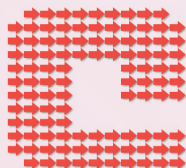

IL SERVIZIO DI EMEROTECA VIRTUALE AL CASPUR ED IL SUO NUOVO MOTORE DI RICERCA



di Gino Farinelli, Riccardo Fazio, Ilaria De Marinis e Stefano De Luca

L'Emeroteca Virtuale del CASPUR (EV) consente l'accesso a più di cinquemila periodici accademico-scientifici a testo completo per un totale di otto milioni di articoli. Il software sin qui utilizzato, Science Server, appare datato e sta causando un aumento eccessivo nei tempi di risposta nelle operazioni di ricerca di un singolo articolo. Tutto ciò ha spinto verso lo sviluppo di un nuovo motore di ricerca più in linea con i tempi di risposta attesi dai destinatari del servizio, anche in relazione ad una mole di articoli ricercabili che si sta avviando a superare il traguardo dei dieci milioni.

Dott. Gino Farinelli
CASPUR
Gruppo Automazione
Biblioteche
gino.farinelli@caspur.it

Dott. Riccardo Fazio
CASPUR
Gruppo Automazione
Biblioteche
fabio.fazio@caspur.it

Dott.ssa Ilaria De Marinis
CASPUR
Gruppo Automazione
Biblioteche
ilaria.demarinis@caspur.it

Stefano De Luca
CASPUR
Gruppo Automazione
Biblioteche
stefano.deluca@caspur.it

• Abstract

CASPUR has been offering since 1999 a web based *Digital Library* service called "Emeroteca Virtuale" (EV), which guarantees full-text permanent access to scholarly scientific publications of five commercial publishers, the biggest ones in a worldwide context. Nowadays EV is a service accessed by more than 350.000 users, mostly researchers belonging to nearly 30 middle and south Italy universities and research bodies. Users can browse and search on this platform up to 8 million full text articles from 5.000 e-journals, mainly in Scientific-Technology-Medicine (STM) area.

Since 1999 EV service is based on a commercial software called *Science Direct*, whose main functionalities are articles browsing and articles searching. This software is going to show its limits with such huge amount of searchable articles, especially when search times are concerned. This is why starting from the second half of 2008 CASPUR staff which is managing this digital library has planned a migration from Science Direct search engine, to a new one, chosen in the context of open source software: Lucene from Apache Software Foundation. Nowadays Lucene is used in many web based applications around the world, especially in the field of scientific digital libraries. Concerning EV service, Lucene integration with Science Server environments is going to end in the next few months. Exhaustive tests have been performed using the whole contents, giving really promising results, with a mean search time 10 to 100 shorter than Science Direct search response. This definitively states the choice of Lucene as a good search engine and represent a valid starting point for "Emeroteca Virtuale" whole service renovation.

Il servizio di Emeroteca Virtuale (<http://periodici.caspur.it>), gestito dal CASPUR sin dal 1999, consente l'accesso permanente a più di cinquemila periodici accademico-scientifici di cinque tra i più grandi esponenti mondiali di editoria scientifica o di società professionali. Attraverso un'interfaccia web gli utenti di circa 30

università italiane del centro-sud, che hanno sottoscritto contratti di accesso alle riviste con gli specifici editori, possono arrivare a visualizzare il testo completo del singolo articolo scientifico (si parla in tali casi di “accesso al *full-text*”).

A fine 2008 l’Emeroteca Virtuale ha raggiunto il traguardo degli otto milioni di articoli e Science Server, il software gestionale su cui si basa il servizio, ha cominciato a mostrare limiti sempre più evidenti, a partire dagli elevati tempi di risposta nelle funzioni di ricerca di un articolo.

Esistono due modalità operative con le quali gli utenti dell’Emeroteca Virtuale hanno la possibilità di reperire gli articoli: attraverso elenchi da cui selezionare la rivista desiderata e successivamente “sfogliarla” come fosse una rivista cartacea; oppure tramite articolate maschere di ricerca all’interno delle quali è possibile indicare titolo della rivista, titolo dell’articolo, autore (autori), anno di pubblicazione, e molte altre chiavi di ricerca.

Elevati tempi di risposta alle ricerche effettuate possono costituire un problema rilevante per gli utenti dell’Emeroteca. Infatti, come diversi studi di *web usability* dimostrano, gli utenti che effettuano ricerche su internet si aspettano di ottenere risposte dai server web in tempi brevi, al massimo dell’ordine di una manciata di secondi; in caso contrario gli utenti possono ritenere che il servizio non funzioni o che la ricerca non sia andata a buon fine e, in questo caso, essere indotti ad avviarla nuovamente (fenomeno dei doppi click), appesantendo così ulteriormente il carico di lavoro del server.

Al fine di mantenere la qualità del servizio offerto dall’Emeroteca su livelli accettabili, in ragione soprattutto della sua diffusione in ambito universitario, si è avviata, nel 2008, una fase di studio e valutazione delle soluzioni open source¹ di *Information Retrieval (IR)*² disponibili in rete che potessero meglio rispondere ai criteri di “scalabilità”, ovvero di poter garantire la loro efficienza e funzionalità al crescere della base dati degli articoli ricercabili, e di “robustezza”, ovvero di poter essere applicate anche in condizioni di intenso utilizzo.

L’attività di ricerca ha portato alla scelta di Lucene³, un potente software open-source di *Information Retrieval*, interamente sviluppato nel linguaggio di programmazione Java dalla Apache Software Foundation (ASF)⁴.

Lucene è tanto potente e versatile quanto complesso da configurare; per tale motivo la stessa comunità dell’ASF ha sviluppato un *enterprise search server*, denominato SOLR, anch’esso open source, che fornisce allo sviluppatore e all’amministratore del sistema un’agile interfaccia web per la gestione e la configurazione di Lucene. Per il suo funzionamento, SOLR necessita di un *Java servlet container*⁵ e, allo scopo, si è deciso di utilizzare il server Tomcat⁶, distribuito anch’esso dalla ASF.

¹ *Open-source* è un termine utilizzato per indicare una tipologia di software non soggetto a *policy* d’uso di tipo commerciale e, soprattutto, di cui è distribuito in chiaro il codice sorgente.

² Con il termine *Information Retrieval (IR)* si intende l’insieme delle tecniche utilizzate per il recupero mirato di informazioni in formato elettronico; nell’Emeroteca Virtuale le “informazioni” sono rappresentate dagli articoli scientifici in full-text.

³ Rif. <http://lucene.apache.org/>.

⁴ L’ASF (<http://www.apache.org>) è una fondazione non-profit formata da una comunità distribuita di sviluppatori che lavorano su progetti di software open source per applicazioni WEB.

⁵ Lo *Java Servlet Container* è un server java utilizzato per lo sviluppo di siti web con contenuto dinamico.

⁶ Rif. <http://tomcat.apache.org/>.



Uno dei motivi che ci hanno spinto a privilegiare Lucene⁷ come nuovo motore di ricerca dell'Emeroteca è legata proprio al sistema SOLR: quest'ultimo riunisce in due soli file, in formato XML, le configurazioni utilizzate da Lucene e fornisce una versatile interfaccia web per il *debugging* analitico del motore di ricerca. L'indicizzazione dei documenti all'interno del sistema di IR avviene su file XML nell'ambito di sessioni HTTP e nello stesso formato sono i risultati delle ricerche effettuate.

Alla scelta di Lucene ha certamente contribuito anche il fatto che nello specifico contesto degli organismi che offrono un servizio simile a quello offerto dall'EV, in ambito internazionale Lucene venga utilizzato dal consorzio delle Università dell'Ohio (OHIO-Link) e dalla *Research Library* dei laboratori nazionali di Los Alamos, con cataloghi contenenti più di dieci milioni di "oggetti digitali", rappresentati da articoli o libri elettronici. In Italia Lucene è utilizzato dalla Biblioteca Nazionale Centrale di Firenze.

Uno degli aspetti nei quali si è intervenuti per la personalizzazione di Lucene all'ambiente di EV è stato quello relativo alle "procedure di *stemming*": queste procedure permettono di indicizzare un termine all'interno del DB utilizzato per la ricerca, memorizzando non solo la parola completa che si vuole rendere ricercabile, ma anche la sua radice. In questo modo la ricerca del termine "cell" può restituire non solo l'articolo che contiene questa parola, ma anche quelli che contengono "cells", "cellular" etc. Dal momento che nessuno degli "algoritmi di *stemming*" adottati da Lucene si comporta come quelli utilizzati attualmente dal software dell'EV (*Science Direct*), si è deciso di riscrivere un algoritmo di *stemming ad hoc*, sfruttando il fatto che il software Lucene è di tipo open-source.

Un'altra area nella quale sono state necessarie modifiche è quella relativa alle maschere di ricerca: normalmente Lucene, e gli *Information Retrieval* in genere, vengono utilizzati per realizzare motori di ricerca "google-like"; la maschera di ricerca è composta da una sola casella di testo e i termini da ricercare vengono posti in OR fra loro (almeno uno); nei risultati vengono mostrati prima i documenti che contengono tutti i termini ricercati e, a seguire, i documenti in cui sono presenti solo alcuni di loro. La ricerca degli articoli in EV è caratterizzata invece da una diversa impostazione: infatti l'utente deve avere la possibilità di effettuare ricerche su più campi contemporaneamente in AND fra loro (tutti contemporaneamente) e, se si inseriscono più termini in un dato campo di ricerca, anch'essi devono essere posti in AND fra loro (AND "implicito"). Anche in questo caso sono state apportate personalizzazioni dell'ambiente offerto da Lucene, in modo da garantirne una transizione "senza discontinuità" verso quello dell'Emeroteca.

Buona parte dell'attività di "integrazione" di Lucene nell'EV ha riguardato l'identificazione dei dati soggetti alle operazioni di indicizzazione: nell'Emeroteca Virtuale ad ogni articolo è associato un file PDF, che contiene il *full-text* dell'articolo (ovvero l'articolo completo), e un file XML con la descrizione di tutte le informazioni legate all'articolo stesso (i cosiddetti "metadati").

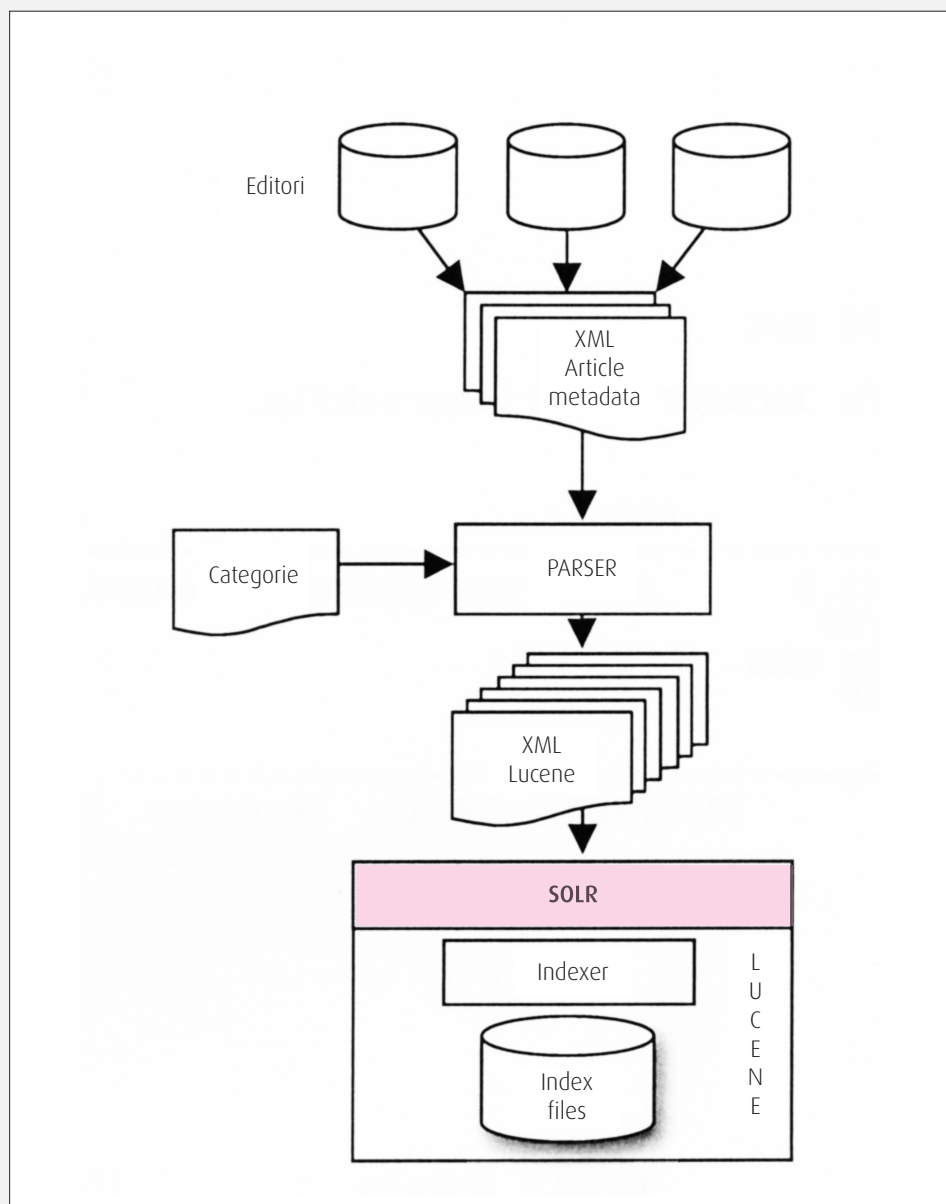
⁷ Il codice ISSN è una stringa alfanumerica di 8 caratteri, identificativa, nel panorama delle pubblicazioni cartacee o elettroniche, dello specifico prodotto editoriale (periodico o monografia).

Struttura e contenuto dei metadati nei file XML sono stati attentamente analizzati in modo da definire quali informazioni degli articoli dovessero essere presenti nelle maschere di ricerca e, quindi, ricercabili dall'utente e quali dovessero invece essere mostrati nella scheda con i risultati.

Lucene prevede che ad ogni informazione da trattare sia associato un tipo di dati (testuale, numerico, etc.) il cui comportamento può essere definito dall'amministratore (ad esempio, se e quale algoritmo di *stemming* utilizzare) e, inoltre, se l'informazione sia da indicizzare, o se sia da salvare.

Ad esempio, il titolo dell'articolo è un tipo di dato che deve essere soggetto allo *stemming* e che deve essere indicizzato e salvato. Questo perché tale informazione deve essere ricercabile anche nelle sue varianti e deve essere mostrata nei risultati della ricerca. Viceversa l'*abstract* di un articolo viene indicizzato in quanto ricercabile, ma non viene salvato, poiché non compare nei risultati; o ancora, il *coverdate* (la data di pubblicazione dell'articolo) è un'informazione che non va indicizzata in quanto non ricercabile, ma deve essere salvata perché viene visualizzata fra i risultati.

Figura 1
 Schema a blocchi dell'integrazione di Lucene con l'Emeroteca Virtuale.





Per quanto riguarda l'EV le informazioni indicizzate in Lucene per ciascun articolo comprendono: l'editore, il titolo della rivista, il codice ISSN⁸, il titolo dell'articolo, l'autore(i), l'*abstract*, le *keyword* dell'articolo, l'anno di pubblicazione ed il DOI⁹.

L'attività di personalizzazione di Lucene nell'ambiente dell'EV si è conclusa con lo sviluppo di un programma (*parser*) che effettua l'analisi dei file XML con i metadati degli articoli ed estrae soltanto le informazioni necessarie all'indicizzazione in un formato adatto. Da ultimo sono state sviluppate tutte le procedure per l'interrogazione attraverso sessioni HTTP del DB di Lucene e per la visualizzazione dei documenti trovati (si veda lo schema a blocchi di Figura 1).

Nell'ultimo trimestre del 2008 sono stati effettuati dei test di indicizzazione degli articoli nel DB di Lucene, sempre più massivi, fino ad arrivare all'indicizzazione di tutti gli articoli presenti in Emeroteca Virtuale (otto milioni circa).

Parallelamente sono stati effettuati sul DB diversi test di ricerca, differenziati per tipologia (ricerca semplice, avanzata, esperta) per valutare le prestazioni e i tempi di risposta di Lucene all'aumentare del numero degli articoli indicizzati.

In tutti i casi i risultati sono stati estremamente incoraggianti, dal momento che i tempi di risposta si sono attestati ben al di sotto del secondo, anche nel caso di *query* relativamente complesse dal punto di vista del numero degli articoli trovati (ad esempio la ricerca effettuata con la chiave "cell" ne fornisce più di 250.000).

Come nota conclusiva si vuole osservare come l'attività di integrazione iniziata con il software Lucene (che andrà in linea entro la prima metà del 2009) costituisca un primo passo verso una riprogettazione completa del servizio di Emeroteca Virtuale sia in termini di contenuti sia di *layout*. Il tutto al fine di fornire agli utenti un servizio più fruibile, efficiente ed in linea con le nuove tendenze (di format e contenuti) che stanno emergendo nel campo degli strumenti di presentazione *on line* delle pubblicazioni scientifiche.

⁸ Il codice ISSN è una stringa alfanumerica di 8 caratteri, identificativa, nel panorama delle pubblicazioni cartacee o elettroniche, dello specifico prodotto editoriale (periodico o monografia).

⁹ DOI: *Digital Object Identifier*; è un identificatore univoco degli oggetti digitali (articolo, presentazione, immagine, video etc.) che ne permette non solo l'identificazione, ma il recupero indipendentemente dalla sua specifica collocazione all'interno della rete.