

CONSIGLIO REGIONALE DELLA PUGLIA
BIBLIOTECA MULTIMEDIALE & CENTRO DI DOCUMENTAZIONE
“Teca del Mediterraneo”

8° WORKSHOP

Servizi e metodi innovativi per le Biblioteche e i Centri
di documentazione del XXI secolo

Tavola rotonda

*Pratiche di indicizzazione semantica poste a confronto:
thesauri, soggettari e classificazioni*

Opac semantici?

Sprechi e potenzialità negli accessi per classe

Claudio Gnoli

Università di Pavia. Dipartimento di Matematica. Biblioteca

Bari, 17/18 giugno 2005

L'indagine

L'indagine "Opac semantici", realizzata presso le Università di Venezia e Pavia con la collaborazione del GRIS e di altri colleghi, si è proposta di valutare l'utilizzo degli accessi per soggetti e per classi nei cataloghi in rete italiani. Tutte le informazioni, i riferimenti e i dati prodotti dall'indagine sono pubblicati sul sito del progetto: <<http://www-dimat.unipv.it/biblio/sem/>>.

In una prima fase è stata messa a punto un'apposita lista di controllo delle funzioni semantiche di ricerca e visualizzazione potenzialmente presenti in un opac. In base ad essa, in un opac semantico ideale la disponibilità di tutte le funzioni avrebbe prodotto un *indice di semanticità* pari a 100. Questo valore è stato confrontato con quello rilevato su un campione di 152 cataloghi reali (su circa 800 repertoriati in "AIB-WEB"), principalmente quelli comprendenti più biblioteche ubicate in diverse regioni (copertura *nazionale*, secondo la terminologia del MAI), o in diverse province (copertura *regionale*), o in diversi comuni (copertura *provinciale*). I rilevamenti sono stati compiuti da 13 studenti dell'Università di Venezia Ca' Foscari e da due volontarie, istruiti e coordinati sia durante il corso di Documentazione elettronica che attraverso un gruppo di lavoro telematico.

L'indice di semanticità più alto fra quelli misurati, totalizzato dal polo SBN della Romagna, ammonta a 54,7 punti su 100, e il grosso dei cataloghi si attesta intorno a 30 punti: segno che complessivamente le potenzialità della ricerca semantica sono poco sfruttate. I risultati, disponibili in dettaglio sul sito, sono già stati presentati in un articolo [Gnoli et al. 2004] e in due incontri pubblici a Firenze e a Venezia [Barazia 2004]. In questa sede ci concentreremo sugli aspetti relativi alla ricerca per classi.

Indicazioni per gli utenti sugli usi della classificazione

In Italia abbiamo una forte tradizione di cataloghi alfabetici per soggetti, ben rappresentata dai colleghi che mi precedono in questa tavola rotonda. I cataloghi sistematici sono invece meno diffusi: spesso infatti la classificazione è considerata solo come uno strumento per collocare, mentre sul piano della catalogazione è pensata come una sorella minore. A ciò contribuisce l'espressione *indicizzazione per soggetto*, che fa pensare soprattutto ai cataloghi alfabetici per soggetti, ma che in realtà traduce l'inglese *subject indexing*, includente la classificazione. Di fatto, soggettazione e classificazione non sono rivali, ma complementari e strettamente interdipendenti, come ha ricordato di recente Revelli [Revelli 2005].

In compenso, esiste un valido gruppo di lavoro che traduce in italiano, cura e pubblica regolarmente le nuove edizioni della Classificazione decimale Dewey, di gran lunga lo schema più utilizzato e conosciuto nel nostro paese. Tale prevalenza della CDD si rispecchia anche tra gli opac esaminati: di quelli che indicano il sistema di classificazione utilizzato, 54 citano la Dewey, mentre solo 7 la CDU, 5 lo schema della Library of Congress, e 6 schemi sviluppati localmente. Tra gli schemi meno diffusi è interessante notare la presenza in più di un opac del sistema Iconclass, della classificazione RVK prodotta dalla biblioteca di Ratisbona, e della Mathematics subject classification dell'AMS. Ma il risultato più notevole è il fatto che altri 94 cataloghi (quindi una maggioranza assoluta) non indichino affatto quale sistema di classificazione utilizzano, sebbene molti di essi contengano dati di classificazione. Talvolta la scelta della Dewey è considerata scontata (ma lo è anche per gli utenti?), talaltra semplicemente non ci si è preoccupati di fornire l'informazione.

Ancora più rara è l'indicazione di quale edizione dello schema sia utilizzata nel catalogo (CDD 19, 20, 21...): la forniscono solo 5 opac su 150. Dietro questa indefinitezza si nasconde una realtà problematica: non è infrequente infatti che i cataloghi più grandi contengano registrazioni provenienti da vari archivi fusi tra loro in epoche successive, alcuni dei quali usano un'edizione della Dewey, altri una diversa, altri ancora hanno vuoto il campo per il numero di classe... Si tratta di inconvenienti comprensibili – nessuno può biasimare un ente perché sta lavorando in direzione di un unico catalogo collettivo, né pretendere che questo sia immediatamente perfetto e coerente – che possono essere fronteggiati inserendo un campo separato relativo al numero dell'edizione, e con gradualità lavori di bonifica. Quello che però dovrebbe essere sempre chiaro è la situazione attuale del contenuto del catalogo, in modo che chi conduce le ricerche possa interpretare correttamente i risultati.

Una situazione ancora più negativa si è rivelata riguardo all'uso della classificazione per le segnature: nonostante, come abbiamo detto, si tratti di un'applicazione frequente, solo 3 cataloghi si preoccupano di informarci della scelta... Negli altri, per l'utente ignaro e non particolarmente appassionato di crittografia, le sequenze alfanumeriche di collocazione potrebbero tanto essere arbitrarie quanto veicolare profondi ed esoterici significati, dai quali egli è comunque escluso. Le collocazioni formate con numeri di classe, inoltre, sono sfruttabili in fase di ricerca solo in 9 degli opac esaminati.

Queste mancanze di informazioni associate alle maschere di ricerca appaiono tanto più gravi perché potrebbero essere risolte con poca attenzione supplementare e nessuno sforzo tecnico: basterebbe che gli opacmaster, invece di limitarsi a far funzionare ciecamente la macchina, si fermassero qualche volta a considerare a che cosa questa debba servire. Diversa è la situazione per le funzioni di ricerca e di visualizzazione, che esaminiamo nel seguito, le quali invece comportano certi requisiti a monte: nel lavoro dei catalogatori, nella struttura della base-dati, nella progettazione dell'interfaccia. È qui che il matrimonio fra biblioteconomia e informatica avrebbe l'occasione di mostrare i suoi frutti migliori, a condizione però non solo che entrambi i genitori siano in buona salute, ma che abbiano imparato ad andare d'accordo...

Ricerche per classi: un problema di equivalenti verbali

Ricerche sistematiche per liste sono risultate effettuabili nel 53,9% degli opac esaminati, e ricerche sistematiche per campi nel 50,7%. Questi dati sono nettamente inferiori a quelli delle corrispondenti ricerche alfabetiche per soggetti, rispettivamente 75,7% e 84,9%. Questa prevalenza dei soggetti non significa necessariamente che nei cataloghi non siano presenti dati di classificazione. Infatti numerosi opac, fra i quali molti che utilizzano il software Easyweb, il numero di classe viene presentato nelle schede dettagliate trovate con una ricerca, quale informazione ausiliaria, ma non è possibile utilizzarlo nella fase di ricerca, né per navigare a partire dalle schede trovata dopo una prima ricerca. Insomma, il dato c'è, ma è praticamente inutile: una volta che avete trovato un libro per altre vie, quale soddisfazione vi arreca il sapere che è descritto dal codice Dewey 499.52?...

Talvolta i documenti di una data classe sarebbero estraibili perlomeno scrivendo un'apposita query nel linguaggio del sistema, tuttavia questa possibilità non è neppure indicata! Ancora: succede in istituzioni anche importanti che i catalogatori coscientemente immettono dati di classificazione (dati che come è noto richiedono un'analisi del documento piuttosto impegnativa e una certa competenza sul sistema di classificazione), ma che poi questi dati non sono visibili affatto in alcuna parte del catalogo. Che livello di soddisfazione

professionale hanno questi colleghi? Faranno in tempo a vedere utilizzato il loro lavoro prima di andare in pensione?...

Un altro dato rivelatore del fatto che sulla classificazione si investa poco è il seguente. Mentre la ricerca per soggetti, se è disponibile, viene sempre presentata come opzione fin dalla maschera di ricerca semplice, invece la ricerca per classi in 20 casi su 35 è presentata solo nella maschera avanzata e non in quella semplice. Questo significa che la classificazione è considerata una funzione sofisticata e non di base. È una scelta spiegabile, se vogliamo, in quei casi in cui non sono disponibili gli equivalenti verbali, sicché l'unico modo di cercare per classi è inserire gli esoterici simboli dello schema, operazione presumibilmente riservata a utenti molto esperti e ai bibliotecari stessi. Purtroppo questi casi sono la grande maggioranza: mentre gli opac che permettono di cercare per simboli sono 101, quelli che permettono di usare gli equivalenti verbali sono solo 31. Anche in fase di visualizzazione dei risultati, in 53 casi sono offerti solo i numeri di classe, mentre solo in 23 casi si vedono anche gli equivalenti verbali. Nelle schede dettagliate di ciascun documento, infine, ci sono i numeri di classe in 111 casi ma gli equivalenti verbali solo in 43.

E qui arriviamo al nocciolo della questione: perché la classificazione sia utilizzabile dall'utente comune, che difficilmente è nato con le tavole della CDD cablate nel cervello, è ovvio che debba essere permessa una prima ricerca per parole, in linguaggio quasi-naturale: a quel punto il sistema proporrà tutte le classi che contengono nei loro equivalenti verbali i termini richiesti, l'utente ne sceglierà qualcuna e arriverà così nella zona pertinente delle tavole; lì, grazie all'ordinamento sistematico prodotto dalla notazione, potrà scorrere le voci in avanti e all'indietro trovando altre classi affini, e scoprendo talvolta che queste fanno al caso suo ancora più delle precedenti; e infine perverrà a un insieme ben selezionato di documenti che parlano di ciò che gli interessa. È chiaro che un'interfaccia che permetta solo un'interazione tramite simboli risulterà di poca utilità alla massa degli utenti, e non ci sarà da stupirsi allora che le statistiche indichino che alla ricerca semantica sia preferita quella per autore e titolo, come riportava Scolari alcuni anni fa [Scolari 1999]. Come dire: preferite usare un treno sempre in ritardo, ultraaffollato e con le porte che non funzionano, oppure la vostra auto?...

Ma perché così spesso la ricerca e la visualizzazione degli equivalenti verbali non sono disponibili? Una parte della ragione è tecnica: gli equivalenti verbali richiedono di essere gestiti parallelamente ai corrispondenti numeri di classe in apposite tabelle. Nel momento in cui un catalogatore crea un nuovo numero Dewey per combinazione di un numero base e di isolati comuni (come "provincia di Bari", "ventesimo secolo", "periodici" ecc.), dovrebbe venire prodotto manualmente o automaticamente anche il corrispondente equivalente verbale, il che implica un certo sforzo organizzativo [Giunti 2001]. La classificazione è uno strumento sofisticato, fatto di diverse componenti, ciascuna delle quali richiede un adeguato trattamento informatico: per trarne i frutti è necessario crederci e investirci adeguatamente [Gnoli 2003] e non limitarsi ad abbandonarla agli automatismi delle macchine come un qualsiasi codice identificativo.

Ragionando sui risultati dell'indagine con Giulia Visintin abbiamo trovato anche un'altra possibile ragione, più a monte: i dati delle classi CDD, spesso derivati da altri opac, circolano in gran parte nella sola forma simbolica; gli equivalenti verbali addirittura, in base alle condizioni poste dall'editore della Dewey, non sarebbero liberamente riproducibili [AIB 1992]. Starebbe allora a ogni singolo catalogo reinserire dei propri equivalenti verbali, operazione come dicevamo costosa, e che inoltre non favorisce certo la standardizzazione di quello che dovrebbe essere un vocabolario controllato.

Esempi virtuosi

Se da un lato è vero che la classificazione è impegnativa, dall'altro ci sono esempi virtuosi che mostrano quanto può essere remunerativa in termini di recupero di informazioni rilevanti. Un software che dimostra di sfruttarla seriamente è quello realizzato dall'Università di Firenze, e poi acquisito anche dalle università di Padova e di Bologna. Non a caso è questo il software che ha ottenuto largamente il miglior indice medio di semanticità: evidentemente metterlo a punto "in casa" ha consentito una più stretta aderenza delle soluzioni informatiche alle necessità della ricerca bibliografica. Gli opac che lo utilizzano permettono ricerche con equivalenti verbali e visualizzazione in ordine sistematico di tutte le classi che li contengono, in qualsiasi posizione e non soltanto all'inizio – una funzione quest'ultima assai importante ma offerta solo da un altro software, Bookmarkweb.

Un modo molto efficace di guidare gli utenti nella ricerca sistematica è quello di offrire loro interfacce con liste ordinate di voci verbali che, attraverso collegamenti ipertestuali, permettano di estrarre le registrazioni contenenti i simboli corrispondenti. In questo modo, anche quando per i problemi descritti gli equivalenti verbali non sono disponibili nelle tabelle del catalogo, possono comunque essere proposti al livello dell'interfaccia: l'utente clicca sul termine "diritto internazionale" e la query estrae per lui tutti i documenti la cui classe comincia per 341. Ai livelli più alti, questa funzione è disponibile in molti opac Sebina e Unibiblio, che dispongono di pagine ausiliarie con spiegazioni sulla struttura e la funzione della CDD, a partire dalle quali è possibile navigare nei primi livelli dello schema e quindi richiamare i documenti associati. Alle biblioteche che possiedono documenti molto specializzati questo non basterà, ma si possono realizzare menù ad hoc che estraggano nello stesso modo insiemi di documenti di classi rilevanti. È quanto è stato fatto all'Università di Pavia dal Dipartimento di Ecologia del territorio per la propria sala Flore, i cui atlanti botanici sono classificati geograficamente <<http://et2.unipv.it/homepage/biblioteca/fito.htm>>, nonché dal Dipartimento di Matematica per tutta la propria raccolta, seppure solo attraverso una collocazione classificata <<http://www-dimat.unipv.it/biblio/scaffali.html>>. Queste applicazioni sfruttano la struttura degli URL dinamici dell'opac Easyweb, nei quali è compreso il codice di classificazione di volta in volta richiesto [Gnoli 2001].

Una funzione in cui la classificazione rivela il suo grande potenziale informativo è la ricerca di simboli troncati: per la natura stessa della notazione decimale [Gatto 2001], troncadola a destra all'altezza di una determinata cifra si ottengono automaticamente anche tutte le classi più specifiche riferite allo stesso soggetto, un risultato che nella maggior parte dei casi è assai utile. Non è difficile predisporre interfacce che permettano, in modo predefinito oppure a richiesta, un recupero inclusivo di tutte le classi subordinate alla classe scelta. Sul totale degli opac, abbiamo trovato un 19,7% che offre una ricerca per classe troncata automaticamente (perlopiù con i software Sebina, Playlib e Bookmarkweb; solo in alcuni ALEPH, e in un solo Easyweb), e un 30,1% che offre la possibilità di troncatura a richiesta.

È interessante anche la possibilità di troncatura a sinistra: se in molti casi non produce associazioni significative (le classi 343.227, 893.227 e 653.227 non hanno probabilmente alcuna relazione tra loro), questo avviene invece quando i numeri sono costruiti con suddivisioni comuni. Lo stesso standard UNIMARC prevederebbe appositi codici che «facilitate computer manipulation of synthesised numbers to indicate separately parts of numbers for which it may be desirable to search explicitly» (campi 661-668) ma dubito che questa potenzialità sia sfruttata nella realtà dei nostri cataloghi. Ad esempio, nella CDD la notazione -03 equivale sempre ai dizionari, e -0945751 sempre alla provincia di Bari: richiedendo quindi tutte le registrazioni il cui numero di classe comprende la sequenza -0945751-, si potrebbero ottenere tutti i documenti che abbiano a che fare con la provincia

di Bari, compresi i suoi singoli comuni, magari da combinare con altre condizioni. Questa funzione è risultata presente nell'11,0% degli opac, perlopiù con gli stessi software precedenti nonché con quello usato dai poli SBN di Lombardia, Piemonte e Umbria. Tuttavia raramente essa è evidenziata e spiegata agli utenti nel suo significato: più spesso si tratta ancora di sapienza esoterica.

Sempre a livello di notazione, una situazione problematica si verifica quando nel catalogo sono presenti simboli di classe provenienti da più di un sistema: se questi coesistono in una sola tabella producono un notevole rumore informativo, poiché lo stesso simbolo 35 nella CDU significa "amministrazione pubblica", mentre nella MSC significa "equazioni alle derivate parziali"... Ciononostante, dei cataloghi che dichiarano di usare più di un sistema di classificazione, solo un terzo permette di limitare la ricerca a un solo sistema. Anche qui la carenza va risolta a livello delle tabelle della classificazione, utilizzando un apposito campo che specifichi lo schema a cui i dati si riferiscono.

Completamente assenti, infine, si sono rivelati negli opac italiani i rinvii da una classe all'altra: li trovano i catalogatori sui manuali di classificazione, ma non i poveri utenti nei cataloghi. In questo senso molto meglio fanno i repertori di siti Internet come "Google directory" e "Yahoo!", che nei loro alberi sistematici incorporano i rinvii trasversali da una classe all'altra, utilizzati dagli utenti quasi senza accorgersene.

Un potenziale non ancora sfruttato

Complessivamente, dunque, l'indagine ha rivelato notevoli carenze negli accessi semantici della massa dei cataloghi in rete italiani, carenze ancor più gravi per la classificazione che per i soggetti alfabetici. Per essere risolte, alcune di esse richiederebbero maggiore impegno a livello catalografico, molte altre a livello informatico, mentre parecchie potrebbero essere risolte anche solo con semplici accorgimenti e maggiore attenzione al buon senso degli utenti qualunque.

Nei simboli di classe è concentrata in pochi byte una grande ricchezza informativa, grazie alla struttura degli schemi ai quali i simboli si riferiscono (è un po' la stessa filosofia adottata dall'emergente Web semantico), schemi escogitati con il lavoro intellettuale di generazioni di esperti bibliotecari. Ma come abbiamo visto questa ricchezza è in gran parte sprecata, non tanto per difetti intrinseci alle singole componenti informative, quanto per la mancanza di adeguati collegamenti fra loro nel formare il sistema informativo integrato che il catalogo dovrebbe rappresentare. Per cogliere questi problemi a tutti i livelli – non solo quello catalografico, o solo quello informatico, o solo quello delle interfacce – e nelle loro interrelazioni occorre certo un po' di sensibilità; ma probabilmente è proprio questa integrazione fra teoria biblioteconomica e tecnica informatica, entrambe potenti ma meno efficaci finché prese da sole, ad essere il Graal di cui oggi dobbiamo andare alla ricerca.

Riferimenti bibliografici

- AIB 1992:** I diritti di Dewey = AIB notizie. 4: 1992. n 10. p 10
- Barazia 2004:** Opac semantici : indagine sugli accessi semantici nei cataloghi in rete italiani : Venezia, 20 ottobre 2004 [resoconto] / Caterina Barazia = AIDA informazioni. 22: 2004. n 4. p 83-85 | <<http://www.aidainformazioni.it/pub/barazia42004.html>>
- Gatto 2001:** Classificazione e automazione / Eugenio Gatto = Quali spazi per le classificazioni? : tavola rotonda AIB Piemonte : Torino, 20 gennaio 2001 [resoconto (ISKO Italia) – <<http://www.iskoi.org/doc/spazi.htm#gatto>> : 2004-]
- Giunti 2001:** In SBN con Dewey: il catalogo classificato del Polo della Biblioteca nazionale centrale di Firenze / Maria Chiara Giunti = Bollettino AIB. 41: 2001. n 1. p 31-46
- Gnoli 2001:** Scaffali telematici: una rianimazione del catalogo classificato / Claudio Gnoli = Bibliotime. 4: 2001. n 1 <<http://www2.spbo.unibo.it/bibliotime/num-iv-1/gnoli.htm>>
- Gnoli 2003:** La classificazione come investimento nella qualità dell'informazione / Claudio Gnoli = AIB-WEB. Contributi. – AIB <<http://www.aib.it/aib/contr/gnoli4.htm>> : 2003-
- Gnoli et al. 2004:** Di che parla questo catalogo? : un'indagine sugli accessi semantici negli opac italiani / Claudio Gnoli, Riccardo Ridi, Giulia Visintin = Biblioteche oggi. 22: 2004. n 8. p 23-29 | <<http://www.bibliotecheoggi.it/2004/20040802301.pdf>>
- Revelli 2005:** Soggettazione a faccette / Carlo Revelli = Le dimensioni dell'informazione: giornata di studio AIB Piemonte-ISKO Italia : Torino, 12 giugno 2004 : atti. [2] / Caterina Barazia, Claudio Gnoli: cura – ISKO Italia <<http://www.iskoi.org/doc/dimensioni1.htm>> : 2005-
- Scolari 1999:** Efficacia vs facilità: linee di evoluzione degli OPAC / Antonio Scolari (La biblioteca amichevole: convegno: Milano, 11-12 marzo 1999) = Biblioteche oggi. 17: 1999. n 5. p 18-26