

Dig *Italia*

Anno V, Numero 1 - **2010**

Rivista del digitale nei beni culturali

ICCU-ROMA

Rapport sur la numérisation du patrimoine écrit

Remis par Marc Tessier au Ministre de la culture et de la communication
le 12 janvier 2010

Il testo qui pubblicato, conosciuto come Rapport Tessier, è una traduzione letterale, che «DigItalia» diffonde per favorire la conoscenza di un documento di rilevante interesse nel momento attuale del dibattito sugli accordi delle biblioteche e dei governi europei con Google. Il Rapporto è stato elaborato dal Gruppo di esperti presieduto dal Direttore generale di Video Futur, Marc Tessier. Al Gruppo il Ministro della cultura francese, Frédéric Mitterand, ha affidato ad ottobre del 2009 “la mission” di analizzare e valutare l’opportunità e la fattibilità, dal punto di vista dell’interesse generale, del possibile accordo tra la Biblioteca nazionale di Francia (BnF) e Google.

Rapporto sulla digitalizzazione del patrimonio librario*

Presentato da Marc Tessier al Ministro della cultura e della comunicazione il 12 gennaio 2010

Traduzione di Sonia Beraha per Qualitalia S.R.L.

Indice

Introduzione

- I. Studio preliminare: progressi eterogenei in un ambiente incerto
 - I.1 A che punto sono le biblioteche digitali?
 - I.1.1 Un’idea relativamente datata che ha conosciuto uno sviluppo effettivo solo a partire dal 2004
 - I.1.2 Un’idea che si è concretizzata grazie a sviluppi diversificati
 - I.1.3 Un’idea scaturita dall’evoluzione degli utilizzi
 - I.2 Un ambiente caratterizzato da una marcata incertezza
 - I.2.1 Google si trova in un contesto giuridico complesso
 - I.2.2 Un coordinamento insufficiente degli altri soggetti
- II. Gli accordi con Google: una risposta inadeguata
 - II.1 Una risposta inadeguata alla missione delle biblioteche
 - II.1.1 Il Gruppo di lavoro sulla conservazione
 - II.1.2 Il Gruppo di lavoro sull’accessibilità
 - II.2 L’articolazione tra logica pubblica e logica privata
 - II.2.1 Una considerazione insufficiente degli “atouts” delle biblioteche
- III. Le soluzioni possibili
 - III.1 Uno strumento privilegiato da migliorare: Gallica
 - III.1.1 Aspetti istituzionali
 - III.1.2 Rinforzare la presenza di Gallica e dei suoi contenuti sul Web
 - III.1.3 Migliorare il servizio offerto da Gallica
 - III.2. Le condizioni per un partenariato equilibrato con soggetti privati
 - III.2.1 Obiettivi e condizioni preliminari

* Il testo in francese è consultabile all’indirizzo Web
<http://www.culture.gouv.fr/mcc/content/download/3520/23115/file/Rapport%20sur%20la%20numérisation%20du%20patrimoine%20écrit.pdf>.

- III.2.2 “Un libro in cambio di un libro”: una proposta di partenariato basata sullo scambio dei dati digitalizzati
- III.3. Trovare un nuovo slancio di livello europeo
 - III.3.1 Reciprocità e sinergia nelle azioni delle biblioteche
 - III.3.2 Come far evolvere Europeana
 - III.3.3 Una Carta comune per i partenariati pubblico/privato

Sintesi conclusiva

Allegato 1: Lettera d’incarico

Allegato 2: Lista delle persone intervistate

Allegato 3: Le sfide qualitative della digitalizzazione di massa

Allegato 4: Lista delle biblioteche europee che hanno aderito al programma Google Recherche de livres

Introduzione

I lavori del Gruppo di lavoro sulla digitalizzazione del patrimonio delle biblioteche, presieduta da Marc Tessier, si sono tenuti dal 19 ottobre 2009 al 7 gennaio 2010 (vedi Allegato 1: Lettera d’incarico). Sono state condotte una trentina di consultazioni, con la partecipazione dei rappresentanti di importanti biblioteche straniere.

È stato analizzato il contesto tecnico, economico e giuridico in cui si inseriscono gli accordi ed i progetti di accordo tra la società Google e le biblioteche. Tale analisi è stata condotta in una prospettiva di rinforzo della presenza e dell’accessibilità dei libri in Internet.

Secondo il Gruppo di lavoro, questo obiettivo prioritario richiedeva l’approfondimento di una serie di tematiche, a partire dall’esame delle esistenti piattaforme di divulgazione dei libri digitali ed in particolare quella di Gallica, sviluppata dalla Biblioteca nazionale di Francia (BnF). Sulla base di quest’analisi della situazione esistente, sono state valutate le possibilità di ampliare questa piattaforma e modificarne le

modalità di gestione e le funzionalità, allo scopo di consentire al principale soggetto pubblico francese del settore di intraprendere trattative con i partner privati su una base equilibrata. Il Gruppo di lavoro ha concentrato la sua attenzione sull’asse europeo, non ritenendo possibile la costruzione di un’alternativa credibile ai grandi progetti internazionali su una base prettamente nazionale.

Pertanto, la presente relazione si articola in tre fasi:

- uno studio preliminare delle principali biblioteche digitali, ivi inclusa, nel caso di Google Books, la complessa situazione giuridica della rispettiva società di controllo;
- un’analisi degli accordi in essere tra le biblioteche e Google, che non sembrano rispondere adeguatamente alle missioni delle biblioteche;
- una serie di linee di azione, orientate verso tre direzioni: il cambiamento di scala per quanto concerne la digitalizzazione delle opere e le modalità operative di Gallica; una proposta di partenariato con Google Books basata sullo scambio di file digitalizzati, senza alcun diritto di esclusività sui file scambiati ed, infine, il rilancio di un impulso europeo, rivolto alle altre biblioteche europee ed al portale Europeana.

Queste diverse soluzioni sono riportate in un documento di sintesi.

I. Studio preliminare: progressi eterogenei in un ambiente incerto

I.1 A che punto sono le biblioteche digitali?

I.1.1 Un’idea relativamente datata che ha conosciuto uno sviluppo effettivo solo a partire dal 2004¹

L’idea di digitalizzare i libri per creare delle biblioteche digitali non è recentissima: già nel

¹ Vedere l’articolo di Jean-Michel Salaün, *Bibliothèques numériques et Google Book Search*, «Regards sur l’actualité», n. 316, dicembre 2005.

1971, Michael Hart, studente dell'Università dell'Illinois (Stati Uniti), intraprende la prima iniziativa di biblioteca digitale, il cosiddetto "progetto Gutenberg", avvalendosi di una squadra di volontari per rileggere e verificare l'"ocerezizzazione"² delle opere digitali, tutte di dominio pubblico, ovvero non tutelate da diritti d'autore³. Oggi il sito mette a disposizione più di 100.000 libri tramite una rete di partner, oltre a 30.000 opere scaricabili direttamente dal sito, a titolo gratuito. All'inizio, il progetto era essenzialmente anglofono; dal 2007 in poi sono state incluse anche opere in altre lingue.

Questo progetto ha ispirato la realizzazione o i progetti di realizzazione delle grandi biblioteche digitali, prima tra tutte l'idea di Jacques Attali (all'epoca delle primissime riflessioni sulla creazione di una «très grande bibliothèqu» in Francia) di saltare una tappa e creare direttamente una biblioteca digitale francofona. Questo progetto non sarà realizzato immediatamente, ma la Biblioteca nazionale di Francia (BnF) lancerà nel 1997 la prima versione di Gallica, adottando in principio un approccio selettivo ed una digitalizzazione eseguita esclusivamente in modalità immagine. All'inizio la proposta di Gallica comprendeva 3000 libri in modalità immagine, numero destinato ad aumentare gradualmente (vedere più avanti).

I progetti di realizzazione delle grandi biblioteche digitali sono tornati in auge grazie alle iniziative dei grandi motori di ricerca. Infatti, è nell'interesse dei motori di ricerca sottoporre a *Web-harvesting* il massimo quantitativo possibile di contenuti che vanno ad ampliare la loro base di dati e a migliorare l'efficacia e la pertinenza dei risultati.

Google è stata la prima società a lanciare nel-

l'ottobre 2004 una alquanto contestata piattaforma, Google Print, che nel novembre 2005 prenderà il nome di Google Book Search. L'obiettivo era quello di digitalizzare 15 milioni di opere in dieci anni, avvalendosi principalmente⁴ delle opere conservate nelle cinque principali biblioteche partner del progetto e cioè: la New York Public Library, le biblioteche delle Università di Harvard, Stanford, del Michigan, e la Bodleian Library di Oxford.

In risposta a Google Book Search, che non autorizzava gli altri motori di ricerca ad indicizzare gli elementi presenti sulla sua piattaforma, altri soggetti del settore intraprendono progetti inizialmente simili. Nel dicembre 2006 Microsoft lancia il suo programma di digitalizzazione dei libri, con l'idea di associare il suo motore di ricerca Live Search ad una categoria di servizi, tra cui una piattaforma di libri digitalizzati Live Book Search, sulla base di una serie di partenariati sottoscritti con la British Library, la New York Public Library e, anche in questo caso, alcune biblioteche universitarie americane (Università di Cornell, di Toronto e della California). Ma nel maggio 2008 il progetto viene abbandonato poiché Microsoft opta per una riorganizzazione radicale delle sue attività, con la conseguente separazione del motore di ricerca (Bing e non più Live Search) dalle categorie di servizi Live Search.

Anche Yahoo! tenta di sviluppare le sue attività di digitalizzazione appoggiandosi a Internet Archive, un'organizzazione non profit costituita nell'aprile 1996 con lo scopo di archiviare il Web. Insieme creano l'Open Content Alliance (OCA), che raggruppa numerosi partner (biblioteche delle Università della California e di Toronto, Archivi nazionali britannici, Research

² L'"ocerezizzazione", dall'acronimo inglese OCR (Optical Character Recognition, riconoscimento ottico dei caratteri), indica l'operazione che consiste nella scansione di un libro e poi nell'utilizzo di specifici software che consentono di riconoscere i caratteri di stampa di un documento (lettere, simboli o spazi) e di repertoriare ogni parola. Si tratta di un procedimento fondamentale grazie al quale è possibile condurre ricerche su tutte le parole contenute nel testo (la cosiddetta ricerca "a tutto testo").

³ Conformemente alla legge americana: quindi trattasi di opere pubblicate prima del 1923.

⁴ Ma non solo: sin dall'inizio sono stati stipulati accordi anche con alcuni editori.

Library Group, nonché varie società informatiche). Il sito sperimentale dell'OCA⁵ consente di accedere ad oltre un milione di libri di pubblico dominio, anche in questo caso, in buona parte anglofoni.

Anche l'iniziativa di Google provoca la reazione di alcuni stati per il tramite delle organizzazioni internazionali.

A fronte dell'iniziativa della Francia e di cinque altri stati europei, tra cui la Germania, l'Unione Europea ha avviato nel marzo 2006 la realizzazione della Biblioteca Digitale Europea, nell'ambito della Strategia di Lisbona (iniziativa i2010). Nel 2008 è stato aperto il portale Europea, con un duplice obiettivo: offrire l'accesso gratuito al patrimonio digitale europeo con 10 milioni di documenti online entro il 2011 e proporre, eventualmente, l'accesso a pagamento ai contenuti su cui gli editori partner vantano diritti d'autore.

L'Unesco, da parte sua, ha annunciato nel dicembre 2006 il lancio della World Digital Library, che di fatto è più simile ad una ampia banca dati multilingue di informazioni culturali, molto selettiva, che non ad una biblioteca di libri digitali.

1.1.2 Un'idea che si è concretizzata grazie a sviluppi diversificati

Dall'analisi dei diversi progetti di biblioteche digitali in corso, emerge che i progressi compiuti sono alquanto eterogenei, sia in termini di piattaforme che di soggetti. Il panorama di seguito illustrato non è esaustivo, in quanto incentrato principalmente sui siti contenenti risorse francofone di rilievo⁶.

Google Book Search, oggi Google Books, nome che manterremo nel presente rapporto, è una piattaforma provvista di un motore interno che ospita anche una base di dati. Questo strumento immagazzina ed indicizza il contenuto

dei libri scansionati, trattati ed immagazzinati in formato digitale da Google.

Per quanto concerne l'utilizzo, l'internauta può recarsi nel sito della piattaforma ed eseguire direttamente le proprie ricerche (sempre che stia cercando contenuti provenienti esclusivamente da libri), oppure utilizzare il motore Google, dal quale potrà accedere a risultati composti da pagine Web e da estratti dei libri pertinenti. Il contenuto di Google Books è quindi importante non solo dal punto di vista della piattaforma, ma anche da quello del motore stesso, dal momento che permette di ampliare la base di dati dalla quale effettua le sue ricerche e, conseguentemente, la ricchezza e la pertinenza dei risultati.

Quando viene visualizzato un risultato proveniente dalla base dati di Google Books, l'utente, cliccando sul link, apre un'interfaccia che gli consente di visualizzare livelli di informazioni diversi a seconda dello *status* dell'opera. Per i libri di dominio pubblico, l'opera può essere visualizzata integralmente e scaricata in formato PDF e testo Epub; per le opere coperte da diritti, invece, a seconda che siano stati stipulati accordi tra Google e gli editori o meno, l'utente può leggere qualche pagina dell'opera e cliccare sul link che rimanda al sito dell'editore (nel caso di editori partner), oppure può accedere solo ai riferimenti dell'opera, eventualmente accompagnati da brevi estratti (i cosiddetti "*snippets*"), nel caso degli editori che non hanno firmato alcun accordo. In ogni caso, i dati vengono sempre visualizzati con link che rimandano a siti di librerie e biblioteche, sulla sinistra della schermata.

Il sito si avvale essenzialmente di due fonti. La prima è rappresentata dalle biblioteche che hanno sottoscritto accordi di digitalizzazione generalmente relativi a libri non coperti da diritti. Ma Google è riuscita a digitalizzare, tramite il patrimonio delle grandi biblioteche ameri-

⁵ In Europa, la parte del sito che consente l'accesso alle opere è disponibile solo in versione beta.

⁶ L'allegato 3 propone un confronto dettagliato delle funzionalità offerte da Gallica e da Google Books.

cane, anche alcune opere tutelate da diritti, senza ottenere il consenso preliminare degli aventi diritto, il che ha dato luogo ad un importante contenzioso sia negli Stati Uniti che in Europa, soprattutto in Francia (vedere paragrafo 1.2.1). La seconda fonte è quella degli editori partner. Google si procura anche i metadati (informazioni di identificazione dell'opera) e crea un'immagine che ricostituisce una falsa copertina, quando non dispone del contenuto digitalizzato, al fine di fornire l'accesso ad un minimo di informazioni (titolo, autore, ISBN, numero delle pagine, ecc.) sul libro. Quindi, effettuando una ricerca di un'opera recente appartenente ad un editore non partner si accederà ad una pagina contenente informazioni varie ed eventualmente commenti degli internauti e link ai siti di librerie e biblioteche. All'inizio del 2010, Google Books annuncia che la piattaforma consente di effettuare ricerche sulla versione integrale di oltre 10 milioni di libri⁷. Di questi libri, 2 milioni sono stati digitalizzati sulla base di accordi con gli editori e 1,5 milioni sono di dominio pubblico. Le altre opere, coperte da diritti, sono state digitalizzate senza il consenso dei rispettivi titolari.

La biblioteca digitale Gallica è stata sviluppata dalla BnF a partire dalla metà degli anni Novanta, nel contesto del progetto su grande scala voluto da François Mitterrand. Nel 1997, alla sua inaugurazione, proponeva qualche decina di migliaia di documenti, principalmente in modalità immagine. Inizialmente concepita come una biblioteca digitale selettiva a vocazione enciclopedica che proponeva *corpus* di documenti (le riviste di società accademiche, i *voyages en Italie*, ecc.), a partire dal 2005 ha intrapreso una profonda trasformazione, in risposta ai progetti di digitalizzazione di Google. La BnF ha quindi sviluppato una sua politica di digitalizzazione di massa (contratto Jouve "dei 30.000" e contratto Safig "dei 100.000" nel 2007) passando alla modalità testo (contratto

di "ocerizzazione" dei contenuti già presenti in Gallica, detto "dei 60.000").

L'avvio dei dibattiti con il Sindacato nazionale degli editori (SNE), alla fine del 2007, con lo scopo di consentire l'accesso ai contenuti digitali protetti da diritti d'autore attraverso Gallica, ha segnato un'altra importante tappa di questa evoluzione. Gli editori francesi sono presenti nel portale Gallica tramite la segnalazione di quasi 20.000 libri contemporanei digitalizzati. I documenti sono consultabili, in base a determinate condizioni, nel sito dei distributori digitali.

A partire dal 2005, l'archivio Gallica si è arricchito anche di contenuti di stampa periodica (quotidiani di grande formato del XIX secolo) per mezzo di un importante contratto di digitalizzazione specifica (3,5 milioni di pagine per una ventina di testate) finanziato dal Senato.

Alla fine del 2009, Gallica offre l'accesso a oltre 950.000 documenti, di cui circa 370.000 in modalità testo. Tra questi documenti si contano 145.000 libri (monografie), 650.000 periodici e 115.000 immagini.

930.000 documenti provengono dalle collezioni della BnF, i restanti dagli editori che hanno aderito al progetto e dalle biblioteche partner. La BnF si è impegnata a garantire l'accesso ai documenti digitali di altre biblioteche, ospitandoli oppure procedendo a *Web-harvesting* con il protocollo OAI-PMH. Tale offerta è ancora modesta, con meno di 4000 documenti appartenenti a biblioteche partner accessibili da Gallica (0,4 % del totale di Gallica). I documenti non coperti da diritti sono segnalati in Europeana, per la quale Gallica è uno degli aggregatori per la Francia.

I principali lavori tecnici attualmente in corso sono la modernizzazione dell'interfaccia di consultazione (nel dicembre 2009 è stato proposto un nuovo visualizzatore), la modernizzazione del motore di ricerca (nel corso del 2010) ed il rinforzo della capacità di immagazzinamento e divulgazione, allo scopo di migliorare

⁷ <http://googleblog.blogspot.com/2009/10/tale-of-1000000-books.html>.

la qualità della risposta agli internauti. La BnF si sta anche impegnando per strutturare i dati digitali ed i metadati associati, in un contesto internazionale. Nel 2009 sono anche stati stipulati nuovi contratti di digitalizzazione (relativi sia a documenti specializzati che a libri rari e preziosi).

Parallelamente, negli Stati Uniti, la Library of Congress ha portato avanti, a partire dai primi anni Novanta, un'ambiziosa politica di digitalizzazione, che ha ricevuto ingenti finanziamenti pubblici (provenienti dal Congresso) e oltre 45 milioni di dollari da privati, principalmente sotto forma di donazioni. È stato così realizzato il programma American Memory (<http://memory.loc.gov/ammem/index.html>, ovvero una biblioteca digitale che raccoglie oltre cinque milioni di documenti liberamente accessibili, soprattutto manoscritti, documenti iconografici e contenuti di stampa periodica, sulla base degli obiettivi della politica documentaria definita dalla grande biblioteca nazionale nordamericana. Questi documenti, suddivisi in un centinaio di collezioni tematiche, provengono dalla Library of Congress, ma anche da altre istituzioni culturali americane. Di recente il Giappone ha respinto una proposta di partenariato con Google per la digitalizzazione dei libri ed ha deciso, alla fine del 2009, di intraprendere un proprio programma nazionale di digitalizzazione basato sui finanziamenti pubblici insieme alla Biblioteca della Dieta che svolge il ruolo di biblioteca nazionale. Si prevede che nel 2010 saranno erogati 90 milioni di euro a fronte di un miliardo di euro circa per tutto il progetto.

La biblioteca digitale Europeana è di fatto un portale di consultazione e non un sito che ospita i contenuti stessi. Il suo sviluppo è stato affidato alla fondazione olandese EDL (European Digital Library), parzialmente finan-

ziata dalla Commissione Europea nell'ambito dei concorsi a progetto, ed in parte da alcuni Stati Membri.

Il portale Europeana è stato inaugurato nel novembre del 2008 (versione beta, <http://www.europeana.eu>). Sono consultabili circa sei milioni di documenti, tra cui pochissimi libri (meno di 200.000). I contenuti proposti dalla Francia, soprattutto attraverso il portale Collections del Ministero della cultura, la biblioteca digitale Gallica (vedi sopra) ed il sito dell'INA, rappresentano attualmente circa la metà del totale dei documenti accessibili tramite Europeana. Il lancio della versione operativa è previsto per il secondo semestre del 2010 con un obiettivo di 10 milioni di documenti online. Oltre un migliaio di istituzioni culturali europee aderiscono ad Europeana, ma con livelli di partecipazione ed offerte di contenuti estremamente disuguali. La Commissione Europea sta attualmente considerando l'evoluzione di Europeana ed a tal fine, nell'agosto 2009, ha proposto la consultazione pubblica Europeana – next steps. Le questioni principali vertono sui contenuti che il portale deve offrire agli internauti, le possibili modalità di finanziamento e di gestione, le soluzioni fattibili ed accettabili per coinvolgere in maniera più efficace il settore privato in questo progetto ed aumentarne la capillarità.

Dal canto loro, oltre all'offerta proposta nell'ambito di Gallica, i principali editori francesi⁸ stanno mettendo a punto un'offerta digitale in grado di rispondere alle aspettative degli internauti, nel pieno rispetto dei diritti d'autore. La realizzazione di questo progetto presuppone un'evoluzione delle diverse professioni del settore, ingenti investimenti finanziari (con il sostegno dello stato, principalmente tramite i nuovi sussidi erogati dal Centro nazionale del Libro a favore della digitalizzazione, creati nel 2008) e l'identificazione dei diritti effettiva-

⁸ Gli editori del settore scientifico/tecnico/medico hanno intrapreso da molti anni l'accesso digitalizzato alle loro collezioni, soprattutto nell'ambito del diritto, della medicina e delle scienze. La novità è rappresentata dall'estensione alla letteratura generale, di cui si parla in questo caso.

mente detenuti dalle rispettive parti per lo sfruttamento digitale delle opere.

Lo sviluppo di questa offerta digitale (stimata a fine 2009 a circa 40.000 titoli francesi disponibili) nel 2009 si è tradotto nel lancio di diverse piattaforme di distribuzione (Numilog del gruppo Hachette, Eden-Livre che raggruppa Flammarion, Gallimard e La Martinière, sito E-Plateforme di Editis, L'Harmathèque di L'Harmattan, ecc.). Queste diverse piattaforme di distribuzione (B to B) vanno ad aggiungersi ad un'offerta precedente, costituita invece da aggregatori digitali indipendenti dagli editori (soprattutto Cyberlibris e Numilog, prima di essere acquisita da Hachette) rivolta direttamente agli internauti (B to C)⁹. A medio termine, gli editori francesi, preparano la transizione verso una filiera produttiva nata in digitale. L'offerta digitale editoriale dovrà essere costituita, almeno in un primo tempo, soprattutto da titoli recenti¹⁰.

1.1.3 Un'idea scaturita dall'evoluzione degli utilizzi

Tutte queste nuove biblioteche e piattaforme di libri digitali non sarebbero nate senza lo sviluppo di nuove modalità di utilizzo, tipiche della ricerca sul Web. La rapida diffusione di Internet ha comportato profondi cambiamenti in termini di accesso alla conoscenza ed all'informazione. Vi sono in particolare due tipi di utilizzo che spiegano in parte l'interesse suscitato dallo sviluppo delle biblioteche digitali e che potrebbero influenzare profondamente le considerazioni in materia di strutturazione di queste biblioteche.

La prima modalità di utilizzo è rappresentata dal ricorso ormai prioritario ai motori di ricerca. I motori di ricerca costituiscono oggi strumenti

la cui efficacia è universalmente riconosciuta e consentono agli internauti di accedere alla massa di conoscenze disponibili in rete. Oltre a Google, in passato sono stati utilizzati ampiamente altri motori di ricerca oppure lo sono tuttora, anche se a livelli diversi (Altavista, Yahoo!...), altri ancora stanno emergendo (Bing); resta il fatto che lo strumento del motore di ricerca è indispensabile per gli internauti, anche nell'ambito delle loro abitudini di consumo culturale. Questo successo è attribuibile principalmente alla combinazione di più fattori: un modello economico molto forte, gratuito per l'utente, la semplicità di utilizzo di questo tipo di strumento e la potenza dell'algoritmo basato su una concezione specifica della pertinenza e su un'infrastruttura tecnologica estremamente potente ed efficace.

Due sono gli elementi che definiscono l'efficacia di un motore di ricerca: la pertinenza e la potenza. Ora, i principali motori del Web, soprattutto Google, hanno inizialmente optato per la potenza. Ricordiamo brevemente i principi fondamentali del funzionamento di un motore di ricerca come Google, utili per comprendere meglio le strategie da adottare in materia di digitalizzazione:

- innanzitutto, un robot esplora automaticamente e ad intervalli regolari la rete; il robot segue tutti i link ipertestuali che trova per poi recuperare ed indicizzare tutte le risorse utili. La prima fase consiste quindi in una "visita" estesa della rete¹¹, volta a reperire documenti e pagine Web;
- la fase successiva consiste nell'indicizzazione dei documenti raccolti che potrebbero essere oggetto delle ricerche degli internauti per mezzo delle parole chiave contenute. Occorre quindi estrarre le parole più

⁹ Hachette ha annunciato il lancio del sito Mybook, rivista «B to C» completa di funzioni commerciali e di community.

¹⁰ Un progetto come quello dell'editore Gallimard (digitalizzazione di oltre 25.000 titoli) rappresenta per il momento un'eccezione.

¹¹ Per vari motivi, principalmente di ordine tecnico, una parte del web non è accessibile a questi robot; si parla in questo caso di "Web profondo" o "Web invisibile".

- significative di ogni documento che sono poi classificate in base ad un dispositivo di indicizzazione del motore di ricerca. Parallelamente, le parole estratte sono oggetto di una ponderazione, che corrisponde generalmente alla frequenza di apparizione di tale parola nel documento (ma possono essere utilizzati anche altri criteri);
- la fase della ricerca è quella che, in seguito alla richiesta degli internauti, fornisce i risultati classificati in ordine di pertinenza, valutata in base ad algoritmi specifici per ogni motore. Google utilizza principalmente il modello del *page rank*: la pertinenza di un documento è stabilita soprattutto in base alla sua notorietà sul Web. Tale notorietà viene calcolata con l'ausilio di un calcolo complesso che attribuisce ad ogni pagina Web un indice di popolarità "affidabile". Questo punteggio è per sua natura evolutivo, in quanto è il risultato di un'analisi globale e permanente delle consultazioni sul Web e dei "percorsi di click" attraverso i vari link.

Quindi, i risultati sono presentati a seconda della loro popolarità. Nel frattempo, sono in corso delle ricerche volte a sviluppare altri tipi di motori, maggiormente incentrati sulla pertinenza, grazie ad analisi semantiche. L'idea è quella di associare al termine cercato altre parole di contenuto semantico o logico vicino, allo scopo di rispondere alla domanda posta, quando invece i motori attuali si limitano a cercare la concordanza tra le parole chiave della richiesta ed il loro indice. Infine, l'evoluzione verso un Web semantico dovrebbe consentire la creazione automatica di link tra i documenti digitalizzati (per esempio la versione digitale di un libro, la menzione di tale libro in un articolo, la biografia dell'autore in Wikipedia, ecc.) ed anche la gerarchizzazione di tali documenti. Ma la creazione di questo Web semantico presuppone un importante lavoro iniziale sulla qualifica dei diversi dati del Web, che sembra essere ancora ben lontano dal vedere la luce.

Tuttavia, l'accesso potenzialmente universale alle risorse online proposto dagli attuali motori risulta sufficientemente allettante per gli internauti che appoggiano questa modalità di ricerca, a maggior ragione quando si tratta di opere. Infatti, la possibilità di reperire immediatamente contenuti online, con il conseguente risparmio di tempo, per non parlare degli spostamenti che l'acquisizione del libro cartaceo comporta, sembra essere molto più interessante sia per i ricercatori che per il pubblico in genere, rispetto agli eventuali punti deboli qualitativi relativi alle modalità di ricerca dei motori. Pertanto, lo sviluppo di tali utilizzi è sufficientemente invogliante da spingere ad una rapida digitalizzazione dei contenuti per metterli a disposizione sul Web. Per quanto concerne i motori di ricerca, questa motivazione è ulteriormente rinforzata dalla volontà di disporre di un numero maggiore di documenti per aumentare la ricchezza e la pertinenza delle loro risposte e, conseguentemente, accrescere la gamma documentaria delle loro risorse pubblicitarie.

Lo sviluppo dei social network

Nel mondo dei social network, l'utente crea il suo universo e lo contestualizza. La logica è molto diversa da quella del motore di ricerca: in alcuni settori, e soprattutto per quanto concerne l'informazione, non è più l'internauta che va a cercarla, ma è l'informazione che gli arriva tramite il suo network. Inoltre le informazioni che riceve sono già filtrate e contestualizzate a seconda dei suoi centri di interesse o delle persone che fungono da risorse del suo network. L'internauta può anche, a sua volta, proporre ai membri del proprio network la sua biblioteca ideale e divenire egli stesso fonte di informazione (un'applicazione di questo tipo è proposta, per esempio, da Facebook).

Questo utilizzo relativamente nuovo rappresenta un interesse specifico per i libri digitalizzati, in quanto si tratta di un'altra possibile modalità di accesso, diversa da quella dei motori. Le informazioni relative al libro ed al suo contenuto passano attraverso modalità di ricer-

ca diverse dagli algoritmi o dal Web semantico. È necessario che le biblioteche prendano in considerazione il rapido sviluppo di questo nuovo tipo di utilizzo, con la conseguente proposta di servizi *ad hoc*.

1.2 Un ambiente caratterizzato da una marcata incertezza

1.2.1 Google si trova in un contesto giuridico complesso

La digitalizzazione (nell'ambito dei patrimoni delle biblioteche partner) delle opere coperte da diritti d'autore, senza l'ottenimento del consenso preliminare dei rispettivi titolari, è stata oggetto, a partire dal 2005, di un contenzioso negli Stati Uniti. Il progetto di accordo transazionale a cui sono giunte le parti in data 28 ottobre 2008 deve essere ancora convalidato dal giudice, ma, nel frattempo ha suscitato un coinvolgimento emotivo di livello internazionale, tale da giustificare l'introduzione di emendamenti.

Le opere digitalizzate coperte da diritti appartenenti alle collezioni delle biblioteche universitarie americane sono interamente indicizzate dal motore; la ricerca "a tutto testo" genera la visualizzazione di brevi estratti presentati sotto forma di pezzi di carta strappati (i cosiddetti "snippets"). Nel 2005, le associazioni americane dei titolari di diritti (American Publishers Association e Author's Guild) hanno intentato un'"azione di categoria" contro Google, agendo a nome delle intere "categorie" rappresentate (ovvero tutti gli editori e gli autori).

Si è trattato di un processo basato sulla violazione dei diritti d'autore: la società Google si è vista accusare di violazione del *copyright* per aver riprodotto e pubblicato questi libri senza autorizzazione preliminare. Google si è difesa da tali accuse avanzando l'argomentazione del *fair use* (utilizzo equo), un'eccezione molto generica prevista dal diritto americano. Google si è anche dichiarata disposta a ritirare le opere su richiesta dei rispettivi aventi diritto (*opt out*), una pratica molto contestata in quanto contraria ai principi della proprietà intellettuale;

è stata anche denunciata la lentezza con cui Google darebbe seguito alle richieste di ritiro.

Il giudice non ha più dovuto pronunciarsi nel merito: nell'ottobre 2008, dopo una procedura molto costosa durata tre anni, le parti hanno reso pubblico un progetto di accordo transazionale di categoria (*Class action settlement agreement*) che, se confermato dal tribunale, dovrebbe risolvere il contenzioso. Per una specificità del diritto americano, questo accordo potrebbe vincolare tutti i componenti delle "categorie" rappresentate, fatte salve quelle che si ritireranno espressamente. È stata quindi lanciata una vasta campagna pubblicitaria internazionale per comunicare agli autori ed agli editori che, grazie ad un documento di oltre 300 pagine, redatto in inglese, la situazione dei loro diritti sui libri sarebbe cambiata.

Questo primo progetto comportava due aspetti molto importanti. Da una parte, avrebbe estinto, tramite un sistema di risarcimento, tutte le azioni legali passate e future intentate contro Google per i fatti inizialmente contestati. Dall'altra parte, andava a creare modalità per lo sfruttamento commerciale da parte di Google di tutti i libri digitalizzati. Infatti, Google avrebbe potuto liberamente sfruttare i libri non disponibili nei grandi canali di vendita americani, fatte salve eventuali obiezioni esplicite, presentate dagli editori per i singoli titoli. Questo sfruttamento sarebbe stato limitato al territorio americano, sulla base dell'indirizzo IP dei consumatori. Nel caso dei libri "rivendicati" dai rispettivi aventi diritto, se i titolari avessero autorizzato lo sfruttamento da parte di Google, si sarebbero visti versare il 33% del fatturato generato (pubblicità, accesso a pagamento ai singoli titoli da parte dei privati, accesso a pagamento all'intero database da parte delle istituzioni). Il 33% del fatturato generato dai libri "non rivendicati" sarebbe stato versato ad un *Registro dei diritti sui libri (Books Rights Registry)*, creato al fine di incoraggiare i titolari dei diritti di tutto il mondo ad iscriversi.

Questo primo progetto ha suscitato una serie di vivaci polemiche, in quanto disconosceva i

principi della proprietà intellettuale, obbligando all'*opt out* i titolari dei diritti, qualora non avessero accettato lo sfruttamento dei loro libri da parte di Google. Inoltre, poneva la società in una situazione di monopolio sul territorio americano, relativamente allo sfruttamento delle opere "non rivendicate"; infatti, l'accordo transazionale avrebbe attribuito a Google e solo a Google, una licenza di sfruttamento che deroga dal diritto comune. La società Google presentava questo aspetto del progetto di accordo come l'unico modo disponibile per risolvere la questione delle opere "orfane", ovvero le opere i cui aventi diritto non sono stati identificati e che, vista l'impossibilità di ottenere il loro consenso, non possono, in linea di principio, essere sfruttate (vedi paragrafo 1.2.2).

Il giudice incaricato dell'eventuale convalida del progetto di accordo transazionale ha ricevuto numerosissime obiezioni e commenti. Il governo francese e quello tedesco, e poi il governo americano, hanno infatti espresso le proprie riserve sui termini e sugli effetti della transazione. Le parti hanno quindi deciso di apportare degli emendamenti ed il 13 novembre 2009 è stato presentato un nuovo progetto di accordo. La modifica più significativa consiste nel fatto che l'accordo d'ora in avanti si baserà esclusivamente sui libri inizialmente pubblicati negli Stati Uniti, nel Regno Unito, in Canada ed in Australia, ed anche sui libri registrati (dagli editori di tutto il mondo) presso l'Ufficio dei diritti d'autore degli Stati Uniti (il che significa, per parecchi editori francesi, una parte considerevole del loro catalogo). Poi è stato apportato qualche cambiamento volto a contenere la natura monopolistica del sistema. In ogni caso, l'economia generale del progetto non è stata rimessa in causa.

Il giudice deve pronunciarsi in merito a questo progetto emendato il 18 febbraio 2010. Se il testo di questo accordo transazionale sarà convalidato, Google sarà in grado di mettere a punto un'immensa piattaforma per la commer-

cializzazione di libri, per la stragrande maggioranza anglofoni, ed inoltre vanterà diritti di esclusività per lo sfruttamento di una gran parte di tali libri, vale a dire tutti quelli che non saranno "rivendicati" presso il *Registro dei diritti sui libri*. Nel frattempo, i commenti che il governo americano ha presentato al giudice non sembrano essere stati presi in considerazione nell'ambito del progetto emendato e non è da escludere un intervento delle autorità americane che disciplinano la concorrenza.

Un contenzioso simile in Francia

Il 6 giugno 2006, il gruppo La Martinière ha intentato una causa contro le società Google Inc. e Google France sulla base di capi d'accusa globalmente simili a quelli avanzati dai titolari dei diritti d'autore americani (vedi sopra), ovvero la violazione dei diritti d'autore per la riproduzione non autorizzata e la pubblicazione di brevi estratti sotto forma di *snippets*. I querelanti sono stati appoggiati nell'ottobre del 2006 da un intervento volontario del Sindacato nazionale degli editori (SNE) e nel novembre 2006 da un intervento volontario della Société des Gens de Lettres (SGDL, Società dei letterati).

Dal canto suo, Google ha ritenuto che, ai fini delle accuse relative alla digitalizzazione, conveniva applicare il diritto americano, ove le operazioni si sono svolte nel territorio degli Stati Uniti e che, conseguentemente, tale digitalizzazione non rappresentava una violazione nella misura in cui rientrava nell'eccezione del cosiddetto "*fair use*". Ai fini invece delle accuse relative alla pubblicazione non autorizzata, tale pratica rientrava nell'ambito dell'eccezione della "breve citazione" riconosciuta dal diritto francese¹².

L'udienza si è tenuta il 24 settembre 2009 e la Corte d'Appello di Parigi ha emesso la sentenza il 18 dicembre, constatando che il diritto applicabile è quello francese, sia per quanto concerne le operazioni di digitalizzazione che per la riproduzione di brevi estratti. Detto ciò, ha

¹² Articolo L 122-3 del *Codice della proprietà intellettuale*.

concluso che la società Google Inc. si è effettivamente resa colpevole di violazione dei diritti di autore con la riproduzione senza autorizzazione preliminare prima e con la riproduzione di opere protette da diritti poi. Infatti, ha ritenuto non applicabile al caso di Google l'eccezione della breve citazione, in quanto gli estratti sono pubblicati in maniera aleatoria, escludendo pertanto qualsiasi finalità di natura informativa.

Conseguentemente, la società Google Inc. è stata condannata in prima istanza a versare un risarcimento danni pari a 300.000 € al gruppo La Martinière e un importo simbolico di 1 € alla SGDL e alla SNE. Inoltre, il tribunale ha interdetto la società dal perseguire queste sue azioni, imponendo una penale di 10.000 € per ogni giorno di ritardo.

Va sottolineato che la sentenza si basa soltanto su una lista precisa e ben identificata di libri che era stata redatta da un perito allo scopo di dare inizio alla procedura. Nel frattempo, Google corre il rischio di affrontare una miriade di processi simili intentati dagli editori francesi, considerando che la sentenza è applicabile anche alla loro situazione. Ha dichiarato la sua intenzione di ricorrere in appello contro tale sentenza.

1.2.2 Un coordinamento insufficiente degli altri soggetti

A livello europeo, vi sono delle questioni che restano ancora da precisare. La prima è quella delle opere orfane: un'opera è orfana solo se almeno uno dei titolari dei rispettivi diritti non è stato rintracciato, nonostante siano state condotte ricerche approfondite. Il problema delle opere orfane si è acuito a fronte dei grandi progetti di digitalizzazione realizzati dalle istituzioni culturali. Infatti, un'istituzione non può, salvo specifiche eccezioni¹³, digitalizzare e mettere a disposizione degli internauti delle opere tutelate senza aver prima ottenuto il

consenso dei titolari dei diritti di sfruttamento digitale di tali opere. Ora, la ricerca degli aventi diritto può essere molto costosa ed a volte infruttuosa. Tutti i progetti di digitalizzazione di massa si trovano ad affrontare questo ostacolo giuridico. Si è visto che Google ha aggirato questo ostacolo digitalizzando sistematicamente i libri protetti sul territorio americano, senza metterli inizialmente a disposizione degli internauti, il che consentirebbe alla società di trincerarsi dietro al principio del *fair use*, da cui il contenzioso con gli aventi diritto. Se il progetto di accordo amichevole fosse convalidato dal giudice americano, Google potrebbe sfruttare i diritti della propria digitalizzazione senza essere costretta a cercare i titolari dei diritti d'autore sui libri (vedi paragrafo 1.2.1).

Tale prospettiva fa temere un monopolio sullo sfruttamento digitale dei libri non rivendicati, da cui il desiderio della Commissione europea di fornire alle istituzioni culturali europee la sicurezza giuridica necessaria per portare a buon fine i loro progetti di digitalizzazione di massa. In Francia, in seguito alle *Raccomandazioni* della Commissione specializzata del Consiglio superiore della proprietà letteraria ed artistica¹⁴, il Ministero della cultura e della comunicazione si è impegnato a mettere a punto nel 2010 un sistema che consenta la gestione collettiva delle opere orfane, almeno per il settore della carta stampata e dell'immagine fissa. Altri paesi europei si stanno orientando verso soluzioni in linea con la loro sensibilità culturale nazionale (come nel caso della Scandinavia, con le licenze estese). Il compito della Commissione sarà quello di controllare che la molteplicità di tali sistemi non pregiudichi la circolazione e la libera diffusione delle opere in questione, senza cedere alla tentazione di una nuova eccezione comunitaria poco compatibile con la diversità di tradizioni nazionali spesso sensibili.

La seconda questione riguarda il finanziamento di Europeaana. Il portale europeo ha finora trat-

¹³ In Francia, per esempio, si può citare l'eccezione "biblioteche" (CPI, L 122-5 et L 211-3) o anche il diritto specifico applicabile alle istituzioni preposte al deposito legale (*Code du patrimoine*, L 132-4).

¹⁴ CSPLA, Commissione sulle opere orfane, 18 marzo 2008.

to le proprie risorse soprattutto da bandi di concorso per progetti comunitari, per loro natura non continuativi, il che rappresenta un importante punto di fragilità. Questa situazione potrebbe evolversi a partire dal 2013 con l'entrata dell'Unione Europea in un nuovo ciclo di budget. Al fine di consolidare il funzionamento di Europea e consentirle di entrare in un'ottica strategica pluriennale, è necessario che il finanziamento del portale disponga di risorse costanti. Le fonti dei contributi potrebbero essere tre: la Commissione, gli Stati Membri, i partner privati. Un equilibrio soddisfacente tra questi tre assi deve ancora essere trovato, come anche l'organizzazione di una nuova gestione per il progetto. Queste condizioni sono necessarie per il futuro sviluppo dell'interfaccia. L'equilibrio dipenderà soprattutto dalla volontà più o meno marcata dei diversi soggetti coinvolti a finanziare il progetto.

***Iniziativa private non coordinate:
l'offerta degli editori in Francia***

Dato che il libro digitale rappresenta ancora un mercato emergente, senza un volume d'affari stimato né una regolamentazione da parte delle autorità pubbliche simile a quella esistente da tempo nell'universo fisico (libro cartaceo), le azioni dei principali soggetti economici, in particolare gli editori della letteratura generale, sono state fino a poco tempo fa di natura prettamente sperimentale.

Sinora non sono emerse soluzioni interprofessionali tali da garantire l'interoperabilità delle offerte ed una loro applicazione massiva. Le principali iniziative sono state intraprese dai vari gruppi economici, con la messa a punto, a partire dal 2009, di diverse piattaforme di di-

stribuzione o diffusione di libri digitali (vedi sopra). Rimane da garantire il collegamento tra queste piattaforme B to B ed i portali di vendita delle librerie (B to C). È necessaria una conciliazione interprofessionale per garantire ad ogni rivenditore, una volta in grado di vendere libri digitali (con l'accordo dei distributori), un accesso semplice, omogeneo ed esaustivo all'offerta di tutti gli editori francesi. Il progetto del "hub" recentemente proposto da Dilicom, una struttura interprofessionale già operante come centro di informazioni e gestione degli ordini per i libri cartacei, rientra in tale strategia. Il suo successo sembra tuttavia condizionato all'eventuale consenso unanime degli editori francesi, i quali dovrebbero accettare che la loro offerta sia presentata sul maggior numero di piattaforme commerciali possibile, senza diritti di esclusività né restrizioni di sorta.

Nel frattempo, i vari settori coinvolti proseguono con i loro lavori, sotto la guida dei sindacati di categoria (SNE con la sua commissione digitale, SLF con il progetto di portale della libreria indipendente) e con la partecipazione delle autorità pubbliche (lavori organizzati per mezzo di vari rapporti e missioni¹⁵, con il Centro nazionale del Libro che finanzia varie iniziative).

Tuttavia, dal punto di vista dell'internauta, il panorama dell'offerta può sembrare ancora poco leggibile e poco chiaro.

1.2.3 Il libro digitale non ha una definizione univoca

Infine, più generalmente parlando, il "libro digitale" rimane un oggetto difficile da definire, il che spiega in parte perché in questo settore la digitalizzazione procede più lentamente rispetto ad altri ambiti culturali, come la musica

¹⁵ Rapporto sui libri digitali a cura di Bruno Patino, giugno 2008; rapporto di Hervé Gaymard, Situazione del libro – Valutazione della legge che disciplina il prezzo del libro e prospettive future (marzo 2009); rapporto della missione "Creazione e internet", affidata a Patrick Zelnik, Jacques Toubon e Guillaume Cerruti, presentata il 6 gennaio 2010 al Ministro della cultura e della comunicazione (quest'ultimo rapporto ha raccomandato la realizzazione di una piattaforma comune di distribuzione digitale in base ad un modello B to B. Consultabile in formato PDF all'indirizzo web: <http://www.culture.gouv.fr/mcc/Actualites/A-la-une/Remise-du-rapport-de-la-mission-creation-et-internet>.

o il cinema. Un libro digitale può rappresentare realtà molto diverse tra loro, dalla semplice riproduzione digitalizzata al prodotto digitale provvisto di funzionalità molto avanzate (ricerca, intertestualità, interattività, complementi editoriali). La normativa fiscale definisce il libro come un

«insieme stampato, illustrato o meno, pubblicato con un titolo, avente per oggetto la riproduzione di un'opera dell'ingegno di uno o più autori, a fini di insegnamento e divulgazione del pensiero e della cultura»,

quando invece il contenuto digitalizzato di un libro può essere proposto in base a forme e supporti molto diversi. Il file può essere scaricato integralmente oppure accessibile tramite la sottoscrizione di un abbonamento tematico, può essere "di sola lettura", oppure modificabile, ecc., tutte possibilità diverse che rendono difficile definire in maniera univoca il libro digitale¹⁶ e che moltiplicano le possibili opzioni di presentazione dei contenuti agli internauti. Questa labilità del libro digitale mette in discussione il concetto tradizionale del libro stampato e ne rinnova gli utilizzi potenziali.

L'assenza di una precisa definizione può anche contribuire a rinforzare il senso di indefinitezza che contraddistingue gli accordi sottoscritti tra Google e le biblioteche partner, soprattutto per quanto concerne le condizioni in base a cui queste biblioteche potrebbero utilizzare o non utilizzare le loro copie dei file digitalizzati. La seconda parte di questo rapporto andrà ad analizzare la portata di tale aspetto.

¹⁶ La definizione di "libro digitale" sembra un po' più semplice se si riferisce alla sola trasposizione di un libro stampato: è il percorso scelto dagli autori del rapporto Zelnik che raccomandano di estendere il sistema del prezzo fisso esistente nell'universo del cartaceo anche al libro digitale, una misura ritenuta necessaria (proposta 10: «Etendre le prix unique du livre aux livres numériques dits homothétiques»).

¹⁷ Da notare che sinora nessuna biblioteca nazionale, fatta eccezione, forse, per la Biblioteca di Catalogna, che gode tuttavia di uno statuto speciale, ha concluso accordi di digitalizzazione con Google.

II. Gli accordi con Google: una risposta inadeguata

II.1 Una risposta inadeguata alla missione delle biblioteche

La missione delle biblioteche in materia di patrimonio digitale, è fondamentalmente duplice:

- garantire la perennità, a lungo termine, del patrimonio scritto digitale, ovvero la conservazione e l'aggiornamento dei file digitali, in un contesto di obsolescenza più o meno rapida delle tecnologie;
- consentire l'accesso più ampio possibile al patrimonio digitale. Ciò implica, da una parte, la messa a disposizione e la visibilità di tale patrimonio digitalizzato su Internet, al termine di un processo di "massa" che richiede lo spiegamento di mezzi importanti; dall'altra parte, supporti e strumenti digitali di qualità sufficiente, tali da rendere possibili i diversi utilizzi degli internauti.

Gli attuali progetti di accordo tra Google ed alcune biblioteche¹⁷ lasciano aperti, per ciò che concerne tali obiettivi, una serie di interrogativi.

II.1.1 Il Gruppo di lavoro sulla conservazione

L'obiettivo della conservazione del patrimonio digitale è strettamente collegato alla missione della biblioteca di rendere disponibile tale patrimonio. Da una parte, presuppone una garanzia in quanto alla perennità dei dati digitalizzati e, dall'altra parte, la possibilità di immagazzinare tali dati e trasferirli su diversi supporti digitali, certamente destinati ad evolversi. La qualità della messa a disposizione e la possibilità di moltiplicare le forme di tale accessibilità, dipenderà dalla qualità della conservazione dei dati e dei loro supporti.

Gli accordi in essere si mostrano decisamente insufficienti a tale proposito, in quanto non sfiorano nemmeno la tematica della conservazione e dell'eventuale aggiornamento dei file forniti da Google alle biblioteche, e questo vale sia per l'accordo con la Biblioteca municipale di Lione che per il progetto di accordo (*Memorandum of understanding*) predisposto dalla BnF.

Infatti, sulla base di questi accordi o progetti di accordo, Google deve sì fornire alla biblioteca i file da essa digitalizzati, ma se in un secondo momento dovesse aggiornare questi file, gli accordi non contemplano alcun obbligo di fornire tali aggiornamenti alla biblioteca.

Allo stato attuale, spetta dunque alla biblioteca il compito di gestire gli aspetti legati alla conservazione ed alla perennizzazione dei contenuti digitali, il che comporta un costo affatto trascurabile.

Inoltre, spesso questi accordi non sono abbastanza precisi in merito ai file che saranno effettivamente forniti da Google. Nel caso specifico dell'accordo con la Biblioteca municipale di Lione, il capitolato delle condizioni tecniche speciali contempla solo la trasmissione dei file di immagini e del file di testo "grezzo" (testo non strutturato), senza alcuna sorta di impegno relativamente alla natura dell'"ocerezizzazione" e del trattamento effettuato; non è quindi affatto ovvio che la biblioteca sarà poi in grado di collegare le immagini al testo. Questa vaghezza degli accordi è legata alle restrizioni imposte da Google sui possibili utilizzi dei file da parte della biblioteca, che potrà far scaricare ai suoi utenti solo le immagini dei libri e non i file di testo (vedi più avanti). Pertanto, la biblioteca non sa con sufficiente precisione cosa riceverà esattamente dal suo contraente, il che lascia un punto interrogativo sia per quanto riguarda la conservazione che l'utilizzo dei file da parte della biblioteca¹⁸.

II.1.2 Il Gruppo di lavoro sull'accessibilità

Alla luce della loro missione, e cioè consentire l'accesso al patrimonio digitale, le biblioteche devono preservare la loro autonomia operativa sui file e sul loro utilizzo, e quindi essere in grado di sottoscrivere partenariati diversi.

Ora, gli accordi proposti da Google contemplano una serie eccessiva di limitazioni e di clausole di esclusività, esplicite o implicite.

Se da un lato è abbastanza ovvio che chi (il partner privato) si fa carico della digitalizzazione di determinate collezioni usufruisca poi di una contropartita, come lo sfruttamento commerciale esclusivo dei file, dall'altro lato occorre garantire in qualche modo che tali contropartite non pregiudichino la valorizzazione e lo sfruttamento dei file da parte delle biblioteche. Il problema principale degli accordi (nella fattispecie, il progetto di protocollo di accordo con la BnF ed il contratto stipulato dalla Biblioteca municipale di Lione) risiede nelle formule imprecise utilizzate (o proposte). Innanzitutto, l'entità dell'esclusività commerciale, che è definita per la biblioteca, ma non lo è per Google, la quale non sempre specifica come intende utilizzare i file (al massimo riporta una clausola indicante che le prestazioni eseguite potranno essere compensate con i proventi delle pubblicità incassati a seguito della pubblicazione online dei patrimoni digitalizzati¹⁹). La biblioteca deve consentire l'accesso gratuito alle opere di dominio pubblico.

Il download sistematico (*systematic downloading*) è anch'esso vietato, ma anche in questo caso il concetto è espresso in maniera molto vaga; probabilmente si riferisce ad operatori esterni che potrebbero scaricare tutti i file presenti nel sito della biblioteca, a scopo di rivenderli. Ma l'ambito preciso di tale nozione è poco chiaro, ed eventuali utilizzi legittimi potrebbero essere ostacolati da una restrizione così generica.

¹⁸ Consultare l'Allegato 3 (pag. 47 e successive).

¹⁹ Articolo 6-2 del contratto tra la Biblioteca municipale di Lione e Google.

L'altra difficoltà riguarda le limitazioni esplicite, che potrebbero pregiudicare le iniziative delle biblioteche volte a migliorare l'accessibilità al loro patrimonio digitalizzato. Infatti, la biblioteca non può condividere il contenuto digitale con una terza parte senza aver prima ottenuto l'autorizzazione di Google: per quanto il concetto generale della clausola possa sembrare logico (i file digitalizzati gratuitamente da Google non possono essere ceduti ad un concorrente), tale clausola potrebbe tuttavia impedire lo svolgimento di progetti in corso con altri partner, anche pubblici. Va detto, però, che durante le trattative con la BnF, Google aveva inviato una lettera in cui accettava espressamente che i file da essa digitalizzati fossero riversati in Gallica e repertoriati sul portale Europeana.

Tutti gli accordi stipulati da Google prevedono che gli altri motori di ricerca non possono accedere ai file da essa digitalizzati, allo scopo di indicizzarli e referenziarli²⁰. In altre parole, questa esclusività si traduce, nella pratica, nella mancata indicizzazione e referenziazione del testo dei libri da parte degli altri motori di ricerca. Solo i metadati, di norma prodotti dalle biblioteche partner, sono accessibili ai motori, il che riduce notevolmente la visibilità su Internet dei file gestiti dalla biblioteca e costituisce un grosso ostacolo allo sviluppo delle biblioteche digitali, sviluppo che forse la biblioteca preferirebbe condurre autonomamente. Da una parte, le motivazioni di Google sono comprensibili, in quanto assumendosi l'onere finanziario e tecnico della digitalizzazione si augura di usufruire, in compenso, dell'esclusività sul contenuto digitalizzato, con il conseguente ampliamento della propria base di ricerca e di indicizzazione e l'aumento dei guadagni, ma, d'altro canto, questo significa permettere ad un soggetto che già riveste una posizione dominante sul "mer-

cato" della ricerca di informazioni e dell'accesso ai contenuti digitali di rinforzare ulteriormente questa posizione di dominio.

Tuttavia, sarebbe il caso di interrogarsi in merito all'indicizzazione ed alla referenziazione dei file delle biblioteche digitalizzati da altri. Gli accordi non affrontano tale questione. Google si riserva la proprietà esclusiva dei file digitalizzati in virtù della collaborazione proposta, ma non menziona gli altri file digitalizzati dalle biblioteche, come se gli uni e gli altri "vivessero ognuno di vita propria". Detto ciò, il partenariato con la Biblioteca municipale di Lione e soprattutto il progetto relativo al protocollo di accordo con la BnF si basano solo su una parte minoritaria del patrimonio e la scelta delle opere da digitalizzare, che spetta ai partner pubblici, è obbligatoriamente limitata alle opere di dominio pubblico in grado di essere digitalizzate. Da qui, la questione della loro eventuale integrazione nelle modalità di consultazione e di interrogazione messe a punto da Google Books. È un peccato che gli accordi non disciplinino, allo stesso tempo, gli aspetti della referenziazione e dell'indicizzazione in Google Books, poiché in questo modo la biblioteca si vede costretta ad attuare ulteriori misure per garantire la referenziazione dei suoi file in Google Books e nel motore Google.

La durata delle clausole di esclusività è anch'essa eccessiva; periodi di oltre vent'anni, che oggi sembrano la norma nell'ambito dei contratti stipulati con Google²¹, sono estremamente lunghi, e possono impedire alle biblioteche di espletare la loro missione, e cioè garantire l'accesso.

Il livello minimo della qualità della digitalizzazione non è stato definito con precisione. Invece, per le biblioteche, questo livello deve essere chiarito e chiaramente comunicato.

²⁰ Tutti i contratti noti contemplano una clausola in base alla quale la biblioteca prenderà le misure tecniche idonee onde impedire l'"accesso automatizzato" ai file forniti da Google.

²¹ Il contratto stipulato con la Biblioteca municipale di Lione contempla un diritto di esclusiva commerciale per un periodo di 25 anni.

Infatti, la questione delle possibilità e degli strumenti di ricerca da applicare ai file digitalizzati che saranno sfruttati dalle biblioteche è fondamentale ai fini degli utilizzi e delle loro possibili evoluzioni²². La pratica contemplata dagli accordi consisterebbe nel proporre il file digitalizzato alla biblioteca che potrà accettarlo o rifiutarlo e, in caso di rifiuto, il processo di “ocerizzazione” (riconoscimento ottico dei caratteri) sarebbe rieseguito con la conseguente consegna di un nuovo file. Ma non è stata fissata alcuna soglia minima di qualità, né tantomeno sono state previste penali nel caso di mancato raggiungimento di una qualche soglia.

Infine, la riservatezza degli accordi pretesa da Google, oltre ad essere un principio difficilmente accettabile per le biblioteche pubbliche²³, non ha contribuito a dissipare i dubbi sul bilancio costi/benefici di questi accordi per le biblioteche.

II.2 L’articolazione tra logica pubblica e logica privata

II.2.1 Una considerazione insufficiente degli “atouts” delle biblioteche

Le biblioteche francesi, soprattutto la BnF, presentano dei vantaggi unici che, nell’ambito di una negoziazione con un partner privato, non devono essere sottostimati.

Il primo è quello del patrimonio, molto completo (nel caso della BnF, grazie al suo ruolo di deposito legale). La BnF dispone anche di metadati bibliografici già costituiti sulla base di un catalogo multimediale contenente oltre 10 milioni di notizie bibliografiche relative a documenti. Questo catalogo è il risultato di un’opera di catalogazione portata avanti da validi professionisti per quasi due secoli, per non parlare del lavoro di conversione retrospettiva durato vent’anni, il cui costo non è stato valutato con precisione, ma che in ogni caso si colloca

nell’ordine di una decina di milioni di euro, se non più. I dati più recenti sono redatti in un formato bibliografico molto dettagliato, il cui valore è universalmente riconosciuto.

Inoltre, il catalogo della BnF si basa anche su metadati MADS (Metada Authority Description Schema, 5 milioni di notizie bibliografiche) che descrive con grande precisione gli autori ed il soggetto dei documenti presenti nella biblioteca, i quali rappresentano l’insieme della produzione nazionale, in quanto provenienti in gran parte dal deposito legale (70.000 libri e 40.000 testate, oltre a decine di migliaia di documenti specializzati, nel 2008). Tutto ciò rappresenta un valore, in quanto permette di distinguere gli autori tra loro e contribuire, ove necessario, ad una migliore gestione dei diritti, soprattutto adesso che la problematica delle opere orfane sta emergendo con particolare forza. È chiaro quindi perché, nell’ambito del progetto di protocollo di accordo con la BnF dell’estate 2009, Google abbia mostrato interesse per questi metadati MADS.

Inoltre, la BnF possiede ormai uno specifico *know-how* in materia di digitalizzazione di massa. Nel 2007, con la firma dei primi contratti per la digitalizzazione di massa, l’istituzione pubblica è riuscita a passare da 5.000 documenti digitalizzati l’anno a circa 100.000 documenti l’anno, di cui grosso modo 40.000 libri nel 2009. L’esperienza così acquisita ha permesso di definire le tappe fondamentali del processo di digitalizzazione ed i requisiti da imporre ai fornitori del servizio, di identificare i potenziali ostacoli, di stabilire progressivamente gli obiettivi in termini di livello qualitativo, di specializzare un certo numero di persone nei diversi settori della digitalizzazione (dalla filiera della digitalizzazione vera e propria fino, in misura minore, alle fasi di divulgazione e conservazione dei documenti digitali così prodotti). Questa prima esperienza ha generato dei risul-

²² Consultare l’allegato 3, da pagina 48 a pagina 52.

²³ I documenti del contratto stipulato con la Biblioteca municipale di Lione sono stati divulgati solo in seguito ad un’apposita richiesta del capo-redattore della rivista «Livres-Hebdo» alla Commissione di accesso ai documenti amministrativi.

tati tangibili con una biblioteca digitale che, nel dicembre 2009, contava oltre 950.000 documenti disponibili online.

Infine, l'istituzione è un "marchio" riconosciuto a livello mondiale e potrebbe pertanto "fare da traino" alle altre biblioteche europee, soprattutto tramite Europea.

Ora, gli accordi non sembrano dare molto peso a questi aspetti, eccezione fatta per il progetto relativo alla BnF, in cui una clausola specifica prevedeva che Google trasferisse alla biblioteca i file delle opere francofone di dominio pubblico digitalizzate provenienti dalle altre biblioteche associate. Questo aspetto di scambio dei patrimoni è molto interessante e potrebbe suggerire altri tipi di partenariato (vedi paragrafo III). La Biblioteca municipale di Lione, dal canto suo, ha insistito affinché Google installasse un'interfaccia nel sito della biblioteca (*hosted solution*), al fine di rendere accessibili i file digitalizzati, in attesa di creare la propria biblioteca digitale sulla base dei file recuperati dal partner, ma, secondo l'accordo, tale biblioteca diventerà autonoma solo al termine del periodo di esclusiva, e cioè dopo 25 anni.

II.2.2 Una negoziazione delicata visto il particolare posizionamento di Google

Tutte le negoziazioni con Google sono particolarmente delicate alla luce della posizione dominante occupata dal motore di ricerca. Ovviamente, il diritto della concorrenza punisce solo l'abuso della posizione dominante ma, nella valutazione degli impegni assunti, vanno certamente presi in considerazione gli obiettivi strategici di Google.

Google ha tutti gli interessi a mantenere la sua posizione privilegiata, in questo universo in costante evoluzione. A tal fine, la digitalizzazione di massa dei libri rappresenta un aspetto di grande interesse, in quanto consente al motore di ricerca di accrescere la sua base di indicizzazione, assicurandosi allo stesso tempo l'esclusività dei contenuti digitalizzati, a lungo termine. Si tratta di una forma di ricerca ad "integrazione verticale" tra lo strumento di ricerca ed i

contenuti indicizzati, non più basata solo sulle risorse liberamente disponibili nel Web, ma anche su contenuti esclusivi, il che comporta un indubbio vantaggio competitivo.

In ragione del suo posizionamento, Google è maggiormente motivata a perseguire l'esclusività relativa all'utilizzo dei file digitalizzati basati sui patrimoni delle biblioteche; solo le biblioteche stesse potrebbero garantirne anche loro lo sfruttamento, pur se limitato. Eppure le risorse digitalizzate provengono da collezioni pubbliche e sono di dominio pubblico. In questo contesto, anche se la digitalizzazione viene effettuata dal partner privato, è difficile accettare clausole di esclusività a lungo termine. In pratica, altri soggetti privati dovrebbero poter accedere alle stesse risorse. Ovviamente, i libri "cartacei" non rientrano nelle clausole di esclusività e quindi altre parti possono eventualmente ridigitalizzarli (sempre che lo stato di conservazione delle opere ne consenta la digitalizzazione). Ma i file digitalizzati dovrebbero comunque poter essere utilizzati in altro modo, al fine di rendere possibili altre iniziative alternative e complementari. Pertanto, è essenziale che il contratto non limiti le iniziative di altre parti private.

A questo proposito si utilizza spesso l'espressione "*essential facilities*", derivante dal diritto della concorrenza, ed applicata al settore delle reti (ferrovie, telecomunicazioni, elettricità). Una *essential facility* per definizione non è intercambiabile o sostituibile, in quanto i costi necessari per la creazione di un'infrastruttura equivalente sarebbero proibitivi, in termini di denaro e/o tempo. L'accesso a questa struttura è quindi indispensabile per gli operatori del settore e chi la gestisce si trova di fatto in una posizione dominante o di monopolio. Applicato ai libri digitalizzati, questo concetto giustifica l'obbligo di mettere i file a disposizione delle altre parti.

D'altro canto, si tratta di una società i cui metodi sono contestati. Prima di tutto, Google ricorre, come abbiamo visto, alla pratica della di-

gitalizzazione dei contenuti protetti da diritti, senza l'autorizzazione preliminare degli aventi diritto, appellandosi alla nozione del *fair use*. Il progetto di accordo transazionale di categoria, se fosse approvato dal giudice americano, andrebbe a convalidare i risultati di una pratica comunque illegale.

Inoltre, Google può essere percepita come una minaccia anche per quanto concerne l'immagazzinamento e l'utilizzo dei dati personali degli internauti. Tale timore è dettato dalla capacità dell'operatore di aggregare dati sparsi per tracciare un profilo dettagliato di milioni di persone (percorso professionale e personale, abitudini in materia di consultazione di Internet, partecipazione a forum, ecc.). Google ha ridotto da 18 a 9 mesi la durata della conservazione dei dati personali dei suoi utenti, un termine comunque superiore ai 6 mesi raccomandati nell'aprile 2008 dal G29, un comitato che riunisce le varie autorità incaricate della protezione dei dati personali (tra cui la CNIL) nei paesi europei²⁴. Inoltre la CNIL accusa Google di non rispettare la legislazione francese applicabile.

Questo posizionamento particolare rende quindi difficili, sin dall'inizio, le negoziazioni. In ogni caso, vi sono delle soluzioni che possono essere proposte, con o senza Google, per digitalizzare il patrimonio delle biblioteche francesi e migliorarne così la visibilità e la disponibilità in rete. La terza parte di questo rapporto tratta appunto delle possibili soluzioni.

III. Le soluzioni possibili

La riflessione si inserisce in un contesto ormai rinnovato, con margini di manovra tali da consentire la conduzione di una politica autonoma.

L'annuncio del Presidente della Repubblica, nel contesto delle priorità del Grande Prestito²⁵, relativo ad uno stanziamento specifico a favore della digitalizzazione del patrimonio culturale e soprattutto dei libri, introduce un cambiamento notevole in termini di dimensioni, ritmo e "filosofia" dei progetti di digitalizzazione. Infatti, permette di considerare una politica di digitalizzazione del patrimonio scritto ambiziosa e, allo stesso tempo, autonoma. Inoltre, obbliga a creare un'organizzazione che sia in linea con questa volontà, che sia inserita nel contesto del coordinamento affidato al Ministero della cultura e della comunicazione e che si appoggi a partenariati pubblico-privati con gli editori o, a seconda dei casi, con le aziende che operano nel settore delle reti Internet.

Questa scelta strategica va a modificare le carte in tavola per le biblioteche francesi che, nell'ambito della negoziazione con i partner privati, si troveranno in una situazione più equilibrata. Infatti, potranno gestire con maggiore autonomia la propria politica di digitalizzazione, anziché dipendere da proposte esterne non sempre in linea con i loro obiettivi. Questa nuova autonomia dovrebbe anche consentire loro di gestire meglio la tempistica della digitalizzazione o, perlomeno, di non dipendere esclusivamente dall'agenda dei grandi operatori come Google.

La digitalizzazione lanciata dalla BnF con Gallica rappresenta, a questo proposito, un'esperienza unica e di grande valore, grazie al *know-how* acquisito in materia ed alla presenza di un volume non indifferente di patrimoni già digitalizzati. Le fasi iniziali sono state certo onerose, ma il sito consente oggi di accedere ad oltre 950.000 documenti, di cui

²⁴ Parere del G29 adottato il 4 aprile 2008, http://www.cnil.fr/fileadmin/documents/approfondir/dossier/internet/wp148_fr.pdf.

²⁵ [Su impulso del Presidente Nicolas Sarkozy alla fine del 2009 è stato presentato ed avviato dal governo francese il programma del Grande Prestito (Grand Emprunt) di 35 miliardi di euro, costituito da fondi pubblici e privati, allo scopo di realizzare cinque priorità nazionali: educazione superiore e formazione professionale, ricerca, industria e PME, sviluppo a lungo termine, digitale. Informazioni sul Grande Prestito sono disponibili sul sito del governo francese <http://www.gouvernement.fr/gouvernement/des-investissements> o sul sito <http://www.grandemprunt.net>.]

145.000 libri (vedi paragrafo 1.1.2.). Il ritmo di crescita è in linea con gli obiettivi strategici iniziali: da 5000 documenti digitalizzati nel 2007 si è passati a 100.000 documenti all'anno nel 2009. La mobilitazione di ingenti risorse finanziarie, resa possibile dalla decisione del Capo dello Stato, deve necessariamente accompagnarsi ad un cambio di scala e di ritmo.

Gli obiettivi di questa politica di digitalizzazione devono essere definiti chiaramente sin da ora. A nostro avviso, vi sono due obiettivi di ordine generale che devono essere messi nero su bianco.

Il primo obiettivo consiste nell'evitare il rischio della segmentazione del patrimonio, aspirando ad una digitalizzazione onnicomprensiva o, in ogni caso, quanto più vasta possibile, delle opere non coperte da diritti e delle opere protette da diritti. Il problema della scelta tra la digitalizzazione selettiva e la digitalizzazione di massa è stato risolto osservando gli utilizzi, che testimoniano la fiducia riposta dagli internauti nella "neutralità dei motori di ricerca" e nella capacità dei soggetti culturali di proporre gradualmente i propri contenuti sul Web. Buona parte delle richieste è rappresentata dall'interrogazione di ampie basi dati da parte di un motore di semplice utilizzo. La stragrande maggioranza degli internauti, ovvero il grande pubblico, non sta aspettando una "editorializzazione degli archivi", ma vuole piuttosto trovare ciò che le interessa. Anche se un giorno l'utilizzo semplice del motore di ricerca potrà essere combinato con altre modalità di accesso più strutturate, resta fatto che l'utilizzo più diffuso al giorno d'oggi è costituito dall'interrogazione tramite una richiesta semplice, basata su una o poche parole chiave. Una digitalizzazione esaustiva rappresenta quindi uno degli obiettivi da perseguire, in aggiunta alle strutturazioni documentarie destinate a pubblici più specializzati e realizzate dalle istituzioni pubbliche.

Questo obiettivo è in linea con la vocazione storica della BnF, preposta al deposito legale e che gestisce a questo titolo un patrimonio di riferimento. L'istituzione ha già iniziato a consi-

derare questo obiettivo ed ha avviato il processo di digitalizzazione, scaglionandone le fasi. Una volta identificati i settori prioritari e le tematiche principali, oggi la selezione dei documenti da digitalizzare si effettua soprattutto in base a criteri materiali (opere non coperte da diritti, pubblicate in Francia, condizioni e formato delle opere), in linea con l'obiettivo di una digitalizzazione di massa.

Considerando l'entità del patrimonio in questione (le collezioni della BnF comprendono circa 35 milioni di documenti di cui 11 o 12 milioni di libri; di questi libri, cinque milioni sono ormai di dominio pubblico), si tratta di un obiettivo a lungo termine, tanto più che gli "oggetti" da digitalizzare sono eterogenei e che solo una certa percentuale di libri può essere sottoposta ad una digitalizzazione di massa, allo stato attuale delle tecnologie disponibili. Sarebbe opportuno valutare l'impatto delle condizioni fisiche delle collezioni sulla digitalizzazione effettivamente portata a termine, affinché le opere in condizioni peggiori, spesso le più richieste, non siano assenti dalla biblioteca digitale.

Il secondo obiettivo si riferisce alla posizione del patrimonio francese librario in Internet. Tale patrimonio oggi è visibile soprattutto tramite Google Books, grazie ai patrimoni francofoni digitalizzati delle biblioteche straniere, che non sono completi. D'altro canto, il patrimonio di Gallica è difficilmente accessibile: l'internauta deve necessariamente entrare nel sito di Gallica. Sarà bene quindi fare in modo di non digitalizzare tanto per digitalizzare, ma garantire l'accesso a questo patrimonio digitalizzato, il che implica una riflessione a monte su come i documenti saranno reperiti, ovvero visibili, su Internet. Occorrerà considerare in maniera approfondita l'insieme delle risorse che consentono di ottenere questa visibilità digitale (referenziazione, indicizzazioni, menzioni nei blog o nei siti di *community*, ecc).

La strategia di digitalizzazione sarà quindi definita nell'ambito di un contesto profondamente diverso, su tre livelli.

In primis, l'esistenza di una piattaforma come quella di Gallica consente di appoggiarsi ad uno strumento esistente, ma le cui prestazioni sono ormai insufficienti e la cui dimensione collaborativa deve ancora essere definita, soprattutto per quanto riguarda gli editori e le biblioteche partner.

In secondo luogo, la costituzione di partenariati con soggetti privati (editori, motori di ricerca, piattaforme di diffusione, ecc.) rappresenta una delle condizioni preliminari se si vuole veramente mettere a disposizione il patrimonio digitalizzato su internet; è quindi opportuno definire la struttura di partenariati efficaci ed equilibrati che potrebbero essere costituiti con soggetti privati, che si tratti di Google o di altre società.

Infine, è necessario dar vita a un nuovo slancio europeo, in coordinamento con le altre biblioteche europee coinvolte ed appoggiandosi sul portale culturale comune Europea.

III.1 Uno strumento privilegiato da migliorare: Gallica

III.1.1 Aspetti istituzionali

Un bilancio mediocre

Un primo bilancio della situazione evidenzia i limiti dell'attuale organizzazione della BnF e di Gallica. Vi sono vari aspetti da segnalare che si riferiscono rispettivamente alle risorse, alla cooperazione con le altre biblioteche, alla collaborazione con gli altri soggetti, tra cui, prima di tutto gli editori, ed infine il coordinamento di Gallica.

- a. Per quanto concerne le risorse, il processo di digitalizzazione di massa ha consentito alla BnF di identificare una serie di difficoltà inerenti all'ambiente umano e tecnico. Infatti, questo processo richiede la mobilitazione di mezzi consistenti:
 - a monte, per identificare e selezionare i documenti, trattarli, indirizzarli al fornitore di servizi (tutte operazioni che non possono essere affidate ad un partner esterno) e

per produrre i metadati, soprattutto nel caso in cui i metadati esistenti non siano di qualità soddisfacente;

- a valle, per assicurare il "controllo qualità" dei documenti digitalizzati. Una volta ricevuti i file, la fase del controllo della qualità della digitalizzazione spetta, in linea di principio, alla biblioteca ed i flussi di lavoro potrebbero essere considerevoli. Allo stato attuale, tale controllo è in gran parte automatizzato, allo scopo di utilizzare meno risorse umane, ma tale soluzione non si è rivelata sufficientemente valida.

L'eccessiva mole di lavoro necessariamente associata al processo di digitalizzazione è un elemento fondamentale che deve essere preso in considerazione. Fino ad ora, la BnF si è appoggiata ad un organico praticamente costante, per far fronte alla digitalizzazione ed alle attività quotidiane della biblioteca. Il passaggio ad un ritmo di digitalizzazione più sostenuto riproporrà quindi la questione delle risorse umane dedicate a questo processo nell'ambito dell'organico dell'istituzione. La digitalizzazione è stata inoltre per lungo tempo considerata come un progetto supplementare rispetto agli obiettivi dell'istituzione. I diversi dipartimenti della biblioteca hanno iniziato a sentirla come un loro obiettivo solo di recente, e più precisamente nel 2007, anno in cui la BnF ha stipulato un contratto di "digitalizzazione di massa" per le sue collezioni di libri stampati. Questo "senso di proprietà" spiega in qualche modo la scelta di una relativa diluizione del team tra i suddetti dipartimenti, piuttosto che un servizio centralizzato in seno a Gallica. Al fine di raggiungere il nuovo obiettivo di digitalizzazione di massa ad un ritmo sostenuto, è assolutamente necessario operare una revisione dell'organizzazione, accompagnata da un ambizioso spiegamento di risorse umane dedicate alla filiera digitale.

- b. Per quanto riguarda la cooperazione con le altre biblioteche pubbliche, Gallica, nonostante abbia espresso il desiderio di ospitare altre collezioni oltre a quelle della BnF, deve riconoscere che le altre biblioteche collaborano solo in misura molto marginale, come si evince dall'esiguo numero di risorse provenienti dalle biblioteche partner (e cioè meno di 7.000 documenti a fronte dei 900.000 che il sito conta alla data odierna). Anche se si sta cominciando ad invitare queste biblioteche a partecipare al progetto Gallica, il processo decisionale e l'accesso agli stanziamenti per la digitalizzazione rimangono centralizzati nella BnF, il che non è certo un valido incentivo per le biblioteche partner.

La BnF propone la firma di accordi con le biblioteche interessate, ma esercita un controllo non indifferente e poco trasparente sulle scelte delle biblioteche, ovvero quali opere devono essere digitalizzate, in base ad una forma di collaborazione "verticale discendente". Il rapporto dell'Ispettorato Generale delle Finanze sulla BnF²⁶, pubblicato nel gennaio 2009 propone a tale proposito una maggiore condivisione della strategia di digitalizzazione con le altre istituzioni, soprattutto con la rete dei Pôles associés de partage documentaire, che ha potuto accedere solo in misura estremamente limitata agli stanziamenti della terza tranche del contratto relativo alla digitalizzazione di massa.

- c. Se la BnF deve fungere da attore principale nell'ambito del processo di digitalizzazione, occorre indubbiamente riflettere sulla modalità associativa degli altri partner, pubblici o privati, in termini di definizione degli obiettivi e delle opzioni tecniche e, più generalmente, in termini di armonizzazione delle scelte editoriali.

A titolo esemplificativo, non sarebbe illogico, al fine di semplificare la messa a punto di una filiera di digitalizzazione di massa, non solo delle opere non coperte da diritti, ma anche di quelle protette da diritti d'autore, affidare alla BnF l'incarico di digitalizzare anche quest'ultima categoria. Infatti, la legge del 1 agosto 2006 ha introdotto nel *Codice del patrimonio culturale* (articoli L. 132-4, L. 132-5 e L. 132-6) il diritto dell'organo preposto al deposito legale di digitalizzare, a scopo di conservazione e consultazione in sito, il patrimonio da essa posseduto, senza necessità di ottenere l'autorizzazione preliminare dei depositanti. Grazie a questo incarico, la BnF potrebbe digitalizzare in massa le opere che custodisce, comprese quelle coperte da diritti, opere di cui possiede anche i rispettivi metadati.

Resta inteso che la messa a disposizione di tale patrimonio digitalizzato è concepibile solo in un contesto contrattuale ben definito con gli editori ed i rappresentanti degli aventi diritto, dato che le disposizioni legali limitano in maniera molto restrittiva la consultazione delle opere in deposito legale.

L'insieme così creato costituirebbe una base digitalizzata considerevole, molto interessante da sfruttare, soprattutto per quanto concerne le opere esaurite. Le opere che gli editori non intendono necessariamente ripubblicare in formato cartaceo potrebbero così essere oggetto di un nuovo utilizzo, remunerato e non esclusivo, in formato digitale²⁷. Ciò potrebbe comportare un consistente allargamento della base consultabile dal pubblico. Pertanto, la digitalizzazione del patrimonio di libri della BnF, coperti da diritti e non, troverebbe una giustificazione logica, ma, allo stesso tempo, legale, industriale e commerciale.

²⁶ Ispettorato generale delle Finanze, *Rapport sur la Bibliothèque nationale de France*, n. 2008-M-065-02, gennaio 2009; vedi pagina 14 del rapporto.

²⁷ Tale accesso potrebbe essere monetizzabile, per esempio tramite l'indirizzamento al sito dell'editore, o in seno alle biblioteche stesse, sotto forma di abbonamento.

In quest'ottica, Gallica diverrebbe un sito di accesso a tutto il patrimonio librario, attraverso una piattaforma cooperativa che tiene conto dei diritti dei diversi partner, con condizioni di accesso diverse a seconda dello *status* di ogni opera.

Secondo il Gruppo di lavoro, la posizione di Gallica ed il suo ruolo nei confronti degli altri partner deve essere ridefinita, sia sul piano tecnico che sul piano operativo. La collaborazione estesa ed efficace dei partner al coordinamento di Gallica è una condizione essenziale per accelerare il processo di digitalizzazione e divulgazione digitale.

Verso una piattaforma cooperativa di valorizzazione dei patrimoni e delle opere digitalizzate

Innanzitutto, va detto che Gallica nasce come piattaforma di riferimento, non esclusiva, che consente l'accesso del pubblico al patrimonio digitalizzato dell'insieme dei partner. Le biblioteche e gli editori che lo volessero, potrebbero rivolgersi a Gallica, senza che ciò limiti la loro possibilità di divulgare i loro contenuti digitalizzati attraverso qualsiasi sito da loro scelto, sulla scia della situazione esistente. In compenso, l'accesso al finanziamento pubblico per la digitalizzazione dei libri dovrebbe essere subordinato all'adesione a Gallica, vale a dire ad almeno una delle sue funzionalità e cioè: indicizzazione del contenuto, consultazione degli estratti o, a seconda dei casi, commercializzazione dei file, diretta (su Gallica) o indiretta (tramite l'indirizzamento da Gallica ad un sito esterno di vendita, scelto dall'editore che detiene i diritti). Gallica si propone quindi come piattaforma di divulgazione di riferimento, con il vantaggio di una maggiore visibilità sul Web, grazie all'importanza dei contenuti ed alla reciprocizzazione delle risorse.

a. È in quest'ottica che si potrebbe costituire

il partenariato con gli editori ed i rappresentanti degli aventi diritto, al fine di migliorare la presenza del *corpus* francofono su Internet.

La BnF potrebbe, come abbiamo visto, procedere alla digitalizzazione di massa delle opere raccolte a titolo di deposito legale. Anche se la BnF è in possesso della legittimità giuridica per agire in tal senso a fini di conservazione, e quindi per accedere localmente a tale patrimonio, il dialogo con gli editori rimane comunque un aspetto indispensabile a vari livelli. Gli editori rimarrebbero liberi di perseguire le proprie operazioni di digitalizzazione, che attualmente riguardano principalmente le opere recenti. Inoltre, gli editori potrebbero condurre operazioni di digitalizzazione più accurate oppure applicare norme più restrittive al loro patrimonio, a seconda di quello che intendono farne, ovviamente, nel loro ambito di spettanza²⁸.

Il dialogo da instaurare deve includere anche le tematiche relative alle modalità di accesso al contenuto, fermo restando che i libri coperti da diritti non possono essere oggetto di accesso gratuito tramite Gallica, se non previa autorizzazione dell'editore.

Oltre all'opportunità giuridica offerta dalla legge del 1° agosto 2006 in materia di digitalizzazione e di messa a disposizione dei patrimoni digitalizzati, sarà necessario attuare una fase contrattuale, in cui la digitalizzazione dei libri coperti da diritti sarà preceduta da un dibattito con gli editori in merito alla suddivisione dei compiti e degli oneri a carico delle parti, a seconda degli utilizzi e delle modalità di sfruttamento stabilite.

Infine, in virtù del suo posizionamento al confine tra il settore del patrimonio culturale ed il settore dei diritti d'autore, Gallica andrebbe a ricoprire il ruolo di piattaforma di divulgazione e valorizzazione delle opere

²⁸ E riceverebbero un finanziamento specifico, oltre ai fondi del Centro nazionale del Libro per i progetti di digitalizzazione "accurata".

orfane, a partire dal momento in cui il codice della proprietà intellettuale consentirà lo sfruttamento digitale di tali documenti. Infatti, una raccomandazione del Consiglio superiore della proprietà letteraria ed artistica dell'aprile 2008 suggerisce la messa a punto di un sistema di gestione collettiva obbligatoria delle opere orfane (scritti e immagini fisse), allo scopo di renderle più facilmente disponibili per altri utilizzi, principalmente sul Web. Il 30 settembre 2009, il Ministro della Cultura ha incaricato il suo staff di proporre un seguito legislativo a questa raccomandazione.

Affinchè Gallica possa diventare il centro di un'efficace divulgazione di questi libri protetti da diritti, ma i cui rispettivi titolari non sono stati identificati, sarebbe auspicabile che gli editori prendessero parte alla gestione di Gallica, proprio per garantire il rispetto dei diritti associati a tali opere.

- b. Per quanto concerne le biblioteche, il processo della digitalizzazione in sè (scelta delle opere, volume della digitalizzazione, ecc.) potrebbe rimanere una competenza delle biblioteche partner.

Tuttavia, la BnF dovrà nel frattempo proporre a chi lo desidera soluzioni comuni, offrendo a questi soggetti l'opportunità di partecipare ai suoi contratti di digitalizzazione, sulla scia, per esempio, della sperimentazione in corso con le biblioteche di Alençon, Compiègne, Lione, Cujas e dell'Istituto nazionale di storia dell'arte (tre biblioteche municipali e due biblioteche universitarie) che ha consentito a queste biblioteche di usufruire di un contratto della BnF, con la decentralizzazione di alcuni aspetti del processo di digitalizzazione. In questo modo, si potrebbero condividere le attività della digitalizzazione e soprattutto quelle relative all'immagazzinamento dei file digitali, senza che ciò impedisca alle biblioteche associate di intraprendere eventualmente tali attività in autonomia.

Che tipo di governance?

L'organizzazione da prendere in considerazione deve inserirsi nella struttura che sarà definita dal Ministero della cultura e della comunicazione per coordinare il processo di digitalizzazione del patrimonio culturale e, più generalmente, deve tener conto del ruolo del Commissariato incaricato della gestione del Grande Prestito che dovrà anche coordinare l'insieme delle azioni.

Le proposte del Gruppo di lavoro comportano la valutazione di diversi livelli di azione:

- la digitalizzazione dei libri (scelta delle opere, gestione della filiera della digitalizzazione, stipula dei necessari contratti di prestazione di servizi, produzione di file, ecc.);
- l'immagazzinamento e la manutenzione di tali file;
- la messa a disposizione degli internauti (concezione della piattaforma, scelta delle tecniche di indicizzazione ecc);
- gli sfruttamenti commerciali veri e propri, gratuiti o a pagamento, diretti o indiretti;
- i partneriati con gli operatori in grado di sviluppare l'accesso alle opere nazionali.

a. Spetterà alla struttura di coordinamento organizzata dal Ministero della cultura e della comunicazione definire la portata ed i finanziamenti associati ai primi due livelli sopra descritti: il processo di digitalizzazione e di immagazzinamento.

b. Il Gruppo di lavoro suggerisce peraltro che le scelte dei formati e delle specifiche tecniche, nonché delle procedure e dei partneriati, siano effettuate in seno alla struttura di riferimento riorganizzata, e cioè Gallica, con riserva di approvazione della struttura di coordinamento del Ministero e, a seconda dei casi, del Commissariato incaricato della gestione del Grande Prestito.

In questo modo, l'insieme degli aspetti legati ad una necessaria logica di armo-

nizzazione sarebbe trattato a livello di Gallica²⁹:

- l’armonizzazione dei metadati associati ai documenti, che rappresentano un veicolo di informazioni essenziali ai fini della qualità della ricerca, la referenziazione e l’accesso all’informazione;
- l’interoperabilità dei file e più generalmente dei formati dei dati scambiati (metadati e contenuti), nello specifico, in particolare, tramite la definizione di standard comuni di digitalizzazione, di divulgazione e di immagazzinamento digitale;
- la gestione dell’interfaccia di consultazione, del motore e della referenziazione;
- la struttura dei contratti relativi allo sviluppo ed al funzionamento della piattaforma.

Una delle principali missioni di Gallica, in questo contesto rinnovato, sarà quello di definire le nuove esigenze al fine di adattarsi al cambiamento di scala della digitalizzazione. Questo cambiamento di scala richiederà infatti una profonda trasformazione del funzionamento della piattaforma. I costi associati a tale cambiamento devono essere considerati separatamente dai costi della digitalizzazione vera e propria (nuove infrastrutture, soprattutto per quanto riguarda l’immagazzinamento, sviluppo di nuove funzionalità, ecc.). Gallica potrà poi predisporre una pratica di richiesta di finanziamento e presentarla alla struttura di coordinamento dei costi organizzata dal Ministero.

Sul piano strategico, il Gruppo di lavoro ritiene necessario che Gallica conservi un ruolo centrale per quanto concerne la gestione degli eventuali accordi con i partner privati, poichè la logica di tali accordi si riferisce più all’interesse collettivo che non all’interesse dei singoli partner di Gallica. Inoltre, è necessario che vi sia un unico portavoce, meglio se un interlocutore

avente un certo peso. Gallica potrebbe essere così univocamente identificata in tale ruolo, godendo di una posizione privilegiata nell’ambito delle trattative e di un *know-how* reciproco in materia di partenariato. Questa funzione “motrice” è però fattibile solo se anche la gestione di Gallica viene ridefinita, con un coinvolgimento molto più attivo dei suoi partner, primi fra i quali biblioteche ed editori.

La questione si pone al momento di decidere se Gallica debba rimanere integrata nella BnF, o altrimenti se sia meglio creare una nuova entità, sotto forma, per esempio, di un raggruppamento pubblico a vocazione culturale.

Indipendentemente dalla soluzione prescelta, il ruolo della BnF rimarrà comunque importante dal momento che sarà un partner privilegiato.

- c. Il Gruppo di lavoro raccomanda che Gallica assuma in ogni caso una maggiore autonomia, in modo da diventare una struttura dedicata, che lavora per partner diversificati ed incaricata di definire (di concerto con gli altri partner) le scelte tecniche importanti, di garantire la diffusione dei contenuti e di gestire i finanziamenti. Questa struttura opererebbe sotto la guida di un organo composto dai diversi partner della piattaforma. Questo organo collegiale avrebbe poteri decisionali.

L’organo di coordinamento di Gallica avrebbe in particolare l’incarico di definire le principali evoluzioni della piattaforma, delle norme e degli standard tecnici relativi alla messa a disposizione dei patrimoni, di elaborare una politica di visibilità del patrimonio digitalizzato, di valutare l’interfaccia di consultazione, il motore, e più generalmente, il miglioramento delle funzionalità di Gallica, allo scopo di rispondere meglio agli utilizzi e trovare un accordo sull’“esperienza” proposta all’utente. Dovrebbe an-

²⁹ A tale proposito, sono stati definiti diversi statuti (statuto documentario, statuto della digitalizzazione, statuto OCR). Sarà necessario portare avanti queste riflessioni condivise.

che, a seconda dei casi, concepire una politica di valorizzazione commerciale, con l'accordo degli organismi o dei partner proprietari dei file.

Il coordinamento generale tra i vari partner coinvolti nel processo di digitalizzazione sarebbe assicurato dalla struttura costituita in seno al Ministero della cultura e della comunicazione.

III.1.2 Rinforzare la presenza di Gallica e dei suoi contenuti sul Web

Le attività di digitalizzazione devono essere accompagnate da una volontà di "conquistare" una maggiore visibilità sul Web. Non serve a nulla essere disponibili se non si è visibili³⁰.

A tale proposito, si potrebbe innanzitutto cambiare il nome di Gallica, approfittando del cambiamento di dimensione della politica di digitalizzazione.

Se Gallica è facilmente accessibile tramite i motori di ricerca, cercando il nome "Gallica" (il link al sito è visualizzato come prima risposta in tutti i motori di ricerca utilizzati), lo stesso non accade quando il motore di ricerca viene interrogato inserendo un titolo o il nome di un autore. Per esempio, se si cerca "le Rouge et le Noir", presente nelle collezioni digitali non coperte da diritti di Gallica, non appare alcun risultato proveniente da Gallica. La prima occorrenza dell'opera digitalizzata proviene da Google Books, che propone l'accesso alla copia digitalizzata della biblioteca dell'Università della California (edizione del 1866). Una ricerca in cui si associano i termini "le Rouge et le Noir" e "Gallica" non genera risultati provenienti dal sito, ma consente un accesso indiretto allo stesso, tramite alcuni link dai blog degli internauti, il che dimostra che il sito è stato rintracciato.

La politica di visibilità del sito e di "disseminazione" dei suoi contenuti non protetti da diritti d'autore su Internet è cominciata, ma deve essere rinforzata affinché tali contenuti possano

essere reperiti più efficacemente dai motori di ricerca.

Referenziazione e indicizzazione

Il personale della BnF ha recentemente condotto una riflessione approfondita su come incrementare la notorietà di Gallica, individuando a tal fine tre linee di azione: la moltiplicazione degli accessi, a partire dalla stessa piattaforma, a contenuti diversificati (la cosiddetta strategia dei "link avanzati"), il miglioramento della segnalazione e della referenziazione ed un miglior accesso dei metadati e dell'indicizzazione dell'insieme dei contenuti (indicizzazione "a tutto testo"), ottenuto dai motori di ricerca.

Questa politica, che si basa sia su una migliore referenziazione del sito che su un'indicizzazione dei contenuti tale da renderli facilmente accessibili tramite le richieste nei motori di ricerca, deve essere perseguita e approfondita.

I motori di ricerca non possono accedere a pagine con contenuto dinamico che si creano solo su richiesta dell'internauta, per mezzo di un'interfaccia di ricerca specifica di Gallica. Per far sì che le sue risorse siano maggiormente visibili in seno all'universo digitale, Gallica può avvalersi di una serie di tecniche tra cui: l'attenta scelta del nome da attribuire alle pagine, la personalizzazione delle URL, il miglioramento della referenziazione naturale, la concatenazione delle pagine, la creazione di "pagine di approdo", l'utilizzazione dei metadati, oppure l'acquisto di parole chiave.

Tali azioni devono consentire sia l'indicizzazione dei contenuti da parte di questi motori di ricerca che la consultazione completa dell'opera tramite un link che indirizza al sito di Gallica.

Nuove funzionalità, maggiormente orientate agli aspetti partecipativi e di community

Gli aspetti partecipativi ("wiki") non devono essere sottovalutati. Infatti, rappresentano un arricchimento del sito, in particolare dei meta-

³⁰ Si rimanda nuovamente all'Allegato 3, in cui si opera un confronto dettagliato tra le funzionalità offerte da Gallica e quelle proposte da Google Books.

dati, consentono lo sviluppo di *community* in seno a Gallica, oltre ad una maggiore notorietà del sito.

Gli internauti possono contribuire ad arricchire una risorsa (per esempio con una fotografia o un'opera) tramite link, risorse associate (per esempio tesi sull'opera o sull'autore) o commenti. Anche in questo caso, le competenze dei bibliotecari si possono espandere nel territorio del digitale, con la possibile creazione di etichette (per esempio, nel caso delle biblioteche fisiche, le cosiddette "opere di consultazione"), finalizzate ad evidenziare determinate opere. Anche lo sviluppo e l'armonizzazione dei metadati possono essere rinforzati, adottando un approccio di natura semantica, anch'esso completato dagli internauti.

Gallica potrebbe anche offrire alle *community* (di interessi comuni o di ricerca) funzioni di classificazione, suddivisione, arricchimento. Sebbene i tentativi di organizzazione delle *community* presenti sul Web da parte delle istituzioni siano spesso votati al fallimento, la proposta regolare di nuove funzionalità *user-friendly*, ma innovative, potrebbe consentire a Gallica di assolvere, nel mondo digitale, la sua missione di accoglienza degli appassionati e dei ricercatori.

III.1.3 Migliorare il servizio offerto da Gallica *Un motore migliorato*

Gallica sta attualmente considerando il miglioramento del suo motore di ricerca. Il motore utilizzato attualmente (Lucene) è dedicato alla ricerca a pieno testo: opera su un indice preconstituito ed unico e non interagisce con nessun'altra base di dati. Se la scelta del motore si baserà sui volumi trattati, il motore dovrà essere ripensato radicalmente nella prospettiva di un accesso di massa³¹.

Da una parte, se Gallica intende adattare i suoi strumenti documentari, il motore deve evolversi con l'aggiunta di strumenti di analisi semantica

più avanzati. D'altra parte, una riflessione sull'ergonomia della ricerca deve essere ulteriormente approfondita, al fine di definire con maggiore precisione i parametri del motore in base alle richieste in termini di utilizzo. In altre parole, l'ergonomia deve essere rivista sia dal punto di vista dell'internauta che vorrebbe effettuare una ricerca rapida "in un clic", che dal punto di vista del ricercatore che desidera affinare la ricerca, con l'ausilio di strumenti di smistamento, filtro, posizionamento in un campo specifico di ricerca. Quest'ultimo elemento rappresenta un aspetto distintivo rispetto a Google Books. Ma in tutto ciò, la facilità di accesso al contenuto non dovrà essere assolutamente tralasciata.

Il cambiamento di scala della digitalizzazione rende questa problematica più urgente: sembrerebbe che le attuali scelte tecnologiche, soprattutto in materia di motore, non siano in grado di accogliere questa evoluzione. Questo aspetto è di importanza strategica, poiché le prestazioni del motore rappresentano una condizione essenziale per convivere con piattaforme come Google Books.

A tal fine, si potrebbe pensare ad un potenziamento del motore attuale o all'adozione di un altro motore di ricerca più efficiente.

Lo staff tecnico di Gallica ha già classificato questo progetto come prioritario.

La standardizzazione dei metadati

I metadati sono essenziali per accedere alle informazioni sui contenuti digitalizzati. Pertanto, è indispensabile evitare errori e colmare eventuali lacune³². Oggi, le biblioteche digitali come Gallica e Google Books utilizzano il medesimo standard internazionale di base (il Dublin Core), che garantisce l'interoperabilità dei metadati provenienti da diverse risorse documentarie.

Ma l'utilizzo di uno standard omogeneo non è una garanzia di qualità. La rigidità di questo

³¹ Vedi Allegato 3, pag. 55 e successive.

³² Vedi Allegato 3, da pag. 53 a pag. 55.

standard potrebbe di fatto incoraggiare le biblioteche digitali a semplificare i metadati di cui dispongono o a raggruppare in un'unica categoria ("campo") i metadati strutturati in modo più avanzato. Lo standard non può d'altronde, risolvere i difetti che potrebbero esserci nei metadati iniziali.

Questi metadati iniziali sono strutturati per campo ("autore", "titolo", "editore", "fonte", ecc.). Se, per errore o per la preoccupazione di semplificare, alcuni elementi figurano nel campo sbagliato (per esempio se il nome dell'autore appare nel campo "titolo"), l'inadeguatezza del contenuto e del campo che lo contiene renderà meno pertinente i risultati della ricerca. Questi errori pregiudicano la creazione di link tra i libri (libri dello stesso autore, per esempio) e la qualità delle azioni che possono essere condotte sui risultati della ricerca (classificazione o filtro per autore).

Poichè le fonti di informazione delle biblioteche digitali sono multiple ed eterogenee, la standardizzazione dei metadati rimane un aspetto problematico ai fini della digitalizzazione di massa. Questa standardizzazione richiede l'utilizzo di Authority file unificati, il reindirizzamento delle notizie bibliografiche e l'esercizio di un controllo qualità approfondito sui metadati associati ai patrimoni digitalizzati.

In tutto ciò, sono in gioco la pertinenza dei risultati della ricerca per l'internauta, la possibilità di classificare le informazioni ricercate (che sono collegate alla migliore pertinenza semantica dei motori di ricerca), ma anche lo sviluppo del "Web semantico". Queste poste in gioco dipendono totalmente dalla portata delle attività degli istituti preposti alla conservazione ed alla valorizzazione dei patrimoni, che devono necessariamente partecipare.

III.2 Le condizioni per un partenariato equilibrato con soggetti privati

I precedenti sviluppi non devono far pensare che le grandi istituzioni pubbliche potrebbero condurre la loro politica di digitalizzazione senza costituire dei partenariati con soggetti pri-

vati che operano sul Web. Questi partenariati potrebbero produrre risultati estremamente positivi, non solo sul piano tecnologico, ma, nell'ambito di un settore in evoluzione costante e rapida, anche in termini di massa critica di contenuti, soprattutto se tale partenariato fosse costituito con un soggetto che abbia già condotto iniziative importanti in materia di digitalizzazione dei contenuti.

Questi partenariati dovrebbero anche rispettare un certo numero di condizioni essenziali e garantire allo stesso tempo equilibrio e reciprocità tra le parti. La BnF deve svolgere un ruolo esemplare a tale proposito, legato soprattutto al compito di ricettore del deposito legale.

III.2.1 Obiettivi e condizioni preliminari

Obiettivi

L'obiettivo principale è quello di garantire *un'ampia visibilità del nostro patrimonio su Internet e la sua identità*, ricorrendo ai metodi di indicizzazione e referenziazione specifici dell'universo digitale, ma, allo stesso tempo, sfruttando le complementarità del contenuto, al fine di ottenere un effetto di massa critica nell'universo francofono online.

Il primo aspetto presuppone di non scartare partenariati con soggetti come i motori di ricerca, tra cui ovviamente Google, ma anche Bing, il motore sviluppato da Microsoft, oppure Yahoo! Le nuove generazioni di motori sono orientate verso la presentazione di contenuti sempre più strutturati. Una piattaforma di contenuti ordinati come Gallica potrebbe pertanto diventare un partner molto attraente per questi motori.

Il secondo aspetto interesserà maggiormente i soggetti che hanno già digitalizzato contenuti, in primis gli editori, ma anche Google Books. Il partenariato con gli editori dovrebbe basarsi, come già detto, su una cooperazione rinforzata. Il presente rapporto è incentrato principalmente su questa cooperazione con piattaforme di contenuti di tipo Google Books, ma senza escludere gli altri aspetti.

L'altro obiettivo è *la messa a disposizione più*

ampia possibile delle opere di dominio pubblico che saranno digitalizzate. Poiché queste opere sono di pubblico dominio, il loro contenuto può essere, in qualche modo, considerato come una “infrastruttura essenziale” ai fini del diritto della concorrenza. In tal senso, l’utilizzo dei dati digitali di queste opere non deve essere vincolato da clausole di esclusività, trattandosi del *Web-harvesting* da parte dei diversi robot dei motori esistenti o degli utilizzi delle biblioteche depositarie di questi contenuti non coperti da diritti.

Condizioni preliminari

L’eventuale contraente di Gallica dovrà soddisfare almeno due condizioni preliminari.

La prima è il rispetto del diritto d’autore, essenziale nell’ottica dell’osservanza della legge francese e della lealtà nei confronti dei diversi partner della piattaforma Gallica.

La seconda condizione si riferisce alle modalità in base alle quali il contraente gestirà le informazioni ed i dati personali degli internauti. Riteniamo essenziale definire in anticipo le condizioni (soprattutto per quanto concerne la durata), in base alle quali tali dati saranno conservati e utilizzati. Il tipo di consultazione ed, eventualmente, di lettura, da parte dell’internauta rappresenta un dato sensibile, che impone l’applicazione a priori di determinate precauzioni, in materia di non divulgazione e conservazione limitata dei dati.

III.2.2 “Un libro in cambio di un libro”: una proposta di partenariato basata sullo scambio dei dati digitalizzati

I principi

Gli accordi sinora intercorsi o proposti alle biblioteche in materia di digitalizzazione di massa, principalmente dall’attore principale del settore, e cioè Google, consistevano nell’affidare la digitalizzazione (scansione e “ocerizzazione” delle opere) al soggetto privato, il quale doveva poi mettere a disposizione della biblioteca una copia del file digitalizzato, tuttavia con alcune restrizioni di utilizzo di troppo lunga durata.

Per evitare che le opere di dominio pubblico siano soggette a clausole di esclusività come queste e per consentire alla biblioteca di essere proprietaria a pieno titolo dei suoi file digitali, sarebbe necessario che queste operazioni siano effettuate o finanziate dalla biblioteca stessa. Pertanto, un partenariato con un soggetto che già dispone di contenuti digitalizzati si deve fondare su un terreno diverso.

Un accordo con Google, per esempio, o più precisamente con la sua piattaforma di contenuti, Google Books, potrebbe mirare, non tanto ad assumere lo sforzo della digitalizzazione, quanto a ripartirlo, attuando uno scambio di file di qualità equivalente e di formati compatibili (testi ed immagini).

Le proporzioni di questi possibili scambi possono essere definite, allo scopo di evitare ridondanze, ma anche di poter sfruttare le complementarità (per esempio, edizioni diverse della stessa opera).

Questo consentirebbe a Gallica di raggiungere più rapidamente la massa critica necessaria in lingua francese (Google Books dispone di contenuti francofoni digitalizzati derivati dal patrimonio delle biblioteche americane con le quali ha stipulato accordi ed anche di contenuti francofoni provenienti da biblioteche europee come Gand, Losanna e presto anche Lione). Dal canto suo, Google Books potrebbe arricchire la sua base di contenuti dalle collezioni di *reference* presenti in Gallica. Questo scambio potrebbe essere ulteriormente integrato dal miglioramento reciproco dei metadati e delle rispettive notizie bibliografiche. Infine, l’accordo potrebbe contemplare l’obbligo di citare la provenienza del file nel sito consultato, per esempio tramite la visualizzazione di un logo o di un’apposita dicitura.

Ogni partner potrebbe disporre liberamente dei file ottenuti tramite il suddetto scambio, sulla base di condizioni trasparenti e definite in anticipo.

Equilibrio economico del progetto

Questo aspetto susciterà sicuramente le opinioni più divergenti tra loro per una serie di

motivi: in primo luogo, non sarà possibile far coincidere gli obiettivi di tutte le parti, poi, l'elaborazione di proiezioni finanziarie sicure e trasparenti si potrebbe scontrare con il desiderio del partner di non divulgare le informazioni relative alle proprie fonti di reddito ed, infine, le valorizzazioni praticabili a medio termine non sono sempre facilmente identificabili.

Se questo dovesse rivelarsi necessario, si potrebbe procedere per fasi successive:

- 1^a fase: un accordo sugli standard dei file da scambiare e una valutazione della qualità e della compatibilità dei file su vasta scala;
- 2^a fase: uno scambio di file basato su un determinato numero di documenti (per esempio, nell'ordine di 100.000 opere) ed una valutazione delle consultazioni effettive degli internauti;
- 3^a fase: uno scambio su una base più ampia di file ed eventualmente di *corpus* complementari, a seconda dei casi, accompagnato da un accordo sulle condizioni di sfruttamento commerciale dei file scambiati.

Proposta alternativa

Tuttavia, non va sottovalutata la complessità che lo scambio di file puro e semplice potrebbe comportare sul piano tecnico: la definizione dei tipi di file da consegnare, con eventuale conversione se i file iniziali non corrispondono, la pesantezza ed il costo della conversione (in formato PDF per esempio). Nel caso in cui tale scambio di file non fosse realizzabile facilmente, a causa per esempio di incompatibilità tra i file da scambiare e la piattaforma di destinazione o l'utilizzo che la biblioteca vorrebbe farne, si potrebbe concepire un altro tipo di partenariato, che avrebbe il vantaggio di rispondere all'obiettivo di visibilità dei patrimoni digitali.

I motori di ricerca vedono le pagine dei libri come pagine "Web" che devono quindi necessariamente essere strutturate in maniera da poter essere facilmente indicizzate da questi motori. Il processo di digitalizzazione che Google adot-

ta integra questa strutturazione specifica. Potrebbe essere interessante valutare la creazione di una filiera di digitalizzazione condivisa, una sorta di joint-venture che consentirebbe sia a Google che alla biblioteca di condurre le proprie operazioni di digitalizzazione allo stesso momento, ciascuno con il proprio procedimento, ma condividendo il personale e mobilitando solo una volta le opere da digitalizzare. Nell'ambito di una filiera comune, si potrebbe pensare di condividere totalmente il laboratorio di digitalizzazione, per scansionare i documenti solo una volta. In altre parole, la digitalizzazione "immagine" potrebbe essere condivisa completamente. I processi di "ocerizzazione", invece, potrebbero restare distinti, affinché ogni partner possa scegliere il livello di qualità e le funzionalità necessarie per un'integrazione fluida nella propria piattaforma.

Certo le opere verrebbero "ocerizzate" due volte, ma i costi sarebbero ripartiti e le biblioteche avrebbero la certezza che i file digitalizzati da Google sarebbero trattati in modo da poter essere correttamente indicizzati. Una corretta indicizzazione aiuta a rintracciare un'opera sul Web, ovvero a renderla accessibile e visibile. Questa maggiore visibilità, ottenuta quasi gratuitamente, è da mettere in conto a fronte dei costi relativamente elevati di una politica dinamica di messa in linea efficace sul Web. Infine, ognuna delle parti manterrebbe il controllo dei rispettivi file. Sarebbe opportuno valutare l'apporto iniziale di ogni soggetto (Google o la biblioteca) e stabilire, a seconda del numero di documenti forniti da ognuna delle parti, gli eventuali compensi.

Un partenariato in linea con gli obiettivi delle parti

Una società come Google potrebbe essere interessata ad un partenariato di siffatta natura, nel contesto dell'ampliamento del programma intrapreso con numerose istituzioni patrimoniali, negli Stati Uniti ed in Europa? E parimenti, la messa in linea di fondi letterari francesi richiede necessariamente una referenziazione

come elemento di Google Books? E ancora, un accordo di questo tipo potrebbe contribuire ad accentuare lo squilibrio a vantaggio di Google ed a scapito dei suoi concorrenti?

Intanto, questo tipo di accordo risponderebbe agli obiettivi ufficiali di Google e cioè mettere a disposizione degli internauti un accesso semplificato ed omogeneo alla maggioranza delle opere digitalizzate, senza che ciò comporti un costo diretto per la società, a differenza degli accordi attualmente in essere con le biblioteche che sono onerosi per Google. Inoltre, questo accordo autorizzerebbe lo sfruttamento commerciale, una volta stabilite le regole di ripartizione dei profitti in base allo *status* delle opere, e migliorerebbe la rilevanza dell'offerta. Infine, potrebbe garantire l'accesso a metadati ed notizie bibliografiche preziose, il che andrebbe a compensare i punti deboli di Google Books in materia.

Questo tipo di partenariato è ugualmente interessante anche per la visibilità dei libri non anglofoni sul Web. La migliore offerta digitale e la moltiplicazione dei possibili accessi a tali contenuti (direttamente, tramite Gallica o attraverso un motore di ricerca), risponde all'obiettivo principale di divulgazione del patrimonio scritto digitalizzato. E questo non rappresenterebbe per ogni partner di Gallica una valorizzazione ulteriore delle proprie attività di digitalizzazione ed una maggiore visibilità dell'eventuale sito di tale partner?

Resta il fatto che un accordo richiede l'incontro di due volontà e che non si può prefigurare il futuro.

III.3 Trovare un nuovo slancio di livello europeo

In Europa, un certo numero di biblioteche con forma giuridica diversa hanno concluso accordi con Google (vedi Allegato 4), altre sono interessate a sottoscrivere un partenariato, di solito per usufruire dei finanziamenti necessari ad una digitalizzazione di massa, ma anche spinte dal desiderio di essere efficacemente presenti su Internet, e ciò sia per quanto riguarda le

opere redatte in lingue molto diffuse che per lingue meno comuni, ma referenziate da Google.

D'altra parte, la realizzazione dei progetti francesi di digitalizzazione del patrimonio scritto rischia di creare uno squilibrio in Europa, anche se esistono gli strumenti (soprattutto Europea) e le concertazioni sono spesso in stretto rapporto.

Il Gruppo di lavoro ha consultato vari dirigenti di altre biblioteche europee e ha fatto il possibile per analizzare approfonditamente le poste in gioco e per comprendere sia i dubbi che il reale livello di coinvolgimento dei nostri partner. Ha ritenuto indispensabile sviluppare le basi di una cooperazione attiva, in grado di fare da contrappeso alle iniziative dei grandi attori privati del settore.

La sfida è triplice:

- come finanziare una digitalizzazione di massa in Europa?
- come procedere, cioè sulla base di quali principi, quali tecnologie e quali standard?
- Europea è un'alternativa credibile, di qualità potenzialmente equivalente a Google Books ?

Tutte domande a cui deve rispondere una politica di livello europeo, che potrebbe appoggiarsi su tre linee di azione: la reciprocità delle azioni delle biblioteche, il rilancio di Europea e la definizione di una Carta comune per i partenariati pubblico/privato.

III.3.1 Reciprocità e sinergia nelle azioni delle biblioteche

Le basi su cui fondare un'azione europea

Per quanto concerne le esperienze "individuali" delle biblioteche, innanzi tutto, molte di esse hanno già intrapreso attività di digitalizzazione e sviluppato un loro know-how.

Per quanto riguarda, invece, la volontà di agire congiuntamente, si può sottolineare che siamo davanti alla reale presa di coscienza del valore

della digitalizzazione dei patrimoni, che molte biblioteche già partecipano a progetti e programmi comunitari (come il PCRD, per menzionarne uno tra i programmi relativamente datato, oppure il più recente progetto ARROW, sostenuto dal programma europeo e-Contentplus per le opere orfane, per esempio), e che infine sta nascendo un interesse condiviso per i partenariati pubblico-privati come dimostra il progetto di una Carta delle biblioteche pubbliche che disciplini le condizioni dei partenariati pubblico/privato (sul quale sta lavorando la Conference of European National Librarians – CENL).

La creazione di una rete di biblioteche volontarie

Alla guisa delle iniziative intraprese nel settore dei media, si potrebbe creare una rete di biblioteche che adottino volontariamente un processo comune di digitalizzazione. Questo potrebbe implicare, a più livelli:

- la realizzazione, per chi lo vuole, di centri comuni di digitalizzazione, con l’obiettivo di armonizzare le metodologie ed i file, al fine di renderli interoperabili e di dividere i costi della filiera della digitalizzazione, mettendo in comune competenze e risorse. In questo modo, le biblioteche potrebbero sfruttare i benefici di un’alternativa alle proposte di Google, senza essere costrette ad accettare clausole di esclusività;
- l’eventuale creazione di centri comuni di immagazzinamento digitale, che anche in questo caso permetterebbe la ripartizione dei costi;
- lo scambio dei file digitali tra le biblioteche di questa rete. In quest’ambito, Gallica potrebbe proporre ai partner che lo desiderano di ospitare e divulgare i loro file digitali, estendendo anche alle biblioteche europee una parte di quanto offerto ai partner francesi. La provenienza dei contenuti sarebbe identificata sistematicamente, sia per mezzo di loghi che tramite un’interfaccia adattata alla consultazione dei fondi stranieri,

tale da rendere trasparente la diffusione tramite Gallica. In altre parole, Gallica ospiterebbe in questo caso i fondi, che però sarebbero consultabili solo nel loro ambiente specifico, scelto dalla biblioteca da cui provengono, e questo, senza esclusività. I grandi patrimoni stranieri conservati presso la BnF potrebbero incentivare la collaborazione di altre biblioteche nazionali europee. Pertanto, la digitalizzazione delle collezioni non dovrebbe più essere limitata alle opere inizialmente pubblicate in Francia.

Anche altre piattaforme già esistenti o da crearsi potrebbero svolgere tale ruolo in Europa. Queste piattaforme si scambierebbero reciprocamente i file e stabilirebbero i collegamenti che consentono la consultazione ed il *Web-harvesting* dei loro file.

III.3.2 Come far evolvere Europea

L’accelerazione della digitalizzazione ed il relativo ampliamento dei contenuti di Gallica andrebbero a rinforzare Europea, che è alimentata dagli Stati e dalle istituzioni che l’hanno fondata. Oltre a questo aspetto “quantitativo”, si potrebbero fare delle riflessioni di natura qualitativa.

Europea dovrebbe diventare un luogo di scambio, tramite il quale sviluppare tre linee di cooperazione:

- cooperazione in materia di metadati, formati e modalità di indicizzazione, allo scopo di aggiornare definitivamente le opzioni condivise dai partner della rete suddetta. Il portale beneficerebbe direttamente di quest’opera di armonizzazione;
- supporto alle biblioteche, affinché possano sottoscrivere reciproci accordi di scambio dei file;
- coordinamento delle riflessioni sulle questioni relative alle opere orfane e alle opere esaurite. Anche l’integrazione del programma Arrow è orientata in questa direzione.

Per quanto concerne le attuali funzionalità di Europeaana, il passaggio a modalità di ricerca a tutto testo, perlomeno sui contenuti scritti, andrebbe a migliorare sensibilmente la qualità della navigazione nel portale: i risultati della ricerca diventerebbero più pertinenti per l'utente. Sarebbe bene, inoltre, trovare i mezzi per un'indicizzazione a tutto testo da parte dei motori di ricerca, indicizzazione che garantirebbe una maggiore visibilità del portale e dei suoi contenuti sul Web.

Infine, sarebbe il caso di considerare le modalità di finanziamento di Europeaana e la possibilità, per alcuni progetti di digitalizzazione, soprattutto quelli a cui prendono parte stati con risorse limitate, di ricevere finanziamenti dell'Unione Europea più consistenti di quanto non sia attualmente previsto dal modello comunitario.

III.3.3 Una Carta comune per i partenariati pubblico/privato

Le biblioteche che fanno parte della Conference of European National Librarians (CENL) stanno attualmente considerando un progetto di Carta europea in materia di partenariati di digitalizzazione con aziende private. Il Gruppo di lavoro si congratula per questa iniziativa, che deve essere perseguita ed appoggiata.

A tal fine, suggerisce alcuni aspetti da trattare nell'ambito di questa Carta: le condizioni preliminari, le limitazioni delle clausole di esclusività e di durata, clausole che permettono la visibilità e l'aggiornamento dei file.

Condizioni preliminari

Le condizioni preliminari per la creazione del partenariato devono essere di duplice natura. Da una parte, il riconoscimento della necessità di un accordo in materia di diritti d'autore e di modalità di commercializzazione dei libri coperti da diritti. In altre parole, il partner o i partner privati devono rispettare le legislazioni europee che disciplinano i diritti d'autore.

D'altra parte, il rispetto di una serie di regole in materia di utilizzo dei dati personali degli utenti, al fine di proteggerli da eventuali abusi.

Clausole in materia di esclusività e di durata

Se il partner privato assume l'onere della digitalizzazione dei libri, le clausole di esclusività dovrebbero essere il più limitate possibile e non dovrebbero impedire l'accesso ai file da parte di aziende concorrenti disposte a pagare un diritto di accesso o una *royalty*³³. D'altra parte, è necessario respingere categoricamente le misure tecniche che impediscono l'indicizzazione del testo da parte dei motori di ricerca.

L'accordo dovrà anche prevedere la possibilità di accedere liberamente ai file dal portale Europeaana ed anche dai siti delle altre istituzioni europee partner.

Allo scopo di non pregiudicare i potenziali utilizzi, sarebbe bene controllare d'altra parte la qualità dei file forniti dal partner, cercando di definire tale qualità a priori, ma anche le modalità di verifica a posteriori.

Infine, dovrebbe essere possibile ridiscutere le clausole di esclusività periodicamente e gli accordi dovrebbero permettere in un certo numero casi un'eventuale risoluzione anticipata.

Clausole in materia di visibilità e aggiornamento dei file

La visibilità dei contenuti è una condizione indispensabile per la loro esistenza su Internet e la loro accessibilità. Quindi, anche queste clausole sono di estrema importanza. L'obiettivo è quello di fare in modo che i file del partner e quelli digitalizzati dalle biblioteche stesse siano oggetto dello stesso trattamento, nell'ambito dei meccanismi di ricerca e visualizzazione.

Per quanto concerne la qualità dei file forniti alla biblioteca, sarebbe opportuno avere sia file immagine che file testo di qualità simile a quella utilizzata dal partner privato. Anche i riferimenti che consentono di collegare questi due tipi di file dovrebbero essere forniti, per consentire alla

³³ L'accordo americano tra Google e gli editori prevede delle clausole di questo tipo.

biblioteca uno sfruttamento efficace.

L'accordo deve prevedere anche il reciproco aggiornamento dei file digitali, almeno per tutta la durata dello stesso. Se il partner aggiorna le sue tecniche di "ocerizzazione", dovrà fare in modo che i file che ha digitalizzato per la biblioteca beneficino di tali progressi. Allo stesso modo, se la biblioteca modifica le basi tecniche dei suoi file, dovrà estendere tali benefici al partner.

L'accelerazione del processo di digitalizzazione dei libri è ormai un imperativo che richiede nuove modalità organizzative. In questo contesto, Gallica deve diventare una vera e propria piattaforma di collaborazione con le biblioteche e gli editori partner. A questa prima fase, che rappresenta una forma comprovata di partenariato pubblico/privato, potrebbero aggiungersi nuove forme di accordo equilibrato con altri soggetti privati, come i motori di ricerca o le piattaforme di opere digitalizzate come Google Books. Infine, si sta sviluppando una riflessione approfondita a livello europeo che va sostenuta e continuata.

Sintesi conclusiva

La prospettiva di una digitalizzazione di massa delle opere a stampa, indipendentemente dal loro *status* (protette da diritti d'autore o meno), rappresenta una grande opportunità per la divulgazione della cultura francese. Comporterà una profonda revisione delle politiche pubbliche nel campo delle opere scritte, soprattutto nei confronti del grande pubblico che potrà accedere alle opere in base a modalità del tutto nuove, sia per le ricerche personali, sia per le attività quotidiane. È questa la posta in gioco, politica e culturale, del processo in corso.

La digitalizzazione di massa presenta una serie di costrizioni e di limitazioni:

- costrizioni, legate da un lato alle modalità

di organizzazione industriale implicate, dall'altro, alle esigenze qualitative da definire, sia a livello degli standard adottati che a livello dell'affidabilità dei file digitali, della loro indicizzazione e conservazione;

- limitazioni, in quanto non deve pregiudicare le altre modalità di digitalizzazione, indicizzazione e ricerca richieste dai pubblici specializzati (ricercatori, universitari, professionisti ecc.) o dai diversamente abili (ipovedenti, ecc.).

Il vasto programma intrapreso dalla società Google a livello mondiale ha dato uno slancio decisivo. Qualsiasi politica pubblica, per quanto ambiziosa, non può ignorare i grandi passi compiuti da questo operatore privato, sia a livello mondiale che europeo. Alcuni procedimenti adottati da tale società, in particolare nei rapporti con le biblioteche pubbliche e gli editori, sono alquanto criticabili. A tale proposito, le limitazioni imposte sulla divulgazione dei file ricevuti dalle biblioteche, la durata delle clausole di esclusività commerciale, l'imprecisione delle scelte tecniche e la riservatezza dei contratti stipulati con le biblioteche sono tutte condizioni difficilmente accettabili allo stato attuale, soprattutto per una biblioteca nazionale. Quindi, la Francia deve mantenere il controllo del processo di digitalizzazione e soprattutto di divulgazione dei contenuti digitali.

La decisione del Presidente della Repubblica sul finanziamento delle politiche di digitalizzazione getta le basi per una politica nazionale, autonoma e rispettosa dei diritti di tutte le parti coinvolte. Tuttavia è necessario che ogni istituzione, pubblica o privata, si impegni a contribuire alla nascita di un polo francofono in grado di confrontarsi con le piattaforme di ricerca e commercializzazione messe a punto dai gruppi mondiali come Google, Microsoft, Amazon... È tutta l'organizzazione della filiera della digitalizzazione che deve essere concepita a misura di quest'ambizione. Infatti, questo cambiamento di scala avrà implicazioni significative, non solo di natura quantitativa, ma anche qualitativa: imporrà di ri-

pensare la potenza delle infrastrutture non solo per quanto concerne i volumi da gestire, ma anche l'evoluzione delle professioni e l'acquisizione di nuove competenze.

Il Gruppo di lavoro propone tre linee di azione, che non si escludono l'un l'altra:

- partire dallo strumento già esistente, Gallica, realizzato dalla BnF, ma trasformandone radicalmente la gestione e le funzionalità;
- proporre alla società Google un'altra forma di partenariato, basato su uno scambio equilibrato dei file digitali, senza clausole di esclusività;
- rilanciare lo slancio europeo, sia a livello delle altre biblioteche europee coinvolte che del portale europeo comune Europeana.

Il Gruppo di lavoro vuole sottolineare che deve essere data la massima priorità alla valorizzazione del patrimonio librario sulle reti digitali ed alla sua visibilità per gli internauti di tutto il mondo. Le altre tappe del processo (filiera della digitalizzazione, immagazzinamento e manutenzione dei dati digitali), anch'esse importanti ed, in ogni caso, più onerose, potranno continuare ad essere gestite dalle istituzioni esistenti, in particolare dalla BnF. A tale proposito, sarebbe auspicabile che si possa usufruire della concentrazione fisica, nei locali della BnF, delle opere raccolte a titolo di deposito legale, al fine di avviare un processo di digitalizzazione di massa, non solo per quanto riguarda il patrimonio di opere non coperte da diritti, ma anche per le opere orfane o esaurite; ferma restando, ovviamente, la firma preliminare di un accordo quadro con gli editori, gli aventi diritto e le autorità pubbliche in merito alle condizioni di questa

digitalizzazione e, ovviamente, allo sfruttamento successivo.

Il Gruppo di lavoro ritiene essenziale la creazione di un organo di cooperazione che riunisca le biblioteche pubbliche e gli editori, in una logica di partenariato pubblico/privato. Dovrà anche fare spazio agli aventi diritto e agli altri soggetti della catena dell'editoria.

Questo organismo sarebbe investito della responsabilità di concepire, realizzare e utilizzare una piattaforma comune, da cui gli internauti potrebbero accedere all'insieme delle opere e consultarle, se lo desiderano. Inoltre, avrebbe la responsabilità di organizzare l'accesso alle opere e di concepire le interfacce con altre piattaforme, come i siti di *community*, i motori di ricerca, i siti di commercializzazione e via dicendo. Infine, in ragione del suo posizionamento al confine tra il settore del patrimonio culturale e quello dei diritti d'autore, dovrebbe svolgere un ruolo di piattaforma di diffusione e valorizzazione delle opere orfane, quando il codice della proprietà intellettuale consentirà lo sfruttamento digitale di tali documenti.

A tal fine, il Gruppo di lavoro ritiene opportuno attuare una riforma radicale, sulla scia del lavoro di qualità che ha consentito alla BnF, attraverso Gallica³⁴ di posizionarsi su Internet, ma modificando radicalmente le attuali prospettive. L'obiettivo deve essere quello di creare una base di opere digitalizzate di lingua francese, di qualità simile a quella di Google Book per la lingua inglese. Solo in questo modo la Francia potrà evitare un confronto troppo squilibrato con aziende di livello mondiale e fungere da esempio in Europa.

Il Gruppo di lavoro non aveva l'obiettivo di definire i dettagli della forma giuridica di tale organismo di cooperazione e della sua posizione rispetto alla BnF. Tuttavia, evidenzia la necessità di soddisfare una serie di condizioni: la collegialità del processo decisionale e la trasparenza delle modalità di commercializzazione che devono essere autorizzate da ogni partner coinvolto.

³⁴ La missione suggerisce un cambiamento di denominazione.

Il mondo di Internet è organizzato su base multipolare, ragion per cui ogni partner deve poter conservare il proprio sito o perlomeno organizzare direttamente la commercializzazione delle opere di cui è titolare, a seconda del loro *status*. Inoltre, per potere accedere ai finanziamenti pubblici a sostegno della digitalizzazione, ogni partner dovrà assumere una serie di obblighi:

- depositare i propri file sulla piattaforma cooperativa;
- adottare formati e standard tecnici compatibili con quelli definiti da tale piattaforma;
- delegare alla piattaforma i diritti che consentono al grande pubblico di indicizzare e sfogliare i file, ovvero, autorizzare, a seconda dei casi, lo sfruttamento commerciale di alcuni di questi file sulla piattaforma, senza pregiudicare i possibili utilizzi commerciali da parte degli editori stessi, tramite portali di vendita di loro scelta.

Le prospettive offerte dai partenariati pubblico/privato ne trarrebbero un considerevole giovamento. Da accordi spesso squilibrati, in quanto finanziati solo dalla parte privata, si passerà d'ora innanzi a formule di "scambio reciproco" che andranno ad arricchire le basi documentarie di entrambe le parti, evitando duplicazioni nel processo di digitalizzazione e favorendo la visibilità del *corpus* francofono. A titolo illustrativo, in questo modo le opere francesi sarebbero ampiamente indicizzate in Google Books, e viceversa, la piattaforma nazionale si arricchirebbe di opere già digitalizzate da Google, ovvero di quelle appartenenti ai patrimoni delle biblioteche straniere partner.

Il Gruppo di lavoro ha tentato di definire la possibile struttura di tali accordi reciproci e le condizioni preliminari dei mandati eventualmente affidati ai negoziatori, soprattutto per quel che riguarda il diritto d'autore, la tutela della privacy e la limitazione delle eventuali

clausole di esclusività (sia in termini di entità che di durata).

Nel caso di un'eventuale trattativa con Google, il Gruppo di lavoro non è in grado di stabilire la disponibilità della società a negoziare sulla base dei suddetti termini, sebbene la stessa abbia manifestato un evidente interesse nell'ampliamento del proprio contenuto francofono. Un'offerta di negoziazione presentata a nome della Francia avrebbe comunque un doppio vantaggio:

- far risaltare agli occhi dei partner europei la nostra volontà di costituire dei partenariati equilibrati, con questa società la cui presenza in Europa è indiscutibilmente forte;
- in caso di rifiuto, mettere l'altra parte in posizione di dover meglio precisare la propria strategia ed i propri obiettivi in merito al patrimonio librario europeo.

Le risorse finanziarie investite rinforzeranno la posizione avanzata della Francia in Europa in materia di digitalizzazione del patrimonio librario. È fondamentale sfruttare questa situazione per creare un effetto "traino", sia a livello dell'Unione Europea e degli Stati membri, sia a livello di istituzioni pubbliche e private, ove queste ultime potrebbero mettere in comune i mezzi tecnici e condividere la piattaforma sviluppata dalla Francia. La Società Google oggi è in grado di proporre una struttura collaudata, in base a condizioni economiche senza pari. Le iniziative intraprese in Francia devono riuscire a colmare questo divario e creare un'alternativa a vantaggio dell'insieme dei partner europei.

Il Gruppo di lavoro propone di considerare tre linee di azione:

- fare in modo che i partner europei traggano beneficio delle economie di scala realizzate grazie agli investimenti francesi, proponendo alle biblioteche europee interessate di condividere, eventualmente tramite

- la piattaforma cooperativa e la BnF, l'insieme del processo di digitalizzazione;
- perseguire ed indubbiamente rinforzare Europea (alla fine);
 - elaborare una Carta che inquadri i partenariati pubblico-privato nella serie dei dibattiti già avviati tra numerose grandi biblioteche nazionali.

Il Ministro

21 ottobre 2009

Egr. Sig. Marc TESSIER
Direttore Generale
Vidéo Futur Entertainment Group
27, rue d'Orleans
92200 Neuilly-sur-Seine

Egregio Direttore Generale,

La visibilità del *corpus* librario francofono sulle reti è un obiettivo maggiore che richiede la condivisione di competenze ed esperienze molto diversificate.

Gli aspetti a valle, ovvero le condizioni di un accesso aperto, e quelli a monte, cioè la qualità della digitalizzazione e la pertinenza delle modalità di indicizzazione, devono essere posti sullo stesso piano. Alcuni ritengono che Google privilegi gli aspetti a valle a scapito della qualità, mentre le biblioteche pubbliche punterebbero in primo luogo a valorizzare il loro *know-how* a monte, a svantaggio delle modalità di consultazione di massa. Come spesso accade, tali analisi sono sommarie, ma ci richiamano al nostro impellente obbligo di concepire una modalità di organizzazione e di partenariato che preservi questo equilibrio. Questo è quanto il Gruppo di lavoro ha tentato di concepire.

Elenco degli allegati

Allegato 1: Lettera d'incarico

Allegato 2: Lista delle persone intervistate

Allegato 3: Gallica e Google Books: confronto delle funzionalità – a cura di Alban Cerisier

Allegato 4: Lista delle biblioteche europee che aderiscono al programma di ricerca Google Recherche de Livres

Allegato 1: Lettera d'incarico

Repubblica Francese
Ministero della cultura e della comunicazione

A partire dal 2005, la Biblioteca nazionale di Francia, ha deciso di trasformare la Biblioteca digitale Gallica in un importante strumento di divulgazione dei contenuti francofoni ed europei sul Web, in particolare attraverso la sua partecipazione al progetto Europea, sostenuto dalla Commissione Europea e di cui la Francia è oggi il principale contributore.

Questo obiettivo ha implicato per la BnF la messa a punto di una politica di digitalizzazione di massa particolarmente ambiziosa, caldamente sostenuta dallo Stato, tramite i finanziamenti stanziati dal 2007 in poi dal Centro nazionale del Libro. Oggi, la Biblioteca digitale Gallica mette a disposizione più di 850.000 documenti online, di cui circa 140.000 libri.

Animato dal desiderio di proporre agli internauti un'offerta diversificata e moderna, il Ministero della cultura e della comunicazione ha sottoscritto un partenariato con oltre un centinaio di editori francesi, rendendo possibile, attraverso Gallica, la consultazione di opere protette da diritti di autore. Questo partenariato è attivo dal 2008 e consente la consultazione di oltre 10.000 libri.

Questa politica di digitalizzazione implica ingenti finanziamenti statali (oltre 25 milioni di euro stanziati dall'inizio del progetto) e la BnF sta lanciando o considerando nuovi lavori di digitalizzazione: collezioni specialistiche, libri rari, contenuti di stampa periodica, audiovisivi, ecc.. Per questa ragione, un recente rapporto dell'Ispettorato generale delle finanze sulla Biblioteca nazionale di Francia ha stabilito che tale istituzione

«dovrebbe rinnovare il dialogo, su basi più equilibrate, con Google o altri partner privati ed abbandonare l'attuale logica di concorrenza, nella prospettiva di un eventuale finanziamento condiviso della digitalizzazione di massa di opere non coperte dai diritti».

Come Lei sicuramente saprà, nelle ultime settimane sono stati pubblicati numerosi articoli, alcuni di tono polemico, in merito ai contatti intercorsi tra la Biblioteca nazionale di Francia e Google, criticando l'idea di un partenariato tra la BnF e la società americana.

Al fine di consentire al governo di tutelare al meglio l'interesse pubblico e di prendere la posizione più idonea, gradirei affidarle l'incarico di analizzare e valutare l'opportunità di un tale possibile partenariato. Pertanto, desidererei che fossero definiti:

- i tipi di collaborazione ipotizzabili tra la BnF e Google;
- gli aspetti da considerare come non negoziabili dalla parte francese, di natura tecnica, giuridica o economica;
- l'opportunità di un accordo tra la BnF e Google, da tre punti di vista, della visibilità, della cultura e dell'accesso ai contenuti francesi su Internet, dell'interesse economico e finanziario per lo Stato e per i contribuenti e del messaggio politico da trasmettere alla comunità internazionale;
- le proposte alternative da Lei ritenute opportune in materia di partenariati privati, sulla base degli elementi che avrà modo di analizzare.

Il Gruppo di lavoro che Lei presiederà sarà composta da:

- Emmanuel HOOG, presidente dell'Institut National de L'Audiovisuel;
- Olivie BOSC, conservatore in capo delle biblioteche BnF, professore incaricato presso l'Institut d'Etudes Politiques di Parigi;

- Alban CERISIER, direttore del patrimonio e dello sviluppo digitale presso l'editore Gallimard,
- Francois-Xavier LABARRAQUE, direttore Strategia e Sviluppo di Radio-France;

Sophie-Justine LIEBER, *maitre de requetes* al Consiglio di Stato, sarà il relatore dei Suoi lavori.

Tutti i dipartimenti del Ministero della cultura e della comunicazione ed in particolare la Direzione del Libro e della Lettura, saranno a Sua disposizione per assisterLa nell'espletamento di tale incarico. La prego di volermi consegnare il rapporto entro il 15 dicembre ed un rapporto provvisorio il 25 novembre.

Colgo l'occasione per porgerLe i miei più distinti saluti,

Frédéric Mitterand

Allegato 2: Lista delle persone intervistate

Jean-François Aguesse, Direttore dello Sviluppo e Socio, Synodiance;

Francis Amand, Capo Servizio presso la Direzione Generale della concorrenza, il consumo e la lotta alle frodi;

Patrick Bazin, direttore della Biblioteca municipale di Lione;

Arnaud Beaufort, Direttore generale aggiunto della Biblioteca nazionale di Francia, Direttore dei servizi e delle reti;

Bernard Benhamou, delegato agli utilizzi di Internet, Ministero dell'Istruzione Superiore e della Ricerca;

Philippe Colombet, Direttore del programma Recherche de Livres, Google France;

Robert Darnton, Direttore della biblioteca dell'Università di Harvard;

Olivier Daube, Responsabile del Polo Sorveglianza degli utilizzi e Innovazioni, Radio France;

David Drummond, Vice-presidente senior, direttore dell'ufficio legale, Google;

Serge Eyrolles, Presidente del Sindacato

Nazionale degli Editori;
Jeannette Frey, Direttrice della biblioteca cantonale e universitaria di Losanna;
Antoine Gallimard, Presidente e Direttore Generale, Editions Gallimard;
Emmanuel Gureghian, Bertin Technologies;
Theo Hoffenberg, Presidente e Direttore Generale, Softissimo;
Hervé Hugueny, giornalista di Livres Hebdo;
Jean-Noël Jeanneney, ex presidente della Biblioteca nazionale di Francia;
Dominique Lahary, direttore della Bibliothèque départementale de prêt de Val-d'Oise, portavoce dell'Interassociazione Archivi, Biblioteche, Documentazione;
Bernard Lang, Direttore della ricerca presso l'Istituto nazionale della ricerca del settore informatica;
Pierre Louette, Presidente e direttore generale, AFP;
Michel Marian, Capo del Gruppo di lavoro per l'informazione scientifica e tecnica e le reti documentarie, Ministero dell'Istruzione Superiore e della Ricerca;
Nicolas Masson, Bertin Technologies;
Alexandre Moatti, Ingegnere capo di Mines, ex segretario generale del comitato di gestione della Biblioteca Digitale Europea;
Marc Mossé, Direttore degli affari pubblici e giuridici, Microsoft France;
Elisabeth Niggemann, direttrice generale della Deutsche Bibliothek, presidente del comitato europeo dei direttori delle Biblioteche nazionali, presidente della fondazione European Digital Library (Europeana);
Arnaud Nourry, Presidente e Direttore Generale, Hachette Livre;
Richard Ovenden, Conservatore della biblioteca Bodleienne, Università di Oxford;
Bruno Patino, Direttore di France Culture;
Alain Pierrot, I2S;
Michael Popham, Università di Oxford;
Bruno Racine, Presidente della Biblioteca nazionale di Francia;
Stéphane Ramezi, Direttore del settore multimediale, Radio France;

Pascal Rogard, Direttore generale della Société des auteurs et compositeurs dramatiques (SACD);
Claude Rubinowicz, Direttore generale dell'agenzia del patrimonio immateriale dello Stato (APIE);
François Stasse, Consigliere di Stato, ex direttore generale della Biblioteca nazionale di Francia;
Sarah Thomas, Direttrice della biblioteca Bodleienne e delle biblioteche dell'Università di Oxford;
Jacques Toubon, ex ministro, co-responsabile del Gruppo di lavoro creazione e Internet.

Allegato 3: Le sfide qualitative della digitalizzazione di massa

Considerazioni su Gallica e Google Books

Allegato a cura di Alban Cerisier

Gallica e Google Books sono entrambe biblioteche digitali, in quanto conservano e mettono a disposizione del pubblico, su un sito Web dedicato, collezioni di libri digitalizzati, di cui esse stesse hanno curato e gestito le operazioni di digitalizzazione.

Ambedue propongono una ricerca a tutto testo sulle opere e/o sui riferimenti che le descrivono (metadati) e permettono di accedere online o tramite downloading alla versione integrale o a estratti dei testi stessi, direttamente dalla piattaforma oppure indirizzando l'internauta ad altre piattaforme di lettura online.

Google Books ambisce ad essere un portale bibliografico in quanto integra nel suo *corpus* riferimenti ad opere il cui contenuto non è ancora stato digitalizzato né indicizzato (soprattutto opere recenti).

Le due applicazioni propongono funzionalità diverse ed evolutive, sia in termini di formulazione e trattamento delle richieste dell'internauta che di restituzione e messa a disposizione delle opere. Queste opzioni tecniche e funzionali sono al servizio di concezioni uniche di valorizzazione del patrimonio bibliografico digitalizzato e determinano utilizzi differenziati e,

più generalmente, sono la dimostrazione della diversità delle strategie di posizionamento adottate dalle istituzioni in questione.

Si potrebbe dire lo stesso delle modalità di digitalizzazione, in quanto la scansione di massa delle opere implica, allo stato dell'arte, scelte qualitative problematiche. Occorre quindi delineare la struttura funzionale e tecnica di queste biblioteche digitali, sia sul piano degli elementi documentari prodotti che su quello del loro accesso e valorizzazione.

- I. Le sfide qualitative della digitalizzazione di massa
 - a. La digitalizzazione di massa di una parte del patrimonio.
 - b. La coppia immagine/testo.
 - c. La strutturazione del libro.
 - d. La qualità dell'immagine scansionata.
 - e. Il riconoscimento dei caratteri.
 - f. La correzione a posteriori?
 - g. Una qualità doppiamente problematica (conservazione ed utilizzi).
 - h. Un elementare principio precauzionale

- II. Le sfide qualitative della valorizzazione
 - a. Il corpus
 - b. I metadati
 - c. I motori di ricerca
 - d. Le restituzioni
 - e. La consultazione
 - f. La referenziazione

I. Le sfide qualitative della digitalizzazione di massa

Una digitalizzazione di massa per una parte del patrimonio

È necessario differenziare i lavori di digitalizzazione di massa dalle iniziative più mirate rivolte a *corpus* ridotti e basate su standard qualitativi elevati e modulabili (trattamento di immagini e di testi). Un'istituzione come la Biblioteca nazionale di Francia, ed anche altre biblioteche europee, ha condotto, conduce e continuerà a condurre operazioni specifiche su contenuti atipici

(cartine e mappe, antichi manoscritti stranieri, ecc.). È importante continuare a perseguire le iniziative di questo tipo, soprattutto alla luce dell'impossibilità di sottoporre a trattamento di massa un numero consistente di documenti che hanno caratteristiche fisiche particolari (dimensioni, stato di conservazione ecc.).

Invece, lo scopo dei grandi lavori di digitalizzazione di massa come quelli intrapresi negli ultimi anni da Google Books e dalla BnF, è quello di digitalizzare, nell'ambito di una filiera di produzione industriale, un quantitativo molto elevato di opere. Ogni singola pagina deve essere sottoposta a scansione, senza distruggere né rovinare l'opera, che deve essere poi restituita all'istituzione a cui appartiene e che la reinserirà nelle sue collezioni. Nessun'opera viene tagliata: un operatore gira le pagine manualmente (molte migliaia di pagine al giorno per ogni persona).

Questi "stabilimenti" di digitalizzazione di massa possono essere concepiti solo se la Biblioteca in questione dispone di un catalogo informatizzato delle proprie opere, altrimenti non è possibile garantire la tracciabilità delle opere tra la Biblioteca e la filiera produttiva, e all'interno della filiera stessa.

Il personale delle biblioteche effettua una selezione allo scopo di individuare, tra le opere candidate alla digitalizzazione, quelle che non possono essere considerate in ragione del loro stato di conservazione, del loro valore e della rispondenza ai criteri della digitalizzazione industriale (Google esclude dalla sua filiera qualsiasi opera che abbia anche un solo foglio ripiegato, il quale una volta aperto, ecceda le dimensioni del libro). Sulla base dell'esperienza sinora acquisita dalla BnF in materia di digitalizzazione di massa, sembra che la percentuale di opere respinte sia molto elevata.

La messa a disposizione omogenea delle opere richiede trattamenti differenziati.

La digitalizzazione di massa, anche manuale, coinvolge quindi solo una parte delle collezioni e pertanto rappresenta solo una delle compo-

nenti della suddivisione in lotti necessaria per organizzare la conversione digitale dei patrimoni bibliotecari. Occorre quindi controllare che questi tipi di trattamento differenziato a monte, non implicino poi dei sistemi eterogenei di messa a disposizione del pubblico da parte della biblioteca. Questo è uno degli scopi identificati nei partenariati pubblico/privato e soprattutto in seno agli accordi di partenariato tra Google Books e le biblioteche partner.

La coppia immagine/testo

L'attività di digitalizzazione comporta due fasi: prima, ogni pagina, copertina inclusa, deve essere sottoposta a scansione immagini e poi, dall'immagine così ottenuta, si estrae, in seguito ad un processo di ri-trattamento e ottimizzazione condotto con l'ausilio di specifiche applicazioni, un file di testo per mezzo del riconoscimento automatico dei caratteri stampati (Optical Character Recognition, OCR). A questo punto, l'opera è stata "ocerizzata". Questa fase può richiedere il ricorso ad un'immagine intermedia, diversa da quella destinata all'utente, specificamente adattata alla realizzazione dell'OCR.

Anche il *text layer*³⁵ estratto dall'immagine può essere ottimizzato, ricorrendo a specifici strumenti di correzione. Queste due attività sono ancora oggetto di ricerca e sviluppo³⁶ ma sono già in uso da parecchi anni, soprattutto nel settore della produzione grafica tradizionale (ri-stampa di opere degli archivi editoriali).

Al termine di questa operazione, per ogni pagina del libro vi sono uno o più file immagini (a seconda se si è scelto di scansionare l'opera in bianco e nero, nella scala dei grigi o a colori, e quindi a seconda delle diverse definizioni) ed un file di testo, generalmente in formato XML, basato su un modello di descrizione dei dati standardizzato. La sfida della digitalizzazione di massa consiste nel rendere solidali questi due "stati" del libro.

Alcuni testi presenti in Gallica, ovvero i file digitalizzati della generazione precedente (prima del 2007), non sono provvisti di questa coppia immagine/testo, ma solo di un'immagine. La BnF ha già effettuato buona parte di tale conversione (il cosiddetto contratto "dei 60.000").

La strutturazione del libro

Al fine di consentire un collegamento tra l'immagine della pagina ed il file di testo che ne deriva, i parametri grafici di ogni parola sono memorizzati nel file testo quando viene generato, sotto forma di tag retraibili. Dei tag consentono anche di ripristinare la strutturazione logica della pagina (localizzazione dell'intestazione, dei piè di pagina e dei margini, per esempio), di individuare gli spazi in cui si colloca un'illustrazione, ecc. L'insieme di queste operazioni è reso possibile dalla mappatura preliminare del file immagini e dall'utilizzo di un formato descrittivo standardizzato di tipo XML. Gallica raccomanda l'adozione dello standard Alto, che pare sia stato in parte utilizzato anche da Google.

Per restituire al libro la sua integrità logica, occorre anche associare ogni coppia di file (immagine/testo) alla pagina corrispondente nell'opera. Quindi, il file numero 7 relativo ad un libro non corrisponde necessariamente alla pagina numero 7 dell'opera (infatti, potrebbero esserci delle pagine bianche, delle pagine non numerate, doppie numerazioni in cifre romane e arabe e via dicendo). Pertanto, deve essere creato un terzo file, anch'esso in XML per ricostituire la sequenza logica del libro e collegarla all'insieme dei file e delle cartelle che compongono il libro nella sua forma digitale. Questa è la spina dorsale dell'opera, essenziale per poter informare l'internauta, per esempio, che la citazione è stata trovata nella pagina tot del libro tot. Questo file descrittivo può essere utilizzato anche per identificare o tipizzare le pagine im-

³⁵ [Si tratta del primo prodotto (il primo livello) della digitalizzazione.]

³⁶ Questo è uno dei motivi per cui Google ha acquisito la società reCaptcha (acquisizione annunciata nel settembre 2009), specializzata nella conversione di immagini in testo, soprattutto nell'ambito della sicurezza dei siti e della protezione antispy.

portanti del libro: pagina dello stesso autore, pagina di occhio, pagina del titolo, pagina dell'indice, ecc. Tale fase è decisiva, in quanto consentirà di escludere alcune pagine sia dalla restituzione del libro, sia dal campo di indicizzazione del motore di ricerca.

A quanto pare, purtroppo, né Google né Gallica hanno eseguito questa tipizzazione in maniera approfondita; resta il fatto che l'automatizzazione di questa fase è problematica. Tuttavia, anche le potenzialità dello standard Alto non sono state sfruttate; per esempio, sarebbe stato possibile eliminare dal campo del motore di ricerca l'indicizzazione dei titoli ricorrenti, cioè quelli che si ripetono su ogni pagina dei libri. Sembrerebbe che questa operazione non venga sempre effettuata nei laboratori della digitalizzazione di massa.

Alcune parti del libro potrebbero essere trattate in modo speciale, per esempio gli indici. La localizzazione degli indici nel libro può essere eseguita automaticamente, ed allo stesso modo, un apposito programma potrebbe ricostruirne la struttura logica. In ogni caso, i risultati ottenuti sino ad ora con tali trattamenti sono decisamente imperfetti, come dimostrato dall'escrabile qualità degli indici di Google Books che sono pieni di errori e di lacune e pertanto, sono spesso inutilizzabili. Gallica ha optato per un trattamento manuale di tali indici che, anche se non perfetti, sono sicuramente più accettabili. Gli indici "ocerizzati" sono sottoposti ad una sistematica correzione da parte di un operatore. Si tratta di un'operazione particolarmente sensibile. Infatti, la strutturazione degli indici può essere molto complessa e problematica da un punto di vista logico ed editoriale. Questa fase è importante se si vuole concepire uno sfruttamento di parti delle opere, sulla base della strutturazione del libro stampato, nel contesto di una consultazione o di un download dei singoli capitoli.

La qualità dell'immagine scansionata

Tutte le fasi della filiera devono essere sottoposte a controlli di qualità, nell'ambito di protocolli dettagliati che definiscono i tipi di avvisi e di azioni da intraprendere in caso di rilevamento di anomalie. Per esempio, occorre verificare che non siano state omesse pagine e definire una procedura per l'eventuale "riparazione".

Anche la qualità dell'immagine prodotta rappresenta un elemento sensibile. La qualità si gioca a diversi livelli: la definizione dell'immagine, l'adattamento dell'immagine all'OCR, l'intervento manuale per non deteriorare l'immagine, ecc. In varie occasioni è stata notata la mediocrità della digitalizzazione effettuata da Google Books³⁷, dimostrando così la necessità di rinforzare i controlli automatizzati nell'ambito della filiera e conseguentemente fare in modo che le istituzioni pubbliche siano più esigenti nei confronti dei fornitori del servizio. Infatti, questo oggi rappresenta uno degli aspetti della loro missione di conservazione del patrimonio culturale, fisico e digitale. In questo contesto, il livello di definizione dei file prodotti può essere un elemento fondamentale dell'iniziativa. La BnF ha recentemente imposto ai suoi fornitori una digitalizzazione non più in 300 Dpi ma in 400 Dpi, più costosa e più voluminosa, ma in grado di produrre risultati più soddisfacenti nella prospettiva di fornire un servizio di stampa su richiesta. Le sfide della conservazione perenne e dello sfruttamento sono quindi strettamente correlate.

Il riconoscimento dei caratteri

L'altra grande sfida qualitativa riguarda l'OCR, che può essere più o meno soddisfacente a seconda delle modalità di stampa della pagina, delle variazioni tipografiche interne, degli alfabeti e delle grafie sillabiche, delle lingue e così via. Chi esegue la digitalizzazione deve assumersi degli impegni qualitativi precisi dopo

³⁷ È sufficiente scaricare un file Epub da Google Books per verificare di persona. Nella sotto-cartella "immagini" della cartella OEBPS appare una miriade di scatti in cui si riconosce addirittura la mano quantata dell'operatore. Queste immagini si ritrovano nel libro digitale, come elementi grafici non analizzabili dall'OCR.

aver esaminato il *corpus* candidato alla digitalizzazione; inoltre, la percentuale di OCR (percentuale di errori nell'opera, calcolata in base a campionamenti condotti dal fornitore, fatti salvi i casi di illeggibilità), che è molto variabile da un libro all'altro in questo tipo di lavoro, deve essere comunicata alla biblioteca. Anche se la percentuale di riconoscimento dichiarata può sembrare molto elevata (99,9% per esempio, ovvero due errori grossolani per una pagina di 2000 caratteri in media), va tenuto presente che un livello qualitativo simile sarebbe poco accettabile per un libro stampato, debitamente verificato e corretto.

Il collegamento tra l'immagine ed il testo che ne deriva serve a compensare le lacune dell'OCR. Se il motore di ricerca indicizza il *text layer*, è l'immagine del libro che viene mostrata in prima istanza all'internauta, in seguito alla sua ricerca. L'immagine non è, per sua natura, suscettibile di errori. Grazie alle coordinate della parola immagazzinata nel file di testo, l'applicazione di ricerca per pagina può evidenziare facilmente lo spazio della pagina in cui si trova la citazione o le parole cercate. Ma l'indice rimane comunque difettoso. Bisogna anche considerare che, allo stato dell'arte, qualsiasi ricerca effettuata nell'ambito di una biblioteca digitale è, in senso stretto, approssimativa, in quanto non vi sono garanzie che la parola cercata si trovi solo nei punti segnalati dal motore di ricerca. Una risposta affidabile al 100% richiederebbe una fase costosa e laboriosa di correzione manuale dei file di testo ottenuti con l'OCR. Questo tipo di trattamento, allo stato attuale, è applicato in misura marginale.

I libri digitali proposti agli internauti

I file di cui Google Books e Gallica offrono la visualizzazione o il download sono il risultato di questo lavoro di preparazione. Possono essere semplici immagini integrate nell'applicazione di ricerca, con un *text layer* invisibile per l'internauta, oppure essere costituiti dal *text layer* stesso, non corretto, o da un'immagine PDF creata sulla base dei diversi file sottoposti a

scansione (non interrogabile, non indicizzabile: Google Books), o da un PDF testo/immagini, generato sulla base delle scansioni e del file testo dell'OCR (interrogabile e indicizzabile nel caso di Gallica), o ancora da un file di testo, come nel caso del file Epub. Questo file Epub che negli ultimi anni si è imposto come standard per i libri digitalizzati (anche se molto imperfetto rispetto al livello di precisione editoriale del libro di origine, per non dire assolutamente non all'altezza), si presenta come una cartella compressa che comprende l'insieme del testo dell'opera sotto forma di file diversi. Può essere letto dai PC (con l'aiuto del software gratuito Adobe Digital Edition), ma anche dalla maggior parte dei terminali mobili di lettura e dagli I-Phone. Va precisato che la qualità di questo file (e quindi la sua leggibilità) è strettamente legata alla qualità del trattamento OCR. Se il file non è stato sottoposto a revisione, conterrà gli stessi errori di trascrizione, che possono essere molto penalizzanti per il lettore (come è accaduto con qualche file Epub consultato ai fini di questa indagine dal sito Google Books).

Il servizio automatizzato di lettura a voce alta, disponibile su Gallica, si appoggia anch'esso su questo *text layer* prodotto dall'OCR. Conseguentemente, la qualità dell'audio dipende dal trattamento iniziale: se l'OCR è scarso, il testo diventa incomprensibile. Questi problemi non si verificano se, come nel caso delle opere recenti, ci si basa su un file di testo non prodotto da una scansione, ma fornito dal compositore dell'opera a stampa, considerata priva di errori.

Gallica e Google Books non forniscono mai il file immagini fonte agli internauti, né in modalità visualizzazione né in modalità download. Entrambe generano, a valle della filiera della digitalizzazione, un file molto più leggero in 75 Dpi. La rapidità della restituzione lo impone, ma in questo modo si limita la divulgazione dei file idonei alla stampa su richiesta o, più generalmente, alla commercializzazione.

Nell'ambito dei ben noti accordi sottoscritti tra

Google Books e le biblioteche, il motore di ricerca americano inoltra una copia dei file sorgente della digitalizzazione ai suoi partner³⁸, ma non li autorizza a fornire agli internauti un file che contenga sia immagini che testi (per esempio PDF immagine/testo). La biblioteca può quindi offrire ai suoi utenti solo il download del PDF immagine³⁹, non indicizzabile e non interrogabile (come d'altronde fa Google Books attualmente sul suo sito) "al pezzo" e per un "utilizzo individuale". Questo è un limite non indifferente⁴⁰, che trova una sua giustificazione nella volontà manifesta del motore americano di non rendere accessibili ai concorrenti i file da esso digitalizzati. Tale misura precauzionale sembra basarsi sulla stessa preoccupazione che spinge Google ad applicare misure tecniche di protezione contro il download massivo sul suo sito ed a esigere dalle biblioteche partner, per quanto concerne il loro sito, garanzie dello stesso tipo⁴¹, forse ancora più rigide, a seconda dell'interpretazione delle clausole contrattuali, e molto vincolanti in termini di responsabilità a carico delle istituzioni. A questo proposito, sussiste un rischio di squilibrio tra l'utilizzo che potrebbero fare dei loro file e quanto Google Books si è riservata.

Una correzione a posteriori?

La maggior parte delle carenze qualitative sopra esposte (assenza di tipizzazione delle pagine, qualità mediocre dell'OCR, errori nelle immagini, indici difettosi) potrebbero essere corrette a posteriori, ricorrendo all'intelligenza collettiva di

una rete (wiki), risottoponendo le scansioni originali ad una filiera di trattamento OCR resa affidabile o, ancora, lavorando su soluzioni di correzione automatica basate sul confronto delle opere digitalizzate a più riprese.

Il ricorso alla correzione collaborativa dell'OCR presenta comunque un certo numero di rischi per l'integrità dei testi, in quanto può comportare una dissociazione progressiva del file di testo rispetto all'immagine di riferimento. Questo rischio, di cui si parla poco, non è trascurabile. La valutazione di queste possibilità di miglioramento a posteriori deve ancora essere fatta, anche se sono già state intraprese delle iniziative in questo ambito sul Web.

La questione della qualità è doppiamente problematica (conservazione ed utilizzo)

Conservazione

Gli attuali processi di digitalizzazione di massa, nonostante i recenti progressi ed i nuovi utilizzi che si sono creati, sono caratterizzati da un livello qualitativo alquanto mediocre. L'idea di una conservazione digitale perenne, al di là degli aspetti legati alla sola infrastruttura tecnica di immagazzinamento ed al carattere evolutivo dei formati informatici, è rimessa in discussione a causa dell'irregolarità dei risultati ottenuti in termini di immagini, ed ancora di più, di conformità testuale. Se la nozione di copia di conservazione è intesa soltanto nel senso di un facsimile analogico delle opere ci si potrebbe accontentare sempre che il lavoro sia fatto con accuratezza. Se invece, questa stessa nozione

³⁸ Senza indicare di quale file si tratta esattamente, aspetto sul quale le biblioteche dovrebbero vegliare, il CCTP dell'accordo di Lione prevede che Google inoltri alla Biblioteca municipale il testo «in formato grezzo e senza arricchimenti tipografici», mentre l'atto di impegno contempla la fornitura di una «copia dei file e dei metadati digitali creati dal fornitore» (art. 6-1), senza ulteriori precisazioni. Cosa significa? Si tratta veramente del file strutturato e coordinato che permette il collegamento tra l'immagine ed il testo prodotto dall'OCR? Google dispone a valle di un file più completo di quello fornito al suo partner?

³⁹ Articolo 24 del CCTP relativo all'accordo Google/BM di Lione.

⁴⁰ Questo limite è comunque superabile, almeno in parte, con la messa a disposizione in download dei file Epub.

⁴¹ Clausola contemplata nei contratti con l'Università del Michigan e l'Università della California: «University shall implement technological measures [...] to restrict automated access to any portion of the University Digital Copy or the portions of the university website on which any portion of the University Digital Copy is available».

si riferisce alla restituzione di un *corpus* testuale strutturato, siamo ancora ben lontani dall'obiettivo e la filiera deve ancora fare molti progressi. Per quanto concerne il trattamento dei testi, anche semplici (senza raffinatezze semantiche), le biblioteche digitali, che ricorrono fondamentalmente alla digitalizzazione di massa, adottano pratiche approssimative.

Utilizzi

Questi elementi tecnici influenzano gli utilizzi. Per utilizzi di tipo documentario (ricercatori, studenti, o per una ricerca accurata di citazioni), le biblioteche digitali sono ormai diventate degli strumenti indispensabili, sia per l'indicizzazione dei testi che per la loro restituzione integrale in modalità immagine. Tutto quello che potrà contribuire a rendere più affidabile l'indicizzazione ed a migliorare il motore di ricerca, tramite un migliore trattamento delle opere a monte (gestione degli elementi strutturali del libro, ottimizzazione dell'OCR) si rivelerà utile e risponderà all'obiettivo, non della conservazione, ma della comunicazione e valorizzazione dei patrimoni delle biblioteche. Queste ultime devono quindi partecipare maggiormente alla definizione degli impegni in termini di qualità del servizio per quanto riguarda i prossimi grandi lavori di digitalizzazione e non sottostare a condizioni poco trasparenti imposte dai fornitori e/o dai partner.

L'attuale tecnologia di digitalizzazione di massa, non è in grado di soddisfare un utilizzo come quello di una normale lettura (del tipo "vorrei accedere gratuitamente ad un'edizione di *Le Rouge et le Noir* leggibile sull'iphone), in quanto il formato testo prodotto (Epub) è difettoso. È impensabile per una biblioteca nazionale, e poco compatibile con la sua missione di valorizzazione del patrimonio culturale nazionale, l'idea di distribuire massicciamente testi classici della letteratura francese costellati di errori. In questo ambito, sarebbe meglio proporre una selezione delle opere che potrebbero essere sottoposte ad un trattamento più accurato, con una percentuale di riconoscimento del carattere

vicina al 100% (come qualche prototipo di *ebook* proposto da Gallica); o considerare dei partenariati con gli editori di questi testi, se esistono, per proporre versioni scaricabili gratuitamente, ma di qualità, che oltre tutto sono già disponibili in altri siti Web.

Un principio precauzionale elementare

Natura evolutiva dei file

Nel frattempo, è probabile che le filiere di OCR e le possibili correzioni dei testi a posteriori, per mezzo di un trattamento di massa, continueranno a migliorare negli anni. Il fatto che gli accordi sottoscritti tra Google Books e le biblioteche partner non prevedano mai, nel periodo di esclusività commerciale concesso a Google Books, l'obbligo a carico di quest'ultima di condividere con la Biblioteca partner gli eventuali progressi tecnologici, tramite la fornitura periodica dei file sottoposti a rielaborazione, è alquanto sconcertante. Per tale ragione, si verrà a creare un divario enorme tra l'offerta di Google Books e quella della Biblioteca partner sul proprio sito. A questo proposito, sussiste un rischio molto significativo di emarginazione delle istituzioni.

II. Le sfide qualitative della valorizzazione

Il corpus

Sia i libri soggetti a diritto d'autore che quelli di dominio pubblico sono visibili su entrambe le piattaforme.

Le modalità di interrogazione e di restituzione dipendono essenzialmente dal loro *status* giuridico. In Google Books come in Gallica, le opere considerate di dominio pubblico sono interrogabili integralmente e possono essere lette nella loro totalità. Nella grande maggioranza dei casi provengono dalla digitalizzazione di copie conservate nelle biblioteche (biblioteche che aderiscono al progetto Gallica oppure, nel caso di Google, biblioteche partner).

Lo *status* ed il trattamento delle opere coperte da diritti variano a seconda che l'istituzione abbia sottoscritto degli accordi con gli aventi diritto o meno.

In caso di accordo con i titolari dei diritti (partenariato SNE – Sindacato Nazionale degli Editori /BnF per Gallica, con il sostegno del Centro nazionale del Libro; Programme Editeurs di Google o ancora divulgazione dei libri coperti da diritti Creative Commons), la biblioteca dispone degli elementi necessari per indicizzare l'opera nella sua integralità e renderla disponibile all'internauta, totalmente o in parte nella propria applicazione di ricerca (Google) o nell'applicazione indicata dall'avente diritto (Gallica). È il titolare dei diritti che stabilisce, in entrambi i casi, il numero delle pagine liberamente consultabili.

In mancanza di accordo, invece, Google Books si autorizza comunque a digitalizzare ed indicizzare le opere coperte da diritti sulla base di una copia scansionata nell'ambito del suo Programme Bibliothèques. Tale copia è interrogabile sulla sua piattaforma, ma è possibile visualizzarne solo tre brevi estratti corrispondenti alla ricerca effettuata, sotto forma di *Snippets*, ove le giurisdizioni nazionali non si sono opposte al principio di tale pubblicazione, risultato di una digitalizzazione non precedentemente autorizzata dal titolare dei diritti (Caso Google/La Martinière/SNE/SGDL).

Da notare infine che ultimamente Gallica ha inventariato un gran numero di stampe e di immagini. Se questa scelta è perfettamente coerente per quanto riguarda le collezioni della Biblioteca nazionale di Francia, è il caso di chiedersi se un approccio multi-supporto potrebbe soddisfare realmente le aspettative del pubblico, soprattutto in considerazione dell'"inquinamento" che potrebbe comportare nell'ambito di una semplice ricerca. Google, sinora, non ha adottato questo approccio.

I metadati

I metadati sono gli elementi che descrivono l'opera. Google Books e Gallica si basano sullo stesso standard internazionale, il Dublin Core,

un modello che garantisce l'interoperabilità dei metadati che descrivono le diverse risorse documentarie. Questo modello concettuale, semplice ed efficace, è costituito da quindici elementi, opzionali o meno, ripetibili, che consentono una descrizione formale, intellettuale e giuridica del documento⁴².

Google Books utilizza uno standard complementare di descrizione che consente di identificare il possibile livello di visualizzazione dell'opera (*viewability*), la sua possibilità di esportazione su siti terzi (*embeddability*), ed i commenti associati (*review*).

Ma l'utilizzo di uno standard omogeneo non è sufficiente di per sé a garantire la qualità dei metadati.

Il carattere restrittivo dello standard rischia di obbligare le biblioteche digitali a semplificare i metadati di cui potrebbero disporre a monte, oppure ad assemblare dei metadati meglio strutturati in un'unica categoria. Questo è l'effetto imbuto. In questo modo, nell'elemento "formato" si troveranno degli elementi descrittivi che si riferiscono sia alla descrizione dell'opera fisica (in-8°, 16 pag.), che al file digitale (applicazione/PDF), informazioni queste non attinenti tra loro.

Lo standard non può, d'altra parte, colmare le lacune dei metadati iniziali. In Gallica, si nota anche la ripetizione nell'elemento "Titolo" del nome dell'autore, ripreso anche nell'elemento "Autore", e delle diverse persone che hanno contribuito all'opera. L'insieme delle persone che hanno contribuito ad un titolo non risulta come contributore (elemento "Contributore") ma anch'essi, sono riportati, senza alcuna distinzione nell'elemento "Titolo". Questa inadeguatezza del contenitore e del contenuto contribuisce a rendere meno pertinenti i risultati della ricerca ed a vanificare la ricerca di "Contributori" e gli smistamenti legati a questa categoria. Alla stessa stregua, pare che Google Books riporti nell'elemento "Titolo" informa-

⁴² L'autore è indicato dall'elemento "Creator", l'editore dall'elemento "Publisher", i contributori dall'elemento ripetuto "Contributor", la provenienza del documento dall'elemento "Source", ecc.

zioni relative alla collezione o alla serie di cui fa parte l'opera (13^{mo} volume dei *Classiques Vaubourdolles*): questa scelta è poco strutturale, anche se riconosciamo il vantaggio di una tale flessibilità in termini di efficacia e rapidità.

Poiché le fonti delle informazioni delle due biblioteche digitali sono molteplici ed eterogenee, la standardizzazione dei metadati rimane un elemento problematico negli stabilimenti della digitalizzazione di massa⁴³. Sarebbe auspicabile che tale standardizzazione passasse dall'utilizzo di MADS unificati, cosa che invece non accade. Questo difetto è causa di errori e pregiudica la creazione di collegamenti tra i libri (libri del medesimo autore, per esempio), come anche la qualità delle azioni eseguibili sulle liste di opere (classificazione o filtro per autore). Questa eterogeneità è evidente nell'offerta di Gallica relativa alle opere coperte da diritti, in quanto le fonti di informazione, provengono, senza alcun controllo, dagli editori e dai divulgatori stessi.

Il cuore del sistema: la qualità dei metadati

La realizzazione delle biblioteche digitali deve necessariamente essere accompagnata da una correzione delle notizie bibliografiche create in base alla conversione retrospettiva dei cataloghi cartacei, affinché l'internauta possa usufruire appieno della loro valorizzazione documentaria. Al cambiamento dei supporti corrisponde necessariamente un cambiamento degli elementi che li descrivono e questo ad un triplice scopo: l'interoperabilità dei cataloghi tra biblioteche digitali, il miglioramento della pertinenza semantica dei motori di ricerca, la realizzazione del Web semantico (OWL, RDF/XML du W3C). Questa triplice sfida rientra totalmente nel perimetro delle attività condotte

dalle istituzioni preposte alla conservazione ed alla valorizzazione dei patrimoni. Le biblioteche coinvolte devono capire che la loro partecipazione attiva è fondamentale perché questo è e sarà il loro valore aggiunto di oggi e di domani. La scarsa qualità dei metadati di Google Books si esplicita principalmente nel tipo di imprecisioni sopra descritte (eterogeneità del trattamento delle categorie "redattore" o "autore", assenza di informazioni sulla provenienza, ecc.). Ma questo non si riferisce alle aberrazioni spesso citate dagli osservatori, soprattutto per quanto concerne le datazioni e le categorizzazioni delle opere. Infatti, tali errori sono molto rari in Gallica la cui referenziazione è molto più affidabile, in quanto collegata in buona parte al catalogo della BnF. Per esempio, Gallica sa gestire molto meglio le serie di pubblicazioni rispetto a Google Books.

È difficile stabilire con certezza se questi errori di Google Books provengano dai metadati raccolti presso la biblioteca prima della digitalizzazione, oppure da un'attività complementare di indicizzazione insita nel processo della digitalizzazione⁴⁴. La frequenza di questi errori rivela anche la fragilità della filiera di indicizzazione di Google Books e la mediocrità del controllo qualità condotto dalle biblioteche partner.

Questi punti deboli dei metadati, al di là delle incertezze documentarie che provocano nell'ambito degli utilizzi tradizionali dei ricercatori, comportano un rischio giuridico. Infatti, gli errori di datazione sono all'origine della messa a disposizione gratuita di opere considerate di dominio pubblico, che però si rivelano essere ancora coperte da diritti. La responsabilità di Google Books e delle biblioteche partner potrebbe pertanto essere chiamata in causa; Gallica ha assunto il rischio⁴⁵.

⁴³ Si veda, per esempio, il trattamento dell'autore Théodore de Banville in Google Books.

⁴⁴ Quando, per esempio, Google Books per la scheda associata al *Dictionnaire de la noblesse* del 1775 descrive un'opera su *Le Génie des eaux chez les Dogons*; o una copia della *Revue musicale* di Henry Prunières e di André Cœuroy, dove si cita varie volte Maurice Ravel e de Paul Dukas, viene datata 1827...

⁴⁵ Quando, per esempio, considera erroneamente di dominio pubblico il *Pour Thaelmann* (Éditions universelles, 1935) che contiene, tra l'altro, discorsi di André Gide e di André Malraux.

Controllo qualità

Non si darà mai troppa importanza al controllo qualità condotto sui metadati associati ai patrimoni digitalizzati, controllo che spetta sia all'operatore della digitalizzazione, il quale sottoscrive un Impegno sulla qualità del servizio, che allo staff delle istituzioni preposto alla revisione delle opere digitalizzate. È ormai chiaro che i criteri della digitalizzazione di massa sottovalutano l'importanza di questo lavoro.

Per esperienza, si sa che l'accesso immediato alle pagine delle opere permette di correggere e completare *de visu* i metadati. L'utente potrà così accontentarsi di una qualità discreta di descrizione dei libri messi a disposizione dalle biblioteche digitali. Ma alcuni errori non sono "compensabili" dall'internauta, nella misura in cui incidono direttamente sulle modalità di trattamento delle richieste formulate tramite i motori di ricerca: è possibile correggere solo quello che si può vedere.

Da notare che i metadati delle notizie bibliografiche associate alla presentazione delle opere sono riportati nelle proprietà stesse dei file digitalizzati, come elementi descrittivi interni al file (proprietà di un PDF, per esempio). La divulgazione dei file implica conseguentemente la divulgazione degli errori di indicizzazione.

I motori di ricerca

Allo stato attuale, Gallica utilizza il motore di ricerca *open source* Lucene (Java, Fondation Apache), dedicato alla ricerca a tutto testo. In Gallica, Lucene effettua le sue ricerche in base ad un indice unico preconstituito, senza interagire con alcuna base di dati.

La scelta di Lucene si rivela piuttosto operativa per quanto concerne i volumi effettivamente

trattati e le prestazioni desiderate, tuttavia deve essere riconsiderata se Gallica vuole fare evolvere i suoi strumenti documentari, soprattutto con l'introduzione di strumenti di analisi semantica più avanzati. Ciò richiede il potenziamento di Lucene o l'adozione di un altro motore di ricerca. Lo staff tecnico di Gallica ritiene prioritario questo obiettivo. Per quanto concerne la questione delle prestazioni e della loro misurazione, sembra che Gallica, dinanzi a ricerche complesse o basate su espressioni lunghe, fatichi a fornire risultati pertinenti in tempi simili a quelli offerti da Google Books, sebbene le dimensioni del *corpus* gestito dal portale nazionale siano decisamente inferiori. Oltre alla robustezza dell'infrastruttura ed alla potenza del motore di Google Books, questa differenza può anche essere percepita come l'espressione di scelte documentarie diverse: la messa a punto dell'intelligenza documentaria (smistamenti, faccette) non è sempre compatibile con la rapidità dei tempi di risposta. È la classica dialettica tra pertinenza e performance, un aspetto che diventerà particolarmente critico alla luce della crescita del *corpus* digitalizzato previsto per Gallica.

Allo stato attuale, l'approccio di Gallica non è privo di ambiguità rilevanti, che pregiudicano la comprensione immediata da parte dell'utente degli elementi di risposta forniti dalla biblioteca⁴⁶. Ne scaturisce un'impressione di inefficienza di Gallica, mentre in realtà si tratta solo di scelte di parametri e di sviluppi del tutto rettificabili. Questi effetti hanno forse pregiudicato la reputazione di Gallica, che si è comunque sforzata e si sta sforzando tuttora per correggerli.

Google Books e Gallica propongono due livelli di ricerca, semplice o avanzata; quest'ultima

⁴⁶ Per esempio, nel caso di una ricerca effettuata su un insieme di parole (una citazione, come: «Ausein de l'infini, nous élançons notre être», tratta dal *Poème sur le désastre de Usbonne* di Voltaire), Gallica risale alle opere pertinenti, come Google, ma non dà priorità alla restituzione immediata della sola frase. Gallica indica le opere in cui si trova la citazione; è solo selezionando una delle opere e poi cliccando sulla pagina proposta che si arriva alla citazione. Quindi, ci vogliono tre clic, contro un solo clic di Google Books. La percezione della pertinenza da parte dell'internauta è ulteriormente disturbata dal fatto che ogni parola riportata nella citazione è sottolineata quando il motore la trova nei metadati e nell'opera.

consente di incrociare le ricerche effettuate sul testo dei libri stessi con quelle sui metadati e quindi di affinare a monte le ricerche. La qualità dei metadati diventa quindi un *atout* di prim'ordine. È in questo ambito che si potrebbe deplorare, come precedentemente notato, che, al momento della digitalizzazione, non sia stata condotta un'analisi più approfondita delle strutture della pagina o dei libri o che questa non sia sfruttata meglio dal motore; non è nemmeno possibile escludere parti di libri che rischiano di confondere l'indicizzazione. Una digitalizzazione di massa può consentire questi tipi di approccio, senza pregiudicare la produttività della filiera.

L'algoritmo di ricerca dell'una e dell'altra biblioteca digitale, le leggi di ponderazione su cui si basano le modalità di presentazione delle opere nelle liste prima della loro consultazione, sono poco chiare per gli utenti. Per esempio, non è mai indicato chiaramente quante volte la parola o l'espressione ricercata è stata trovata nei libri proposti. Si sa che uno dei principi di base del motore Google consiste nell'applicazione di coefficienti di popolarità (frequenza degli accessi e dei link) e di "affidabilità" che, nel campo della ricerca sui libri, può tradursi nella valutazione del grado di citazione di un libro rispetto agli altri libri conservati nella biblioteca digitale. Si comprende quindi che il livello del *corpus* raccolto gioca a favore della performance del motore. A questo proposito, e non solo a questo, è difficile operare dei confronti significativi tra il motore di Gallica e quello di Google Books.

Nel caso di utilizzi specifici per i quali i livelli di ponderazione impliciti non sono sufficienti, l'assenza di strumenti di trattamento documentario dipendenti dalla richiesta è particolarmente problematica. E questo è uno dei principali punti deboli di Google Books, parzialmente compensato dalla rapidità dei tempi di risposta e dalla semplicità di utilizzo, che facilita il reperimento spontaneo da parte dell'internauta che sa cosa cerca. Tuttavia, Google Books lascia talvolta l'utente scettico per le sue

scelte di parametri. Per esempio, da una ricerca rapida su "Montesquieu" si evince che Google basa la sua ponderazione principalmente sulla presenza della parola nel metadato "titolo" dell'opera e che non sa identificare "Montesquieu" in primo luogo come autore. In questo modo, le opere di Montesquieu rimangono isolate nella moltitudine di testi su Montesquieu, senza che sia possibile, peraltro, capire cosa ne determina l'ordine di presentazione. Non vi sono restrizioni per faccette che consentono di selezionare le notizie bibliografiche sul solo criterio di "Montesquieu" come autore, in base ad un'unica grafia valida; al contrario di quello che Gallica, come anche molti altri strumenti bibliografici in commercio (Electre, per esempio o Cairn) sa fare.

La scarsa visibilità dei parametri utilizzati dal motore di ricerca pone una maggiore difficoltà, per ragioni che, si sa, esulano all'ambito puramente documentario e catalografico, ma sono riconducibili alla responsabilità sociale d'impresa.

Si noterà infine che Google Books integra delle funzioni di aiuto alla ricerca simili a quelle offerte dal motore di ricerca Web Google, anche se non sono sempre pertinenti (proprio perché si appoggiano ad un *corpus* Web non strutturato e non dedicato all'universo del libro). Sono attive sia le correzioni ortografiche automatiche che l'autocompletamento delle espressioni scelte dall'internauta. La differenza tra le due biblioteche è quindi molto marcata su questo aspetto e per una richiesta come "alain rabbe rillet", Google Books propone di cercare "Alain Robbe-Grillet"; cosa che invece Gallica non è in grado, purtroppo, di fare.

Quale motore per quale scalabilità?

È necessario approfondire ulteriormente le considerazioni sul trattamento della ricerca per Gallica, allo scopo di conciliare i parametri del motore con le aspettative percettibili in termini di utilizzo.

La forte scalabilità del *corpus* di Gallica rappresenta una sfida importante nell'ambito delle scelte su cui si basa il miglioramento delle pre-

stazioni del motore di ricerca che sarà adottato in futuro. La sua capacità di trovare un nuovo punto di equilibrio tra performance e pertinenza sarà uno dei criteri che attirerà l'attenzione degli utenti. Bisogna tenere conto di questo aspetto ai fini del dimensionamento del *corpus* target di Gallica, soprattutto in considerazione della sua ambizione di raggruppare più patrimoni. Qui la posta in gioco è molto elevata.

Le restituzioni

Il primo livello di restituzione proposto dalle biblioteche digitali è rappresentato da una lista delle opere in grado di soddisfare la richiesta, semplice o complessa, formulata dall'internauta. Come già detto, la pertinenza e la rapidità di risposta sono tra gli elementi chiave di questa fase. Nel contesto della ricerca rapida, in tutte le biblioteche, le modalità di calcolo della pertinenza sono poco chiare; non si conoscono gli indici di ponderazione utilizzati dai motori di ricerca. Ne risulta un senso di incertezza per l'internauta, soprattutto quando la lista delle opere proposte è consistente.

Queste funzionalità si dividono in due tipi: classificazioni e filtri. Google Book è molto indietro in merito a questo tipo di offerta di servizio documentario per l'internauta. Non è possibile effettuare alcun smistamento sulle liste generate ed i filtri proposti permettono di operare una sola sottoselezione sulla base dell'unico criterio di accessibilità del file digitalizzato. Gallica propone invece, una vasta gamma di smistamenti e filtri (o faccette), tecnicamente fattibili (nonostante qualche bug persistente come, talvolta, doppioni di opere) e si basa su metadati più omogenei ed affidabili.

Google, dal canto suo, propone qualche strumento, diverso da quelli classici, che consente di farsi un'idea sul contenuto del libro. A questo scopo, ad ogni libro (compresi quelli coperti da diritti, non messi a disposizione) è associato una nuvola di parole chiave, risultato dell'analisi del contenuto testuale dei libri. Vengono così evidenziate delle parole chiave (nomi di persone e personaggi, nomi di città, nozioni) che

permettono un approccio tematico delle opere. Ma questo metodo a volte produce risultati dubbi, poiché l'analisi automatica è limitata dalla cattiva gestione dell'inquinamento, quando per esempio, gli elementi di una pagina che raggruppano gli altri autori pubblicati dallo stesso editore sono considerati come parole chiave dell'opera, come nel caso di una raccolta dei *Poèmes élégiaques* di Laurent Tailhade. Ancora una volta, le scelte della digitalizzazione a monte sono inseparabili dalle modalità di restituzione a valle e quindi dagli utilizzi.

Le funzioni di aiuto alla ricerca, un elemento di differenziazione

L'assenza di strumenti documentari adatti alla gestione delle liste su Google Books è sorprendente. Rinforza il senso di confusione e tradisce l'assenza di controllo documentario del patrimonio digitale proposto dai motori di ricerca, non compensato dalla pertinenza e dalla potenza del suo algoritmo. In termini di utilizzo, può indicare un posizionamento di Google Books presso pubblici che hanno meno familiarità con questi strumenti documentari, o comunque maggiormente orientati ad una ricerca precisa di informazioni (comoda ricerca di citazioni...), piuttosto che ad un approccio più esperto della ricerca.

Gallica deve continuare a rinforzare questa dimensione che diventerà una sfida tecnologica importante con l'aumento del numero di riferimenti disponibili e che rappresenta un elemento chiave di differenziazione tra le due biblioteche digitali.

Google Books propone un accesso limitato alle opere considerate coperte da diritti e per la promozione delle quali non ha firmato accordi di partenariato con gli editori, sotto forma di *snippets* (al massimo tre per ogni espressione cercata), che si presentano come pezzi di pagine strappate. Da notare che questo tipo di restituzione, più o meno apprezzata dagli utilizzatori, che si appoggia sull'analisi delle coordinate grafiche delle parole trovate nel *text layer* del file, privilegia un'ottica di sicurezza piuttosto

sto che una logica di utilizzo. In effetti, gli *snippets* sembrano essere degli estratti già ritagliati; questo significa che una determinata parola o frase si troverà sempre nello stesso *snippet* che quindi non viene rielaborato o composto “al volo”. È sufficiente quindi che una parola sia contenuta in una riga sul bordo tagliato perché non sia visualizzata. Basta guardare questi due esempi, in cui è stata cercata l’espressione “Mais un regard” e poi l’espressione “La condition humaine” riportata nella copia dei *Romans* di Malraux (copia della Bibliothèque de la Pléiade) proposta da Google Books:

Con questo sistema, Google Books si assicura che un libro non possa essere recuperato integralmente sulla base una serie di richieste che consentirebbe di ricostituire una pagina riga dopo riga. Ma, in termini di utilizzo, il risultato non è soddisfacente.

La scelta della restituzione sotto forma di *snippets* è probabilmente giustificata dalla mediocre qualità del file di testo che era inizialmente associato alle immagini; l’immagine invece, se posizionata correttamente, rimane sempre presentabile. Tuttavia, si accede agli *snippets* solo dopo aver consultato una lista di risultati di libri che può, sotto ogni elemento di scheda, restituire una frase del testo grezzo. Ed è proprio questo livello di restituzione che aveva messo in allarme gli editori sullo stato reale dei testi digitalizzati da Google e sul rischio per l’integrità delle opere.

Da parte sua, Gallica si rifiuta di andare al di là della restituzione di un estratto in formato testo per le opere coperte da diritti depositate dagli editori. Potrebbe invece farlo, dal momento che queste opere sono, come i libri di pubblico dominio, indicizzati da Lucene, sulla base di un file di testo parzialmente strutturato

messo a disposizione di Gallica dagli editori partner. Per la consultazione degli estratti, quest’ultima rimanda all’applicazione di ricerca indicata dall’editore, e non all’interno del sito Gallica, applicazione che può essere diversa da un editore all’altro e che potrebbe essere standardizzata nel futuro, con l’evoluzione dei partenariati.

La consultazione dei libri

L’utente può accedere ad alcune pagine o all’opera intera. L’applicazione che consente questa modalità di consultazione varia da un sito all’altro. Google Books offre la successione delle pagine in sequenza verticale, proponendo anche altre forme di configurazione (scacchiera, vignette, ecc.) Per quanto concerne i libri coperti da diritti pubblicati in virtù del partenariato con gli editori, alcune pagine non sono consultabili. Google Books utilizza la tecnologia JavaScript per visualizzare le immagini delle pagine, il che evita di dover scaricare un plug-in per accedervi. Google Books fa anche in modo di non rendere accessibile, in questa modalità di visualizzazione, il *text layer* delle opere, per mezzo di un trattamento in back-office della concordanza tra testo e immagine. È possibile, recuperare, tramite attività di programmazione per mezzo dell’applicazione di ricerca l’insieme delle immagini di un libro, ma non è possibile recuperarne il *text layer*, se Google Books non lo rende disponibile nell’altra modalità di visualizzazione (testo grezzo, per gruppo di pagine), proposta per una parte dei libri di dominio pubblico, peraltro scaricabile in formato ePub o PDF (e quindi potenzialmente accessibile dai motori di ricerca esterni), ferme restando le misure di protezione applicate da Google Books per evitare le richieste automatizzate⁴⁷. I tempi di visualizzazione sono eccellenti.

⁴⁶ Per esempio, nel caso di una ricerca effettuata su un insieme di parole (una citazione, come: «Au sein de l’infini, nous élançons notre être», tratta dal *Poème sur le désastre de Usbonne* di Voltaire), Gallica risale alle opere pertinenti, come Google, ma non dà priorità alla restituzione immediata della sola frase. Gallica indica le opere in cui si trova la citazione; è solo selezionando una delle opere e poi cliccando sulla pagina proposta che si arriva alla citazione. Quindi, ci

In Gallica, la consultazione è sequenziale, pagina dopo pagina, ed è, in via teorica, molto simile a quella di Google: accesso alle immagini con occorrenze sottolineate, accesso al testo grezzo, per pagina, e con, lo ribadiamo, visualizzazione della percentuale di riconoscimento dei caratteri, accesso al testo letto, con sintesi vocale. Più recentemente, è stato messo a punto un nuovo tipo di consultazione, basato sulla tecnologia Adobe Flex, con logiche simili a quelle del navigatore *flash* ed in grado di rispondere alla promessa di una maggiore fluidità di consultazione (tramite *prefetch*⁴⁷), di un miglioramento delle funzionalità di zoom, ecc. L'uso che ne faranno gli internauti dimostrerà se si tratta di un vero progresso.

Si sa che Google Books tenta di mantenere il controllo sull'accesso al *text layer* delle opere della sua biblioteca digitale per mezzo di obblighi contemplati nei contratti con i suoi partner. Resta il fatto che l'applicazione consente un livello di accesso al testo delle opere di dominio pubblico uguale a quello di Gallica. Questo conferma che il controllo esercitato da Google Books è volto essenzialmente ad impedire non tanto l'appropriazione occasionale quanto l'appropriazione massiccia dei contenuti testuali che potrebbero sfuggire al suo controllo. E conferma anche che la biblioteca digitale

“pianta i paletti” tecnici necessari per proteggersi da tali abusi negli accordi sottoscritti con i suoi partner.

Il download dei libri

Oltre alla funzione di consultazione online, le biblioteche digitali offrono la possibilità di scaricare i libri nella loro versione integrale, in formato PDF (in bassa definizione, cioè 75 dpi) o Epub. Per quanto concerne il PDF, Google Books propone solo un file immagini non interrogabile dall'internauta, preceduto da un avviso sulle condizioni di sfruttamento del file in questione e filigranato con il marchio del motore di ricerca. Il download è molto facile. Anche Gallica offre un PDF, ma con doppio strato (immagine/testo) che permette quindi di interrogare il testo. Pertanto, anche in questa fase si capisce la “prudenza” di Google Books, volta a proteggere i suoi contenuti testuali. Questi stessi file potrebbero essere sottoposti a OCR al fine di ottenere un *text layer*. Google Books propone il download di un file Epub per un gran numero di opere di dominio pubblico. Questi file si basano sulle stesse fonti di digitalizzazione e quindi riproducono integralmente gli errori dell'OCR che, per tutta una serie di opere, e soprattutto per le più vecchie, rende l'esperienza della lettura molto “particolare”, sebbene Google Books lo neghi⁴⁹.

vogliono tre clic, contro un solo clic di Google Books. La percezione della pertinenza da parte dell'internauta è ulteriormente disturbata dal fatto che ogni parola riportata nella citazione è sottolineata quando il motore la trova nei metadati e nell'opera.

⁴⁷ «Questi libri sono di proprietà [...] di tutte le donne e di tutti gli uomini e noi siamo solo i custodi di questo patrimonio. Tuttavia, si tratta di un progetto costoso. Conseguentemente, e al fine di favorire la diffusione di queste inesauribili risorse, abbiamo preso le misure necessarie per prevenire i possibili abusi da parte dei siti di rivenditori terzi, in particolare implementando delle restrizioni tecniche sulle richieste automatizzate. Vi preghiamo [...] di [...] *non procedere a richieste automatizzate*. Non inoltrate alcun tipo di richiesta automatizzata al sistema Google. Se le vostre ricerche hanno per oggetto i software di traduzione, il riconoscimento ottico dei caratteri o qualsiasi altro settore per il quale sia necessario disporre di grandi quantità di testo, non esitate a contattarci. Per questo tipo di lavoro noi incoraggiamo l'utilizzo di opere e documenti di dominio pubblico e saremmo ben lieti di esservi utili». (Estratto dell'avviso riportato in un file PDF scaricato da Google Books).

⁴⁸ [Tecnica usata nei microprocessori per accelerare l'esecuzione dei programmi riducendo gli stati di attesa.]

⁴⁹ «Despite our best efforts you may see spelling mistakes, garbage characters, extraneous images, or missing pages in this book. Based on our estimates, these errors should not prevent you from enjoying the content of the book. The technical challenges of automatically constructing a perfect

Il crollo qualitativo che costituisce di fatto questo trattamento pone un interrogativo sul futuro di queste biblioteche digitali agli albori del XXI secolo e sulla natura della rielaborazione a cui noi o le future generazioni dovranno sottoporre i file o persino i libri stessi. In ogni caso, relativizza l'impronta di perennità che si vorrebbe attribuire a queste fonti.

Ecco, per esempio, una pagina doppia di un file Epub di un libro francese dell'inizio del XIX secolo proposto da Google Books⁵⁰.

Complementi interattivi

Uno degli elementi che distingue l'offerta di Google Books, oltre al dimensionamento del suo *corpus* multilingue, risiede nel collegamento che effettua tra i libri (del tipo, questo libro è citato in questi altri libri) e con basi di dati o fonti esterne: Worldcat di OCLC, Sudoc per il mondo delle biblioteche (al fine di consentire l'identificazione delle copie presenti in altre biblioteche), eventualmente con Gallica o le risorse del Web

(siti di fan club degli autori, ecc.) con biblioteche di testi antichi e di novità editoriali, con siti degli editori. Pertanto, questi link sono associati ad un utilizzo molto dubbioso, vista l'imperfezione in termini di coerenza ed affidabilità dei rinvii da una scheda all'altra. Ma, se non altro, ciò testimonia un'innegabile volontà di apertura di questa biblioteca digitale.

Le due biblioteche propongono anche strumenti grazie a cui è possibile apportare annotazioni sui volumi, tramite spazi personali o spazi per i commenti. Anche a questo proposito, le funzionalità di Gallica sono alquanto avanzate rispetto al meccanismo di Google Books.

Gallica e Google Books propongono delle possibilità di integrare link ai siti Web. Ma Google Books va più lontano, proponendo già degli API che consentono un'interazione più avanzata dei suoi contenuti con altri siti di risorse (librerie, biblioteche, blog, ecc.), per esempio tramite un sistema di vignetta esportabile, tipo



book are daunting, but we continue to make enhancements to our OCR and book structure extraction technologies. We hope you'll enjoy these books as much as we do». (Estratto dell'avviso riportato nei file Epub di Google Books).

⁵⁰ Pierre-René Auguis, *Les Poètes français depuis le XIIe siècle jusqu'à Malherbe* dovevano tome 4, Imprimerie de Crapelet, 1824, collezione dell'Università del Michigan.

embed. La funzione “selezionare” che appare nelle pagine di presentazione del libro, consente, per esempio, di integrare in un sito terzo un link all’estratto della pagina di un libro. Google Books può quindi sfruttare al massimo l’utilizzo dei suoi servizi nell’ambito delle pratiche di indicizzazione e di ponderazione dei contenuti più popolari.

I servizi di correzione collaborativa dei metadati e dei *corpus* testuali sono più limitati; non è presente alcuna funzionalità di tipo “wiki”. Google Books propone un servizio minimalista di segnalazione degli errori. Gallica ha intrapreso una sperimentazione in tal senso con Wikipedia e Wikisource su una parte del suo *corpus* (1400 documenti).

I link interattivi includono anche i link pubblicitari proposti da Google Books, che appaiono sia nelle liste dei risultati che nei file di presentazione ed anche nell’applicazione di consultazione delle opere. Si tratta di link contestuali, nei limiti della definizione di contestualità del motore americano. La richiesta sulla frase già citata del poema di Voltaire sulla Provvidenza, «Au sein de l’infini, nous élançons notre être», comprende dei link pubblicitari ai siti della Libreria Amazon o alle automobili “infiniti” (presenti anche sul motore Bing di Microsoft), o ancora verso un sito che propone soluzioni estetiche per l’aumento del volume mammario (ritenuto pertinente in base alla parola chiave “seno”) e verso un sito confessionale di preghiere online. Deve essere ben chiaro che qualsiasi accordo di partenariato con Google Books è destinato a promuovere tali campagne, e che gli eventuali proventi, sempre che si trovino accordi di ripartizione dei proventi con Google, sarebbero generati da link come questi, che non hanno nulla a che vedere con quella che vogliamo chiamare promozione del patrimonio culturale nazionale. Proviamo ad immaginare la reazione dei lettori della BnF se si infilasse un opuscolo pubblicizzante la chirurgia estetica in un volume della sala lettura.

La referenziazione

La sezione Books è un tab specifico nella pagina del motore di ricerca Web Google. I risultati prodotti da Google Books possono essere ottenuti tramite una ricerca semplice sul motore Web, ma non in maniera esaustiva; sembrerebbe quindi che Google indicizzi i dati di Google Books, compreso il testo completo delle opere, nell’ambito della sua attività di indicizzazione generale del Web.

Una volta formulata una richiesta sul motore Web, è possibile lanciare la stessa richiesta sia nella casella di selezione “Libri” che nella casella di selezione “Immagini” (link “Visualizza opzioni” e poi “Libri”). È in questa fase che appaiono i link pubblicitari di Google Books. Accedendo direttamente a Google Books si ha un’altra tipologia di accesso, ma si ottengono risultati identici all’accesso precedente (con link pubblicitari diversi).

Se vi è una correlazione stretta tra Google e Google Books, rimane il fatto che Google non può essere considerata come una modalità di accesso sufficiente ai contenuti della biblioteca digitale. Inoltre, i contenuti di Google Books non sembrano essere correttamente indicizzati dagli altri motori Web, come Yahoo! o Bing. Le possibilità di accesso al testo delle opere su Google Books sono solo relativamente limitate; ma Google dichiara, nei suoi avvisi, di aver applicato misure di protezione dei contenuti allo scopo di limitarne l’indicizzazione massiccia da parte di terzi.

I contenuti di Gallica (liste di documenti e libri) non sono indicizzati dai motori di ricerca del Web (Google, Yahoo!, Bing). E questa è una delle problematiche principali che la biblioteca digitale della BnF si trova oggi ad affrontare. A tal fine, sono previste numerose strategie: moltiplicare gli accessi alla base dati tramite la messa a punto di link editoriali?, integrare nelle pagine Web documenti di metadati strutturati che consentano un migliore riconoscimento degli elementi da parte dei motori di ricerca (utilizzo del meta Dublin Core) nel quadro dello sviluppo del Web semantico, divulgare i suoi

metadati (tramite un server OAI-PMH aperto, che consente per esempio a WorldCat di integrare le descrizioni delle notizie bibliografiche BnF). Quest'ultimo approccio, limitato dall'utilizzo del protocollo utilizzato (OAI-PMH), deve essere completato da uno spiegamento massiccio ed automatizzato delle pagine HTML collegate che contengono queste informazioni (il cosiddetto progetto del "pivot documentario" della BnF) e che le rende facilmente indicizzabili dai motori di ricerca. Lo studio preliminare è stato affidato a Cap Gemini. Infine, si procederà ad un'ottimizzazione dell'indicizzazione delle pagine del sito, tramite il classico utilizzo delle istruzioni impartite ai robot di indicizzazione (robot.txt e sitemap.xml). Questo rimangiamento dovrebbe consentire una migliore indicizzazione delle pagine con contenuto testuale, anche se il loro numero (una pagina Web per ogni pagina di un libro), rappresenta un freno non indifferente, legato alla concezione stessa del sistema di consultazione. Attualmente è in corso una fase di test, in seno al progetto europeo TELplus. Tale test dovrà essere accompagnato da un nuovo impegno collaborativo con i grandi motori di ricerca del Web. Gallica sta anche lavorando su strumenti di bookmarking che dovrebbero favorire la divulgazione dei suoi contenuti nei social network del Web.

L'insieme di questi meccanismi rappresenta una delle grandi sfide relative alla performance di Gallica per il 2010. Deve poter compensare l'insufficienza dei risultati ottenuti sinora.

La digitalizzazione di massa è un percorso fattibile, ma non l'unico, per la digitalizzazione; le sue limitazioni, così come i suoi innegabili vantaggi devono essere integrati da considerazioni approfondite sul senso della missione storica delle biblioteche, sia in termini di conservazione che di valorizzazione. Questa è una delle condizioni necessarie per non perdere il filo conduttore del dibattito.

Allegato 4: Lista delle biblioteche europee che hanno aderito al programma Google Recherche de livres

Aggiornata al 21 dicembre 2009 (Fonte: <http://www.google.fr/googlebooks/partners.html>, consultata il 21/12/2009)

- Germania: Biblioteca di Stato della Baviera (2007)
- Belgio: Biblioteca universitaria di Gand (2007)
- Spagna: Biblioteca nazionale di Catalogna (2007); Biblioteca dell'università Complutense di Madrid (2006)
- Francia: Biblioteca municipale di Lione (2008)
- Regno Unito: Biblioteca Bodleienne (Università di Oxford) (2004)
- Svizzera: Biblioteca cantonale ed universitaria di Losanna (2007)