JLIS.it

Recupero della classificazione decimale Dewey da altre basi di dati: un progetto di bonifica del catalogo

Stefano Bargioni, Michele Caputo,
Alberto Gambardella, Luigi Gentile

1 Introduzione

La Biblioteca della Pontificia Università della Santa Croce[1] è una biblioteca di ricerca appartenente alla Rete URBE – Unione Romana Biblioteche Ecclesiastiche.[2] Attualmente essa possiede circa 167.000 volumi corrispondenti a 145.000 record bibliografici catalogati in formato MARC21. Per la gestione della biblioteca si sono succeduti tre Integrated Library System (ILS): Aleph 300, Amicus 3.5.4 e l'attuale Koha[3] 3.2.7. Contemporaneamente all'adozione dell'ILS open source Koha dall'elevata produttività, sono stati introdotti gli *authority records*. La duttilità di Koha ha permesso inoltre di aprire nuovi percorsi di sperimentazione operativa ordinariamente non realizzabili con un ILS commerciale.

Al fine di fornire all'utenza maggiori strumenti di ricerca catalografica in chiave semantica e tenendo presente che l'attività di

[1] http://www.pusc.it/bib.
[2] http://www.urbe.it.
[3] http://koha-community.org.

Traduzione italiana a cura dell'Autore.

soggettazione basata sul Nuovo Soggettario Thesaurus della Biblioteca nazionale centrale di Firenze è recente, si è deciso di sviluppare le potenzialità legate alla classificazione Dewey,[4] già parzialmente adottata in biblioteca da una decina di anni ed assegnata a circa il 25% dei documenti posseduti.

Si è sviluppata così l'ipotesi di incrementare, attraverso l'importazione da altre basi di dati,[5] la presenza della classificazione Dewey nei record bibliografici, utilizzando il codice ISBN[6] come chiave per il recupero delle classificazioni mancanti. Si è proceduto inizialmente all'individuazione di fonti (basi di dati) che soddisfacessero significativamente le nostre esigenze, sia dal punto di vista qualitativo che quantitativo. L'esperienza della catalogazione derivata – un punto di forza di Koha – è stata fondamentale al riguardo. Una volta scelte le potenziali fonti, sia nazionali che internazionali, sono stati individuati i metodi per potervi accedere programmaticamente. La difformità con cui le varie istituzioni pubblicano i loro dati ha comportato la necessità di diversificare i metodi di interrogazione per poter accedere sistematicamente all'informazione. Si va dal caso più moderno della OCLC, che ha dato vita a Classify,[7] un web service sperimentale specifico per la classificazione, ai casi meno semplici in cui si deve ricorrere alle pagine HTML. Per poter

[4] http://dewey.info.

[5] L'importazione di dati da altre fonti bibliografiche si giustifica attraverso il "principio di condivisione" sentito e vissuto praticamente da sempre dai cataloghi pubblici. Questo principio fonda lo scambio di informazione tramite OPAC, Z39.50, interfacce web, ecc., ed ha come scopo anche il confronto e il controllo reciproco delle registrazioni e della identificazione della biblioteca fonte dell'informazione, assicurata, per esempio, in MARC21 dal campo 035. L'utilizzo delle importazione è avvenuto nel rispetto delle eventuali condizioni o raccomandazioni indicate nelle pagine web dei siti interrogati. Diverso potrebbe essere il caso di un utilizzo commerciale dell'informazione recuperata.

[6] http://www.isbn.org/standards/home/index.asp.

[7] http://classify.oclc.org.

controllare la qualità delle classificazioni Dewey ottenute, è stato creato un apposito algoritmo descritto nel paragrafo "Il controllo di qualità". Il processo di ricerca e importazione dei dati andava anche analizzato sotto il profilo del carico che rappresenta sia per il sistema alla fonte sia per quello di destinazione. Le interrogazioni dei server non possono avvenire ad un ritmo eccessivo, e per questo alcuni di essi pubblicano espressamente raccomandazioni agli eventuali software, chiamati *crawler* o *web robots*, che li interrogano.

2 Individuazione dei record da modificare

I record del catalogo da arricchire sono quelli dotati di ISBN (tag 020) ma mancanti di classificazione Dewey (tag 082). La loro individuazione può avvenire in Koha mediante una query SQL (v. listato 1), specifica del database MySQL, applicata al campo `marcxml`[8] della tabella `biblioitems`[9]

Listing 1: Query per la selezione dei record in Koha.

```
SELECT biblionumber, listaISBN
FROM biblioitems
WHERE isbn_presente
AND dewey_assente
AND lingua_008='...'
```

Non trattandosi di una ricerca tramite indici, l'individuazione avviene mediante l'analisi record per record del database. In questo caso dunque si è di fronte a un aspetto del progetto dipendente dalla

[8] Il campo `biblioitems.marcxml` contiene la rappresentazione del record bibliografico nel formato MARCXML, http://www.loc.gov/standards/marcxml/, http://en.wikipedia.org/wiki/MARC_standards#MARCXML.

[9] Gli elementi principali della query sono descritti in tabella 9 a pagina 19.

potenza di calcolo del server su cui risiede l'ILS. Altri ILS permetteranno di reperire il numero di sistema e l'ISBN di un record senza classificazione Dewey in modi molto diversi da Koha, in funzione della struttura dati utilizzata per conservare i dati bibliografici e degli strumenti a disposizione per accedervi.

3 Le fonti

Gli ISBN di ogni record, estratti dalla query, sono stati utilizzati per interrogare sette diverse basi di dati. Le fonti scelte sono elencate nella Tabella 1 nell'ordine temporale di interrogazione.

Siccome lo scopo del lavoro era essenzialmente pratico, non si è cercato di interrogare ogni fonte con lo stesso ISBN. Nel caso in cui venisse reperita e salvata nel record una classificazione Dewey, si è deciso che quella fonte avrebbe prevalso sulle successive, così che il record non sarebbe stato ulteriormente processato. Questa modalità ci è parsa più economica rispetto alle altre due possibili: interrogare tutte le fonti con lo stesso ISBN, o simultaneamente o in successione. Inoltre in diversi casi la ricerca è stata limitata alla lingua prevalente della fonte interrogata, sia per evitare un eccessivo numero di ricerche, sia perché ritenuta più attendibile. Tra le lingue

1	Classify	Classify di OCLC
2	LC	Library of Congress
3	BNF	Bibliothèque nationale de France
4	DNB	Deutsche Nationalbibliothek
5	BNCF	Biblioteca Nazionale Centrale di Firenze
6	BNCR	Biblioteca Nazionale Centrale di Roma
7	BNB	British National Bibliography

Tabella 1: Fonti di classificazione Dewey interrogate.

presenti in catalogo, lo spagnolo non è stato trattato, in mancanza di basi di dati da noi ritenute sufficientemente significative allo scopo. Il metodo adottato non consente di effettuare confronti tra le diverse fonti a parità di condizioni, ma permette pur sempre un'analisi statistica dell'uso della classificazione Dewey nelle diverse fonti, come si vedrà in seguito.

La Tabella rappresentata in figura 1 mostra l'indirizzo, il tipo di dato restituito, il tipo di servizio contattato per ogni fonte e la lingua interessata: Le fonti di tipo diverso da quelle web forniscono gli

	Fonte	Tipo	Connessione	Interrogazione	Dati	Lingua
1	Classify	REST	http://classify.oclc.org/classify2/Classify?summary=false&isbn=*ISBN*		XML	tutte
2	LC	Z39.50	lx2.loc.gov:210/LCDB	find @attr 1=7 *ISBN*	MARC	tutte
3	BNF	Z39.50	z3950.bnf.fr:2211/TOUT-UTF8 user Z3950, password Z3950_BNF	find @attr 1=7 *ISBN*	MARC	tutte
4	DNB	web	https://portal.dnb.de/opac.htm?query=isbn%3D*ISBN*&method=simpleSearch		HTML	ger
5	BNCF	web	http://opac.bncf.firenze.sbn.it/opac/controller.jsp?action=search_avanzatasearch&query_fieldname_1=keywords&fieldname=identifkw&query_querystring_1=identifkw%40%3A*ISBN*		HTML	ita
6	BNCR	web	http://193.206.215.17/BVE/result.php?textexpert=is%3DISBN&comebackb=1&dove=completa&numschede=10&ordschede=4+ia&lastRefinedQueryRPN=is%3DISBN&cercato=is%3DISBN&numresults=1&startp=esperta&query4usr=is%3D*ISBN*&formatoAna=3&vaiformat=Esegui		HTML	ita
7	BNB	Z39.50	z3950cat.bl.uk:9909/BNB03U richiede credenziali, concesse su richiesta	find @attr 1=7 *ISBN*	MARC	eng

Figura 1: Caratteristiche delle fonti di classificazione Dewey interrogate.

estremi della connessione nelle rispettive pagine di spiegazione del servizio. Per le fonti di tipo web, invece, connessione e interrogazione vanno quasi sempre dedotte empiricamente, in genere a partire dalla schermata di interrogazione avanzata del catalogo. Per poter individuare i parametri da inviare, compreso quello dell'ISBN, si può procedere in uno dei modi elencati in Appendice.

Sempre nel caso di pagine web, la tecnica adottata per l'estrazione del dato è particolarmente specifica. Si deve applicare quello che

comunemente viene denominato *web scraping*,[10] *screen scraping* o in generale *data scraping*.
Occorre in sostanza capire se si dispone di un metodo per individuare ed estrarre il dato di interesse dall'interno del codice HTML ottenuto, operazione che gli altri tipi di risposte rendono più facile e standard visto che forniscono dati strutturati. Il Web 2.0 e ancor più l'incalzante web dei linked data fanno auspicare che le fonti di dati offrano non solo interfacce web, essenzialmente destinate alla fruizione dell'uomo, ma soprattutto interfacce con risposte standard strutturate, fruibili da altre macchine e stabili nel tempo.

La logica utilizzata nei programmi di interrogazione delle fonti dati è riconducibile all'algoritmo rappresentato in figura 2.

```
apertura della connessione al proprio database bibliografico
estrazione degli ISBN dei record senza Dewey
apertura della connessione alla fonte dati, se di tipo Z39.50
per ogni ISBN
      interrogazione della fonte dati in base all'ISBN corrente
      se nella risposta è disponibile una classificazione Dewey
            se la classificazione Dewey supera il "controllo di qualità"
                  aggiornamento del record bibliografico
      attesa per evitare sovraccarichi
chiusura della connessione alla fonte dati, se di tipo Z39.50
chiusura della connessione al proprio database bibliografico
```

Figura 2: Rappresentazione della logica utilizzata nei programmi di interrogazione delle fonti dati.

Fa eccezione il caso di Classify, come detto, per il quale il passo di "interrogazione della fonte dati per l'ISBN corrente" deve essere seguito da istruzioni specifiche (figura 3.).

```
se l'ISBN si riferisce a più opere
      ripetere l'interrogazione relativamente alla prima opera
```

Figura 3: Rappresentazione dell'eccezione alla logica utilizzata nei programmi di interrogazione delle fonti dati da Classify.

[10] http://en.wikipedia.org/wiki/Web_scraping.

Il paragrafo 3 dell'Appendice riporta esempi per ognuno dei tre tipi di dati ottenuti come risposta: XML, MARC e HTML. La risposta di Classify[11] è tipicamente di quattro tipi, come da tabella 2.

Response code	Significato
2	ISBN corrispondente a una singola opera
4	ISBN corrispondente a più opere
101	ISBN errato
102	ISBN non trovato

Tabella 2: Tipi di risposte di Classify.

Nel caso di risposta di "ISBN corrispondente a più opere", Classify[12] fornisce un elenco di identificatori OCLC# delle relative opere. È stata preferita la prima di queste, andando a reperire il record descrittivo tramite il suo OCLC# con un'altra interrogazione del tipo: http://classify.oclc.org/classify2/Classify?summary=false&swid=OCLC#, che ovviamente ha response code 2, singola opera. La risposta di Classify per singola opera (se ne veda un esempio al paragrafo 1 dell'Appendice) riporta sia le aggregazioni delle classificazioni Dewey e LCC assegnate all'opera dai numerosi cataloghi che contribuiscono a OCLC, sia un elenco di edizioni, corredate dalla classificazione. È parso preferibile importare la classificazione

[11] Le API di Classify sono descritte in http://classify.oclc.org/classify2/api_docs/index.html e possono essere provate tramite il Classify API Explorer alla pagina http://classify.oclc.org/classify2/api_docs/classify.html.

[12] Le aggregazioni in Classify avvengono per applicazione di FRBR. Alla pagina http://www.oclc.org/research/activities/classify.html (al 21.1.2013) si afferma: "Bibliographic records are grouped using the OCLC FRBR Work-Set algorithm <http://www.oclc.org/research/activities/frbralgorithm.html> to form a work-level summary of the class numbers and subject headings assigned to a work. You can retrieve a summary by ISBN, ISSN, UPC, OCLC number, author/title, or subject heading".

della prima edizione in elenco, perché rispetto alle altre era spesso più completa. Le fonti Z39.50 richiedono sostanzialmente di estrarre il valore del tag della classificazione Dewey, secondo le regole del relativo formato MARC, come da Tabella 4.

Formato MARC	tag	sottocampo del codice	sottocampoo dell'edizione
MARC21	082	a	2
InterMARC o UNIMARC	676	a	v

Tabella 3: Tag della classificazione Dewey in alcuni dialetti MARC.

4 Il "controllo di qualità"

Prima del progetto, il catalogo era popolato da classificazioni Dewey riferentesi alle edizioni dalla 19 alla 23. La scelta di non introdurre né classificazioni di tipo ridotto né classificazioni di edizioni Dewey inferiori alla 19 ha implicato di dover rinunciare a numerose classificazioni trovate, come riportato nelle statistiche della tabella 7 a pagina 14. È parso opportuno privilegiare la qualità alla quantità per ottenere un arricchimento più possibile allineato alla politica di catalogazione. In concreto, oltre a limitare l'edizione alla 19 o superiori, sono state scartate classificazioni con indicatori 1 e 2 diversi dal "0 0" e "0 4".[13] Sono state eliminate anche le classificazioni contenenti caratteri non numerici o mancanti di edizione. Infine le classificazioni sono state normalizzate prima di essere inserite nel record.

[13]Secondo il MARC21, il primo indicatore del campo 082 con valore "0" indica uso dell'edizione completa della Dewey, il secondo indicatore con valore "0" indica Dewey assegnata dalla Library of Congress mentre il valore "4" corrisponde a notazione assegnata da una agenzia diversa dalla Library of Congress.

5 Il tag 035

Contestualmente alla modifica del record, è parso opportuno tenere traccia degli estremi del record da cui è stata tratta la classificazione Dewey importata, tramite l'utilizzo del tag 035 del MARC21, come nel seguente esempio:

Listing 2: Esempio di utilizzo del tag 035 di MARC21.

```
00872nam a2200265 i 4500
001 000000035650
003 IT-RoPUS
005 20121121122621.0
008 041027r19851982xxk u000 u eng c
020 $a 0198247761
035 $a (OCoLC)007946090
040 $a IT-RoPUS $b ita
082 04 $a 111.85 $2 19
100 1 $a Savile, Anthony. $9 70779
245 14 $a The test of time : $b an essay in philosophical
    aesthetics / $c Anthony Savile.
...
```

Nel caso di fonte non MARC21 o comunque senza MARC Organization Code,[14] è stato scelto di assegnare un codice più logico possibile, come da Tabella 4 nella pagina seguente.

L'ID è stato estratto dal record in posizioni diverse caso per caso. Per le fonti Z39.50 si trova nel tag 001, mentre per la Library of Congress si ricorre al tag 010. Anche Classify lo riporta espressamente nel record XML, mentre il reperimento dai record in formato HTML è particolarmente complesso.

[14] http://www.loc.gov/marc/organizations/.

Tabella 4: Codici istituzione per lo 035.

1	Classify di OCLC	OCoLC	ufficiale
2	Library of Congress	DLC	ufficiale
3	Bibliothèque nationale de France	FR-PaBFM	ufficiale
4	Deutsche Nationalbibliothek	DE-101	ufficiale[a]
5	Biblioteca Nazionale Centrale di Firenze	BNCF	non ufficiale
6	Biblioteca Nazionale Centrale di Roma	BNCR	non ufficiale
7	British National Bibliography	BNB	non ufficiale

[a] http://dispatch.opac.d-nb.de/DB=1.2/LNG=EN.

Questa scelta consente di collegare il record bibliografico a quello di un catalogo esterno, utile per costruire un link di interesse sia a livello di OPAC (figura 4 a pagina 12) che di linked data.

Il link nell'OPAC viene costruito, per ogni occorrenza del tag 035, sulla base dei link della tabella 5 a fronte. La permanenza di alcuni è certa (permalink). Negli altri casi, il link, di natura molto più instabile, può essere costruito ricorrendo alla vista di ogni singolo record offerta dal catalogo.

6 Attese durante la ricerca sulle fonti

Come accennato nell'Introduzione, un uso continuo, facilmente ottenibile con interrogazioni automatizzate, può gravare sul server interrogato. La lettura di pagine web di tipo "Terms and Conditions" permette di regolare le condizioni di utilizzo delle fonti. Ad esempio, la Library of Congress richiede esplicitamente[15] che i crawler utilizzino il server Z39.50 con un ritmo inferiore alle 10 interrogazioni al minuto. Il server Z39.50 della Bibliothèque nationale de France chiude il collegamento dopo la decima interrogazione. Il programma

[15] http://lccn.loc.gov/lccnperm-faq.html#n12.

Tabella 5: Costruzione di link nell'OPAC a partire da un'occorrenza di tag 035.

Classify di OCLC - WorldCat	`http://www.worldcat.org/search?q=no%3A`*ID*	permalink[a]
Library of Congress	`http://lccn.loc.gov/`*ID*	permalink[b]
Bibliothèque nationale de France	`http://catalogue.bnf.fr/servlet/biblio ?idNoeud=1&SN1=0&SN2=0&host=catalogue& ID=`*ID*	
Deutsche Nationalbibliothek	`http://d-nb.info/`*ID*	permalink[c]
Biblioteca Nazionale Centrale di Firenze	`http://opac.bncf.firenze.sbn.it/opac/ controller.jsp? action=notizia_view&notizia_idn=`*ID*	
Biblioteca Nazionale Centrale di Roma	`http://193.206.215.17/BVE/ricercaEsperta. php?dove=esperta &cerca=Avvia+la+ricerca& textexpert=di%3D`*ID*	
British National Bibliography	`http://search.bl.uk/primo_library/libweb /action/search.do?vid=BLBNB&fn =search&vl%28freeText0%29=`*ID*	

[a] http://www.oclc.org/worldcatorg/linking/how.htm#oclc-number.
[b] http://lccn.loc.gov/lccnperm-faq.html.
[c] Dedotto dalla visualizzazione di un singolo record al termine di una ricerca qualunque.

Figura 4: Vista di record nell'OPAC, arricchito con Dewey e link prelevati da DNB.

deve pertanto riaprirlo con la stessa frequenza. Il sito della Biblioteca nazionale centrale di Firenze non si presta ad essere consultato senza pause, dato che sembra sovraccaricarsi quasi subito.

È anche opportuno verificare, per le fonti interrogate tramite protocollo `http`, se vi sono indicazioni ai crawler nel file `/robots.txt`, dove a volte si trovano restrizioni anche per la frequenza di accesso.[16]

Pertanto per tutte le fonti sono state definite attese dai 4 ai 6 secondi tra le interrogazioni. Le pause hanno permesso anche di non sovraccaricare il nostro catalogo. Infatti ad ogni modifica di record, il motore di indicizzazione Zebra[17] usato da Koha e il motore di

[16] http://en.wikipedia.org/wiki/Robots_exclusion_standard#Crawl-delay_directive.

[17] http://www.indexdata.dk/zebra.

1	*numero di sistema*	*ISBN*		ISBN non trovato
2	*numero di sistema*	*ISBN*		ISBN errato
3	*numero di sistema*	*ISBN*		ISBN relativo a più opere
4	*numero di sistema*	*ISBN*		Dewey non trovata
5	*numero di sistema*	*ISBN*	Classificazione ed edizione trovate	Non soddisfacenti
6	*numero di sistema*	*ISBN*	Classificazione ed edizione trovate	Record modificato

Tabella 6: Tipi di record di log. Il tipo 2 e 3 sono relativi solo a Classify.

ricerca per liste sviluppato in proprio,[18] intervengono per aggiornare i propri indici e possono rallentare la consultazione dell'OPAC e il lavoro ordinario. Un aspetto da valutare in funzione della potenza di calcolo a disposizione. Il ritmo imposto dalle pause suddette di fatto prolunga il processo di importazione per ore se non per giorni, in funzione del numero di ISBN da elaborare. Questo può comportare degli adattamenti del programma, per esempio parametrizzandolo affinché lavori solo in certe fasce orarie.

7 Log

Il processo di importazione è stato monitorato al fine di raccogliere statistiche sul lavoro svolto. Sono stati registrati i tipi di record di log descritti nella tabella 6.

[18] Koha non dispone al momento di ricerche a scorrimento di indici, note anche come ricerche *browse*. È stato possibile aggiungere questa funzionalità alla nostra installazione di Koha tramite un applicativo basato su Solr (http://lucene.apache.org/solr) e sviluppato dalla nostra biblioteca. Questo *browse* è stato presentato all'incontro internazionale di utenti Koha tenutosi ad Edimburgo a giugno 2012 (http://wiki.koha-community.org/wiki/KohaCon12_Schedule#Adding_browse_to_Koha_using_Solr_.2815-20_min.29) e verrà integrato in successive versioni di Koha, in particolare quando Solr sarà in alternativa a Zebra o lo sostituirà.

8 Statistiche

I log generati permettono di costruire le seguenti tabelle e confrontare le diverse fonti sotto alcuni aspetti.

Fonte	Lingua	Record esaminati	Record modificati	ISBN non trovati	Dewey non trovate	Dewey scartate	Più opere per stesso ISBN	ISBN errato
Classify	tutte	42387	10267	5321	6607	20059	8240	133
LC	tutte	31999	1252	21195	8562	1011		
BNF	tutte	30903	2253	21327	7268	55		
DNB	ger	4193	163	3867	163	0		
BNCF	ita	12017	4088	3643	3542	744		
BNCR	ita	7549	1515	3003	2978	53		
BNB	eng	6215	193	5449	55	518		
Totale			19710					

Tabella 7: Conteggi.

Fonte	Campioni	Ed. 19 (%)	Ed. 20 (%)	Ed. 21 (%)	Ed. 22 (%)	Ed. 23 (%)
Classify	10267	19,86	23,03	36,18	20,13	0,79
LC	1231	28,11	25,83	24,29	19,58	2,19
BNF	2253	0,00	0,09	0,36	99,56	0,00
DNB	163	0,00	0,00	0,00	100,00	0,00
BNCF	4088	9,10	23,46	55,04	12,40	0,00
BNCR	1515	2,38	9,70	87,92	0,00	0,00
BNB	193	16,58	19,69	26,42	28,50	8,81
Totale	19710					

Tabella 8: Distribuzione delle edizioni, relativa alle classificazioni reperite.

La tabella 8 è riprodotta nei grafici raccolti nella figura 5 nella pagina successiva, uno per fonte.

Si notano alcune scelte precise, quali BNF, DNB e BNCR, di privilegiare una sola edizione. D'altra parte, visto quanto è riportato per Classify, mediamente chi ha intrapreso l'uso della classificazione Dewey da tempo, non sembra aver provveduto ad un aggiornamento delle notazioni Dewey nel catalogo, certamente per la complessità

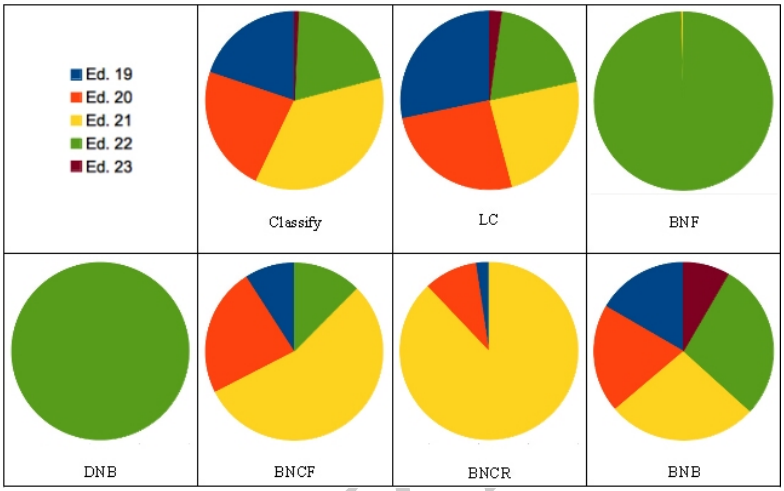

Figura 5: Distribuzione delle edizioni.

dell'operazione. Infine si nota la (ancora) scarsa diffusione dell'edizione 23. Come indicato in precedenza, il catalogo si è arricchito di 19710 nuove classificazioni Dewey in altrettanti record bibliografici. L'aumento è stato del 47,8%, dato che in precedenza i record con tag 082 erano 41255. La distribuzione attuale delle classificazioni Dewey, mostrata nella figura 6 nella pagina seguente, traccia un profilo del posseduto che riflette le aree di interesse delle facoltà e di crescita della biblioteca. La distribuzione delle edizioni Dewey in catalogo è rappresentata dalla figura 7 nella pagina successiva. L'assenza di edizione per un numero significativo di record bibliografici è un caso di disomogeneità catalografica per la cui bonifica si potrebbe utilizzare un metodo molto simile a quello illustrato nel presente lavoro.

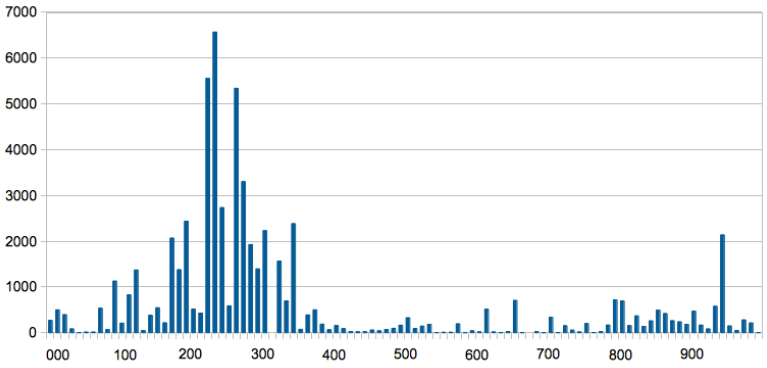

Figura 6: Distribuzione del posseduto secondo le divisioni della classificazione Dewey.

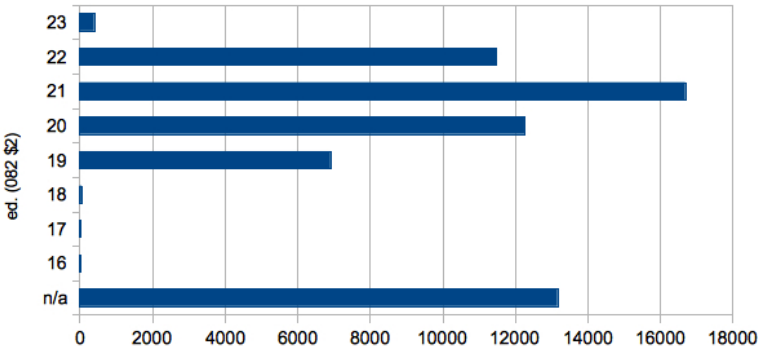

Figura 7: Distribuzione delle edizioni della classificazione Dewey.

9 L'indice Dewey nell'OPAC

Tramite gli indici a scorrimento, mostrati con l'esempio della figura 8 e citati in precedenza, è possibile offrire nell'OPAC un percorso di ricerca semantico basato sulla classificazione Dewey. I conteggi delle ricerche effettuate dall'utenza mostrano che l'indice di maggior utilizzo è proprio quello della classificazione Dewey, superiore anche a quello dell'indice dei nomi, peraltro particolarmente importante per i rinvii dei numerosi autori antichi e dei papi.

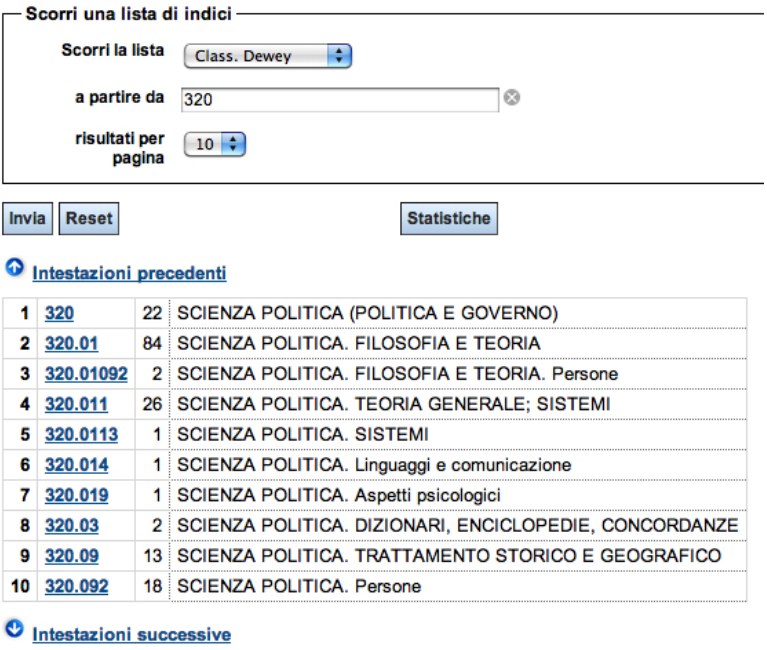

Figura 8: L'indice a scorrimento della classificazione Dewey in Koha.

10 Software utilizzato

I sette programmi di interrogazione sono stati scritti nel linguaggio Perl, ricorrendo alle API di Koha e alle seguenti librerie:[19] LWP per le connessioni HTTP, ZOOM per le connessioni Z39.50, DBI per le connessioni al database MySQL, XML::XPath per il trattamento dei dati XML, WWW::Scraper per il trattamento dei dati HTML, MARC::Record per il trattamento dei record MARC.

11 Conclusioni

Il presente lavoro ha permesso di comprendere il valore e le problematiche del reperimento in rete di informazione che può concorrere a migliorare cataloghi bibliografici. Ordinariamente si considera di interesse la catalogazione derivata per ottenere l'intero record, ma – attraverso identificativi univoci quali l'ISBN o altri – è possibile reperire informazione parziale o "atomica" con cui si possono raggiungere diversi scopi:

- arricchire il catalogo in modo statico, come nel caso presentato;
- arricchire l'OPAC in modo dinamico, recuperando uno o più dati al momento della visualizzazione di un record;
- aumentare la navigabilità per una migliore fruizione da parte dell'utente dell'OPAC;
- contribuire a bonificare situazioni pregresse;
- effettuare controlli di qualità;
- offrire strumenti di supporto al lavoro di catalogazione;

[19] Ogni libreria è documentata e reperibile in http://search.cpan.org.

- aumentare il numero di identificativi univoci presenti in catalogo;

- effettuare confronti tra basi di dati.

12 Appendice

12.1 Elementi della query per la selezione dei record senza Dewey

`biblionumber`	il numero di sistema del record bibliografico
`listaISBN`	`ExtractValue(marcxml,'//datafield[@tag=020]/subfield [@code=a]')` si tratta dell'elenco delle occorrenze del sottocampo $a del tag 020, separate da spazio; normalmente l'occorrenza è unica
`isbn_presente`	`ExtractValue(marcxml,'count(//datafield[@tag=020] /subfield[@code=a])>0')` almeno una occorrenza di 020$a
`dewey_assente`	`ExtractValue(marcxml,'count(//datafield [@tag=082]/subfield[@code=a])=0')` nessuna occorrenza di 082$a
`lingua_008`	substr(ExtractValue(marcxml,'//controlfield[\@tag=\008 \]'),36,3) = 'codice_lingua'

Tabella 9: Elementi principali della query per la selezione dei record bibliografici da trattare.

La funzione ExtractValue,[20] presente in MySQL 5.1.5 o superiori, permette l'interrogazione di dati XML, specificando come parametri il campo da esaminare e una espressione Xpath.[21]

[20] http://dev.mysql.com/doc/refman/5.1/en/xml-functions.html.
[21] http://it.wikipedia.org/wiki/XPath.

12.2 Parametri per le ricerche di tipo web

Per individuare i parametri con cui comporre l'url della ricerca, compreso quello dell'ISBN, si può procedere in uno dei seguenti modi:

- lanciare la query e notare l'url della risposta; se questo non contiene i parametri, cioè nel caso di form con `method=post`, cambiare il parametro `method` al valore `get` tramite "Inspect Element", presente in diversi browser premendo il tasto destro sulla *form*, e lanciare l'interrogazione;
- oppure analizzare la richiesta `http` inoltrata dall'interrogazione tramite un *plugin* per l'analisi del traffico o apposita funzionalità del browser.

12.3 Esempi di risposte

Un esempio di risposta XML da Classify[22] è il seguente:

Listing 3: XML

```
<?xml version="1.0" encoding="UTF-8"?>
<classify xmlns="http://classify.oclc.org">
  <response code="2"/>
  <!--Classify is a product of OCLC Online Computer Library
      Center: http://classify.oclc.org-->
  <work author="Beaucamp, Evode" editions="5" format="Book"
      holdings="69" itemtype="itemtype-book" title="Israel en
      prière : des Psaumes au Notre Père">014271167</work>
  <orderBy>hold desc</orderBy>
  <input type="isbn">2204022659</input>
  <start>0</start>
  <maxRecs>25</maxRecs>
```

[22] http://classify.oclc.org/classify2/Classify?summary=false&isbn=2204022659.

```
<editions>
  <edition author="Beaucamp, Evode" format="Book" holdings="
      40" itemtype="itemtype-book" language="fre"
      oclc=014271167 title="Israel en prière : des Psaumes
      au Notre Père">
    <classifications>
      <class edition=19 ind1=0 ind2=4 sf2=19 sfa=220.6
          tag=082/
      <class ind1="0" ind2="4" sfa="BS680.P64" tag="050"/>
    </classifications>
  </edition>
  <edition author="Beaucamp, Evode" format="Book" holdings="
      21" itemtype="itemtype-book" language="fre" oclc="
      299394640" title="Israel en priere : des psaumes au
      Notre Pere">
    <classifications>
      <class ind1="1" ind2="4" sfa="200" tag="082"/>
      <class ind1=" " ind2="4" sfa="BX2033B42 1985" tag="050"
          />
    </classifications>
  </edition>
  <edition author="Beaucamp, Evode" format="Book" holdings="
      5" itemtype="itemtype-book" language="fre" oclc="
      246374613" title="Israel en prière : des psaumes au
      Notre Père"/>
  <edition author="Beaucamp, Evode" format="Book" holdings="
      2" itemtype="itemtype-book" language="fre" oclc="
      442622354" title="Israel en prière : des Psaumes au
      Notre Père"/>
  <edition author="Beaucamp, Evode" format="Book" holdings="
      1" itemtype="itemtype-book" language="fre" oclc="
      718332441" title="Israel en prière : des Psaumes au
      Notre Père"/>
</editions>
<recommendations>
```

```
    [... omissis ...]
  </recommendations>
</classify>
```

Un esempio di risposta Z39.50[23] (MARC21), nella sua rappresentazione leggibile:

Listing 4: MARC21

```
00932cam 2200253 a 4500
001 500315
005 20050929180451.0
008 851021s1986 nyua 000 0 eng
035 $9 (DLC) 85073338
010 $a 85073338
020 $a 0874472466 (pbk.) : $c $8.95
040 $a DLC $c DLC $d DLC
050 00 $a LB2353.57 $b .A16 1986
082 00 $a 371.2/6 $2 19
245 00 $a 10 SATs : $b the actual and [...] prepare for it.
250 $a 2nd ed.
260 $a New York : $b College Entrance Examination Board : $b
    ...
300 $a 304 p. : $b ill. ; $c 28 cm.
[... omissis ...]
```

Un esempio di codice HTML:[24]

Listing 5: HTML

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http
    ://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
```

[23] Da Library of Congress, lx2.loc.gov:210/LCDB, find @attr 1=7 0874472466.

[24] https://portal.dnb.de/opac.htm?query=isbn%3D9783525563427&method=simpleSearch.

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="de" lang=
    "de" dir="ltr">
  <head>
  <title>DNB, Katalog der Deutschen Nationalbibliothek</title>
  [... omissis ...]
    </head>

  <body onload="doLoad()">
    [... omissis ...]
    <tr><td width="25" ><strong>Link zu diesem Datensatz</
        strong></td>
    <td>http://d-nb.info/977758214</td>
    </tr>
    [... omissis ...]
    <tr><td width="25%"><strong>DDC-Notation</strong></td>
      <td>231.6 [DDC22ger]</td>
    </tr>
    <tr>
    [... omissis ...]
  </body>
</html>
```

la cui versione nel browser è mostrata in figura 9 nella pagina successiva.

Link zu diesem Datensatz	http://d-nb.info/977758214 ID
Titel	Fides caritate formata : das Verhältnis v
Person(en)	Rose, Miriam
Verleger	Göttingen : Vandenhoeck & Ruprecht
Erscheinungsjahr	2007
Umfang/Format	303 S. ; 24 cm
Gesamttitel	Forschungen zur systematischen und ö
Hochschulschrift	Zugl.: München, Univ., Diss., 2004/200
ISBN/Einband/Preis	978-3-525-56342-7 Pp. : EUR 59.90 3-525-56342-6 Pp. : EUR 59.90
EAN	9783525563427
Sprache(n)	Deutsch (ger)
Schlagwörter	Thomas <de Aquino>: Summa theologi
DDC-Notation	231.6 [DDC22ger]
Sachgruppe(n)	230 Theologie, Christentum
Links	Inhaltsverzeichnis

Figura 9: Risultato di una ricerca per ISBN sul catalogo della Deutsche Nationalbibliothek.

Ai fini di una corretta indicizzazione, si invitano i lettori a citare esclusivamente il testo in lingua inglese; l'unico, infatti, che presenta l'indicazione del numero di pagina, l'abstract, le keywords e le date del processo redazionale.

Bargioni, S., M. Caputo, A. Gambardella, et al. "Recupero della classificazione decimale Dewey da altre basi di dati: un progetto di bonifica del catalogo". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art. #8766, p. 1–25. DOI: 10.4403/jlis.it-8766. Web.