DigiBESS: una biblioteca digitale open source

GIORGIO BERTOLLA*
GIANCARLO BIRELLO**
ANNA PERIN***

Architetture aperte per l'archiviazione e la conservazione a lungo termine di opere digitali

Nel 2005 il gruppo di cooperazione BESS (Biblioteca elettronica di scienze sociali del Piemonte), oggi composto da 17 soggetti,[1] ha avviato un progetto di digitalizzazione sperimentale con l'obiettivo di pubblicare online una bibliografia di base di 100 volumi per documentare lo sviluppo sociale ed economico del Piemonte negli ultimi cinquant'anni.[2]
Tre anni dopo, concluso con successo il progetto, lo stesso gruppo di biblioteche ha avviato una seconda fase, più ambiziosa, del programma iniziale. Obiettivo della nuova iniziativa è stato quello di dotare il gruppo di un laboratorio in grado di trattare fondi di dimensioni molto più consistenti. Ci si è cioè posti la domanda se l'attuale contesto tecnologico consentisse anche a soggetti pubblici non specializzati di misurarsi efficacemente con un'attività normalmente delegata al mercato. Ma su questo tema facciamo un passo indietro.
La tecnologia per la riproduzione digitale di opere a stampa e la loro pubblicazione online su piattaforme interrogabili full-text è nei suoi contorni di base sostanzialmente consolidata. È sufficiente acquistare uno scanner di piccole dimensioni, affittare uno spazio web e disporre di conoscenze informatiche elementari per mettere a disposizione piccole collezioni sul web. Questa affermazione vera, ma drastica, banalizza un quadro assai più complesso e soprattutto in continua evoluzione. Perché in questo campo l'autentico salto di qualità sta nella dimensione dei fondi che si intende digitalizzare e nelle opzioni di ricerca che si vuole offrire all'utenza.

Non a caso iniziative di pubblicazione di grandi fondi sono relativamente rare, non solo in Italia, ma anche nel resto dell'Europa. Un notissimo ed interessante esempio in merito è il progetto Europeana.eu., che avrebbe dovuto essere la risposta europea a Google. Partita con un solido appoggio politico, un budget altrettanto significativo, partner qualificatissimi e soluzioni tecniche assai avanzate, Europeana si trova oggi a fare i conti con un grave problema di sostenibilità. L'architettura progettuale, basata sullo sviluppo di un portale federato di diversi archivi/biblioteche online, è considerata ed appare tuttora come un intelligente compromesso tra risorse importanti, ma non infinite, e un obiettivo davvero ambizioso. Eppure molto di recente proprio chi aveva maggiormente sostenuto l'iniziativa avanza dei dubbi e riconosce che l'asticella è forse stata collocata troppo in alto. Così si legge infatti nel *Rapporto* di Hervé Gaymard al Senato francese (18 gennaio 2012):

> Cependant, le portail Europeana n'est pas aujourd'hui à la hauteur des espérances que le projet a suscitées. D'une part, la vision très large du contenu à numériser dans Europeana ne permet pas de définir un périmètre d'action très précis. Loin de se limiter au domaine du livre qui reste minoritaire, Europeana vise l'ensemble du patrimoine culturel et comporte d'ailleurs principalement des images. D'autre part, Europeana ne bénéficie pas des moyens financiers nécessaires à la réalisation d'un projet de cette ampleur. Le portail souffre d'une faible fréquentation, d'un défaut d'organisation des données et d'un manque d'investissement de la part des États, chargés de financer la numérisation.[3]

Inoltre, se è relativamente semplice digitalizzare grandi fondi a scopo di preservazione, non si può dire altret-

* IRES - Istituto Ricerche Economico Sociali del Piemonte, <bertolla@ires.piemonte.it>
** Ceris CNR, <g.birello@ceris.cnr.it>
*** Ceris CNR, <a.perin@ceris.cnr.it>

tanto se l'obiettivo è pubblicarli e renderli efficacemente ricercabili online. Nello stesso tempo, l'evoluzione delle tecnologie porta con sé la promessa che anche biblioteche con budget relativamente contenuti potranno digitalizzare collezioni di grandi dimensioni in tempi relativamente rapidi.
Qualunque biblioteca voglia cimentarsi con questa attività deve rispondere alla domanda "make or buy" che, in altri termini, vuol dire: mi attrezzo e produco in casa o faccio fare da un fornitore esterno? La domanda continua ad essere valida, ma non può essere posta in termini assoluti. La risposta è, infatti: dipende. Dipende, cioè, da una serie non piccola di fattori economici, organizzativi, professionali, nonché di motivazione personale. La valutazione economica, da sola, può essere fatta rapidamente. Se i volumi da digitalizzare sono pochi e non si dispone già dell'attrezzatura, conviene rivolgersi ad un fornitore esterno. Di norma lo stesso vale se invece i volumi sono molti e non si vuole affrontare l'impegno della produzione diretta. In linea di massima la produttività assicurata da una ditta specializzata è più alta. L'aspetto organizzativo si riduce sostanzialmente alla disponibilità di personale che, soprattutto negli ultimi tempi, è difficile reclutare o riconvertire per questioni di budget. La questione relativa alla professionalità è un punto centrale. L'impegno richiesto dalla produzione in casa si giustifica in larga misura se uno degli obiettivi principali del lavoro è far crescere professionalmente il personale. Allora varrà la pena investire nel know-how necessario per disporre delle conoscenze necessarie a presidiare un'area di attività che probabilmente nel futuro sarà sempre più parte della filiera delle biblioteche/centri di documentazione. Le competenze acquisite torneranno utili anche quando si dovrà far ricorso a servizi esterni. Sarà possibile valutare con maggiore sicurezza l'offerta del fornitore e soprattutto il rapporto qualità/prezzo. Una considerazione essenziale da fare è se la biblioteca conta di svolgere il lavoro all'interno di un progetto che coinvolge più partner in rete o no.

Viaggiare in rete

Il gruppo di cooperazione BESS è partito da quest'ultima considerazione. La massa critica raggiunta dai partecipanti all'iniziativa rappresenta un potenziale di collezioni, di competenze e di utenza tale da giustificare l'investimento in una infrastruttura completa.
Soprattutto i partecipanti al progetto BESS costituiscono una rete territoriale e disciplinare. Non facciamo qui riferimento alla cooperazione che è qualcosa di, se non ben definito, certamente più spesso e articolato.[4] Il modello rete che caratterizza il gruppo piemontese consente una maggiore flessibilità, soprattutto una leggerezza funzionale che meglio si adatta ad alcune forme di collaborazione *purpose-oriented*. In tal modo è più semplice attivare la partecipazione di soggetti interessati a iniziative specifiche o a risolvere problemi. Tutto ciò purché l'approccio empirico-collaborativo prevalga su quello teorico ideologico e i partecipanti sappiano far prevalere logiche virtuose anziché opportunismi parrocchiali.[5]
La rete adottata fa perno su un doppio *pivot*: il territorio e la disciplina, cioè istituzioni collegate da prossimità geografica e una specializzazione comune nell'ambito delle scienze umane. Anche per quanto riguarda la digitalizzazione, la scelta di lavorare per area geografica e per affinità disciplinare costituisce un vantaggio evidente, una logistica semplificata e la possibilità di realizzare collezioni on line dedicate a un'utenza omogenea con interessi coerenti. Questo secondo aspetto consente con maggiore agio di valersi di conoscenze professionali ed accademiche che esulano da quelle strettamente biblioteconomiche e quindi di arricchire la biblioteca digitale con contenuti redazionali introduttivi o di commento che aiutino l'utente a orientarsi all'interno della materia.
L'idealtipo di rete a cui il progetto piemontese si è ispirato è quello della Bayerische Landesbibliothek Online sviluppato dalla Biblioteca di stato della Baviera insieme ad un network di altre biblioteche e istituzioni culturali bavaresi.[6] Naturalmente le dimensioni e la dotazione economica del progetto tedesco sono di tutt'altro ordine. Però l'idea di un centro di servizio con una rete di soggetti che contribuiscono a popolare un unico database/punto di accesso è uguale. L'architettura funzionale del programma piemontese prevede un unico centro di digitalizzazione e un unico repository costituiti da: laboratorio di digitalizzazione, hardware e software per il *post processing* (raddrizzamento, scontornamento e conversione formati delle immagini) e il riconoscimento ottico dei caratteri e infine da un repository che utilizza programmi totalmente open source.
Alle considerazione relative alla struttura di governo del gruppo bisogna aggiungere che tutta l'iniziativa è stata resa possibile dal finanziamento concesso dalla Compagnia di San Paolo di Torino che ha consentito l'acquisto del macchinario, la progettazione del software e il reclutamento del personale direttamente addetto alla scansione. L'impegno della Compagnia di San Pao-

lo è tanto più rimarchevole se si considera il fatto che il gruppo di cooperazione attivo da più di dieci anni non ha natura giuridica propria.[7]

Il laboratorio

La dotazione dell'attuale laboratorio, inaugurato nella primavera del 2011, è così costituita:
- scanner automatico Qidenus: per la scansione ad alta velocità di libri di formato max A4;
- scanner planetario Bookeye per grandi formati;
- 6 workstation per lo svolgimento di tutte le fasi della post-produzione;
- una NAS da 15 terabyte per l'archiviazione dei file dei volumi in lavorazione.

L'intero workflow prevede le seguenti fasi:
- scansione dell'originale;
- conversione dei file nativi in .tiff e .jpg ad alta e bassa risoluzione;
- riconoscimento ottico dei caratteri per la produzioni dei files .txt;
- preparazione dei file con i metadati bibliografici e strutturali per l'*ingesting*.

Tutto il materiale scansionato viene controllato per la sua integrità a diversi livelli durante il flusso di lavoro. La prima operazione, svolta tramite software proprietario nel caso del produttore dello scanner automatico Qidenus, verifica la qualità dell'immagine, il *deskewing* (l'orientamento della pagina) e lo scontornamento dello spazio di ripresa esterno alla pagina.
La seconda operazione, dopo la conversione del formato in .tiff, verifica la completezza del libro elettronico e la leggibilità dell'immagine in rapporto all'originale.
Il terzo controllo viene effettuato dopo la produzione dell'OCR per verificare nuovamente l'integrità e le dimensioni del file .pdf. I file .txt prodotti dal programma Abby Fine Reader, che effettua il riconoscimento ottico dei caratteri, non vengono sottoposti a controllo di qualità. La ragione risiede in una serie di considerazioni. Innanzitutto in ragione dell'estrema onerosità del controllo manuale, secondariamente dell'utilità marginale di un testo elettronico controllato a fronte della disponibilità del testo immagine. Infine, perché il livello di qualità del programma OCR è considerato più che adeguato.[8]
Terminata la produzione dei file immagine e testuali, viene prodotto un file per l'*ingesting* (vedi oltre), che contiene i metadati bibliografici e strutturali. Questi sono semplicemente due set di informazioni per identificare e agevolare la navigazione nel libro elettronico. I metadati bibliografici selezionati sono il set base DC. I metadati strutturali sono la tabella di corrispondenza tra i capitoli del libro e il nome del file immagine corrispondente.
L'operatore che svolge quest'ultima operazione è anche quello che effettua l'ultimo controllo di corrispondenza per "impacchettare" tutto il libro elettronico pronto per l'invio tramite un programma ftp (File Transfer Protocol) all'ufficio Ceris-CNR per l'*ingesting*.

Architettura e hardware[9]

L'ufficio IT del Ceris-CNR è stato incaricato di occuparsi di gestire tutta la parte successiva alla digitalizzazione, ossia disponibilità di spazio di memorizzazione per il deposito delle opere digitalizzate adatto alla gestione di grossi volumi di dati, sviluppo delle piattaforme server e implementazione del sistema di repository e infine progettazione, sviluppo e gestione del portale web[10] per la presentazione, ricerca e consultazione dei dati.
L'architettura nel suo insieme presenta una certa complessità per la varietà e la quantità di tecnologie coinvolte. Vi sono nozioni base sistemistiche di *clustering* per la realizzazione dello *storage* ridondato e di virtualizzazione, vi sono aspetti specifici di gestione di applicativi java e php e vi sono gli standard quali xml e xslt, metadati Dublin Core, applicati a oggetti digitali, immagini e testo, inseriti in modelli legati da relazioni semantiche.
Sono state adottate soluzioni quasi esclusivamente open-source che da un lato permettono la possibilità di sfruttare quanto di più innovativo è disponibile nella comunità con un notevole risparmio economico, dando la possibilità di concentrare la spesa sugli apparati e in particolare sull'espansione del numero di harddisk per ottenere la capacità di memorizzazione richiesta, ma per contro richiedono un grosso lavoro di ricerca e adattamento delle varie componenti per ottenere e soddisfare le esigenze specifiche del progetto.
Le principali apparecchiature coinvolte nello sviluppo del progetto sono i server dedicati alla realizzazione del cluster e quelli impiegati come hypervisor per ospitare le macchine virtuali. Il cluster a due nodi attivo/passivo è stato completamente realizzato dal Ceris-CNR. Lo *storage* rende disponibili i vari volumi in cui è suddiviso tramite protocollo iSCSI. Il server che ospita il repository, oltre alla partizione di sistema, è connesso a parti-

zioni direttamente sul cluster; in questo caso il backup consiste sempre nella sola immagine del disco di sistema essendo le partizioni sul cluster già ridondate implicitamente.
Tra le caratteristiche della rete che meritano un accenno è la raggiungibilità tramite protocollo IPv6 del repository e del sito web e inoltre va sottolineato il lavoro di sicurezza che inevitabilmente è richiesto per questo tipo di esposizione dei servizi.
Per lo spazio dati del repository è direttamente il server virtuale a collegarsi tramite iSCSI ai volumi offerti dal cluster senza intermediazioni. Su questa base sono memorizzati gli oggetti del repository, assimilati a dei BLOB (Binary Large Objects), il repository utilizza un sistema di store specifico per questi oggetti. In questo modo è garantita l'affidabilità del sistema, la ridondanza dell'informazione e una veloce procedura di ripristino in caso di guasto del server, essendo sufficiente connettere al volume iSCSI un nuovo server che avrà a disposizione l'intero store del repository.
Si è scelto di separare e distribuire su due server distinti le componenti relative al repository da quelle di presentazione del front-end. La suddivisione risponde alla logica del miglioramento delle prestazioni, della semplificazione delle operazioni di mantenimento dei server e della separazione delle applicazioni java da quelle php: le prime inserite in un unico contenitore Tomcat sul server del repository e le seconde dentro Apache sul secondo server previsto per la parte di front-end. Unica eccezione l'installazione, sul server del repository, del web server Apache con la funzione di reverse-proxy per le applicazioni installate in Tomcat.
All'interno di Tomcat sul server del repository troviamo varie applicazioni, in particolare: adore-djatoka, (trattamento immagini), FEDORA (il repository server), fedoragsearch (il motore di ricerca di fedora), fop (formattazione e rendering), iiv (il viewer online dei libri), imagemanip (trattamento immagini), saxon (trasformazioni xslt) e solr (motore di ricerca e indicizzazione).

I componenti del repository

Il cuore dell'intero progetto è sicuramente il repository, cioè il sistema di memorizzazione e gestione degli oggetti digitali. Il nome, FEDORA, è l'acronimo di "Flexible Extensible Digital Object Repository Architecture"; il software è stato sviluppato in Java ed è frutto della comunità open-source. In particolare il prodotto in sé è una base astratta che definisce un contesto e delle regole per la gestione di oggetti digitali selezionate pensando a sistemi di conservazione di opere digitali a lungo termine, in grado di gestire grossi volumi di dati e in modo flessibile per poter trattare i più svariati tipi di oggetti.
Nel repository gli elementi base sono gli oggetti che possiedono uno o più componenti chiamati "datastream", i quali sono a loro volta dei contenuti, ad esempio un'immagine, oppure metadati che descrivono l'oggetto. I datastream possono essere memorizzati localmente sul server o referenziati tramite un url esterno. Gli oggetti possono dichiarare una o più relazioni con altri oggetti all'interno del repository, tramite asserzioni semantiche del tipo *soggetto-predicato-complemento*: ciò costituisce sicuramente una delle caratteristiche più evolute che offre questa architettura di grande potenzialità.
Quanto descritto, finora abbinato alla possibilità di generare datastream virtuali, ad esempio la generazione di una miniatura da un'immagine ad alta definizione, in aggiunta all'identificazione univoca tramite namespace e Persistent Identifier (PID) degli oggetti, conferisce al prodotto il carattere di repository di contenuti fruibili via web durevoli nel tempo.

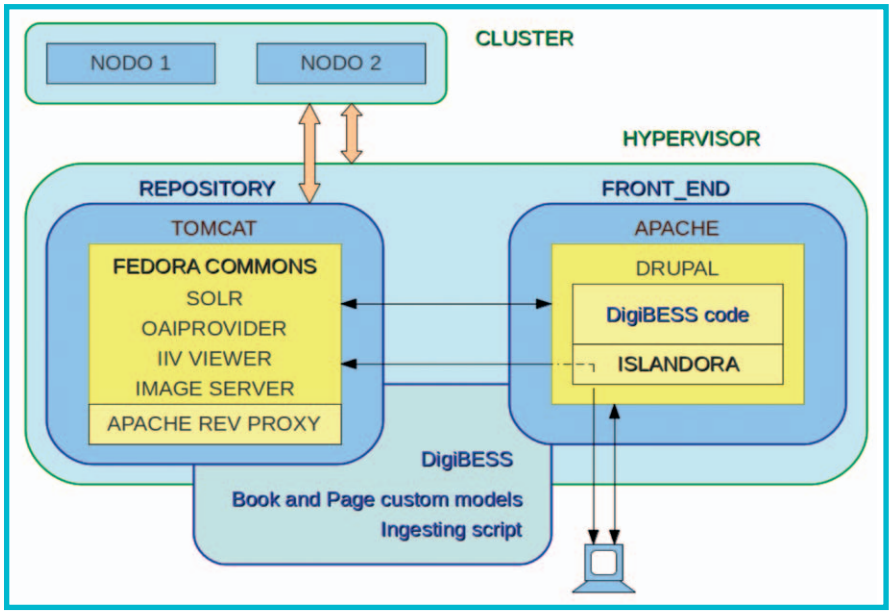

Figura 1 – Architettura

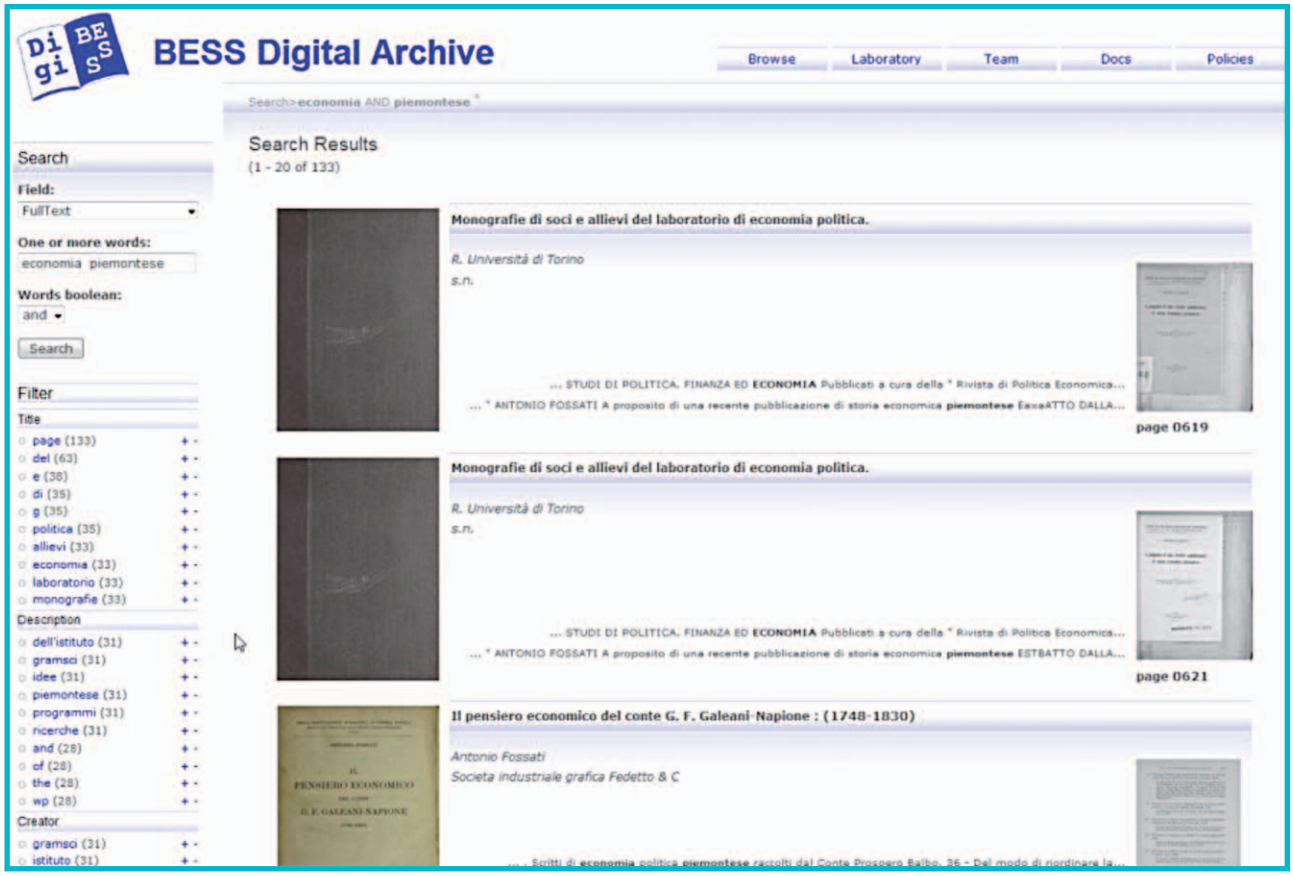

Figura 2 – Ricerca full-text e filtri

Il sistema offre la definizione personalizzata da parte dell'utente di modelli di oggetti potendo così sfruttare le potenzialità dell'ambiente adattato alle esigenze specifiche della soluzione. Un esempio tipo è proprio l'implementazione DigiBESS dove vengono utilizzati modelli quali collezioni, libri e pagine legati da semplici relazioni come "è membro della collezione" o "è parte di", oggetti che prevedono datastream funzione dei prodotti della scansione, ossia immagini ad alta risoluzione, testo OCR e file pdf ricercabili.

Una caratteristica essenziale e preziosa di FEDORA Commons è la possibilità di effettuare l'*harvesting* degli oggetti con interrogazioni OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), importante nell'ottica di condivisione dei contenuti del repository con altre entità. Tale possibilità è già stata sperimentata ed è una applicazione affidabile e funzionante, ad esempio con il portale OpenDOAR.

Come interfaccia web verso gli utenti per presentare gli oggetti digitali, nel nostro caso libri e riviste, è stato scelto Drupal, un CMS (Content Management System) open-source, largamente diffuso e già utilizzato in rete per l'accesso a repository in quanto versatile, aperto e personalizzabile. Dal punto di vista della sicurezza Drupal prevede moduli per l'autenticazione ed autorizzazione in un ottica di possibili integrazioni dell'applicazione che permettano un controllo granulare dell'accesso ai singoli datastream. Questo nel caso siano richieste policy diverse di accesso ai vari contenuti.

Un altro componente dell'architettura è Islandora. Si tratta di un framework open-source sviluppato dalla Biblioteca Robertson della UPEI (University of Prince Edward Island, Canada) e costituisce un sistema completo perfettamente integrato di congiunzione e coordinamento tra il repository FEDORA Commons e il CMS Drupal, rendendo quest'ultimo l'interfaccia tramite la quale amministrare e presentare i contenuti del repository.

La vivace comunità di Islandora ha prodotto, tra i vari moduli e componenti, un paio di elementi particolarmente interessanti per il progetto DigiBESS: il viewer ed il sistema di ricerca. Il viewer è un componente sviluppato in java che permette la lettura online dei libri. Il viewer viene richiamato direttamente dal repository

sotto forma di datastream virtuale e produce la visualizzazione, eventualmente integrata all'interno di una pagina del server web, delle singole pagine di un libro abbinate al relativo testo prodotto dall'OCR, potendo eventualmente effettuare lo zoom delle immagini e navigare tra le pagine del volume.

Per la piattaforma di indicizzazione e ricerca, che costituisce per questo tipo di progetti sicuramente uno dei componenti più apprezzati ed utili per l'utente finale la scelta è ricaduta su Solr, parte del progetto Apache Lucene.

Con una personalizzazione della configurazione di Solr si è ottenuta l'indicizzazione full-text dei datastream relativi ai testi dei volumi e quella a parole dei metadati Dublin Core. Il risultato permette la ricerca per parole chiave nei dati descrittivi delle opere dal sito web con possibilità di filtri suggeriti dallo stesso sistema di indicizzazione (*facet*) ed eventualmente abbinata o in alternativa la ricerca full-text nel contenuto dei libri. Per quest'ultima sono stati adattati i moduli Islandora modificati per produrre a video il risultato della ricerca. Le parti di testo in cui sono state ritrovate le parole vengono evidenziate ed è attivo il collegamento alla pagina specifica del libro.

Il lavoro, molto appassionante e costruttivo, ha permesso di approfondire e conoscere tecnologie nuove, con la sensazione di aver sviluppato qualcosa che integra informatica e biblioteconomia.

Uno degli obiettivi che ci si era prefissato era quello di produrre qualcosa che, coerentemente con la filosofia dalla comunità open-source, fosse aperto come soluzione, di facile replica in altri contesti e riutilizzabile eventualmente da altri progetti. Per tale motivo tutta la documentazione prodotta è disponibile in rete sul sito di sviluppo.[11]

NOTE

[1] <www.bess-piemonte.it>.

[2] <http://elib.bess-piemonte.it/bess/index.jsp>.

[3] *Rapport fait au nom de la Commission des Affaires Culturelles et de l'Education sur la proposition de loi, adoptée par le Sénat, relative à l'exploitation numérique des livres indisponibles du XXe siècle*, par M. Hervé Gaymard, député, <http://www.assemblee-nationale.fr/13/pdf/rapports/r4189.pdf>.

[4] Sulla cooperazione tra biblioteche corre l'obbligo di citare il bel libro di Anna Galluzzi, *Biblioteche e cooperazione. Modelli, strumenti, esperienze in Italia*, Milano, Editrice Bibliografica, 2004. Per chi invece fosse maggiormente interessato agli aspetti più problematici e soprattutto meno scontati della cooperazione si può riflettere su come "though we may cooperate because our own resources are not self-sustaining, in many social relations we do not know exactly what we need from others – or what they ought to want from us... Cooperation [is] a craft. It requires of people the skill of understanding and responding to one another in order to act together, but this is a thorny process, full of difficulty and ambiguity and often leading to destructive consequences..." (Richard Sennett, *Together. The Rituals, Pleasures and Politics of Cooperation*, London, Allen Lane / Penguin Books, 2012).

[5] Pierre Musso, *L'ideologia delle reti*, Milano, Apogeo, 2007.

[6] Klaus Kempf, *Il Münchener Digitalisierungszentrum e lo stato dell'arte degli scanner automatici*, "Biblioteche oggi", ottobre 2008, vol. 26, n. 8, p. 39-45. Per maggiori informazioni si può visitare il sito: <http://www.bayerische-landesbibliothek-online.de/>.

[7] Il finanziamento viene riconosciuto all'Istituto Ricerche Economico Sociali del Piemonte (IRES) che svolge fiduciariamente per conto di BESS tutta l'attività amministrativa ed economica.

[8] In proposito si veda: Markus Brantl – Tommaso Garosci, *OCR: i progetti di digitalizzazione e il riconoscimento ottico dei caratteri*, "Bollettino AIB", vol. 48, n. 4 Dicembre 2008.

[9] I collaboratori del Ceris-CNR che hanno lavorato alla realizzazione del repository sono Giancarlo Birello, Ivano Fucile, Valter Giovanetti e Anna Perin.

[10] <www.digibess.it>.

[11] <dev.digibess.it>.

ABSTRACT

Digital projects represent a tough challenge for today's cash-strapped libraries in Italy. But they also represent an opportunity to review policies and use resources more efficiently. The main vehicle is setting up cooperative programmes. In turn this should leverage on territorial or disciplinary propinquities. The article reports on one such cooperative project currently under way in Piemonte (Italy). After introducing the experiment by way of a short discussion of pros-and-cons of cooperation/networks, the focus turns on the technical platform. The description summarizes the various components and their function within the repository architecture. The project is fully OAI-PMH compliant and all the software layers are open-source allowing for smooth data harvesting.