Google Book Search (désormais Google Print), à la Open Content Alliance et au projet de la bibliothèque numérique européenne. Une attention particulière est aussi consacrée aux débats suscités par les choix de Google Book Search de numériser des contenus soumis aux copyrights sans l'accord explicite des détenteurs des droits.