

Un server Z39.50 per l'Università di Brescia

Zeno Tajoli

CILEA, Segrate

Abstract

Per connettere l'Opac dell'Università di Brescia a MetaLib è stato implementato un server z39.50. Viene spiegato a livello generale in cosa consiste lo standard Z39.50. Vengono spiegate le caratteristiche basilari del software utilizzato, Cheshire II, e le operazioni svolte per rendere operativo il servizio. Infine sono elencati i dati necessari per la connessione e la ricerca attraverso un client Z39.50.

Keywords: Protocollo Z39.50, Bath profile v. 2, OPAC, Cheshire II

La biblioteca centrale Interfacoltà dell'Università di Brescia ha integrato l'Opac dell'ateneo [8] con il nuovo portale UniBSirio per l'accesso alle risorse elettroniche, gestito attraverso MetaLib e SFX[11].

La richiesta che era giunta dall'università era che l'integrazione fosse la più ampia possibile e che fossero sfruttate al massimo le potenzialità dei *software* MetaLib e SFX. In particolare MetaLib permette l'integrazione di risorse esterne sia via http che via Z39.50. Da questi protocolli MetaLib può ricevere sia dei formati non strutturati (tipo una pagina web in html) sia strutturati (tipo una pagina XML). I dati vengono passati a specifici parser che estraggono i dati e li rendono disponibili al resto dell'applicazione. Più il dato è formalizzato secondo regole e standard, meglio viene compreso dall'applicazione. In particolare la funzione che deduplica i record bibliografici identici provenienti da diverse fonti funziona al meglio se riceve dati altamente strutturati. Inoltre MetaLib può gestire le raffinate tipologie di ricerche dello standard Z39.50 se la base su cui avvengono le ricerche le supporta.

Considerando questo scenario si è deciso che la soluzione migliore fosse di costruire un vero e proprio server Z39.50 con dati in standard MARC e con una buona dotazione di indici.

Lo standard Z39.50

Realizzare l'integrazione tramite il colloquio via Z39.50 vuol dire usare uno strumento molto raffinato per gestire la sessione di ricerca. Lo standard è mantenuto ufficialmente dalla Li-

brary of Congress, che cura anche un sito web molto aggiornato sull'argomento [12]. Il protocollo nasce per gestire una sessione di ricerca di dati bibliografici tra un client e un server[1]. In particolare l'accento viene posto sull'uniformità sul lato client; una ricerca con una stessa sintassi può essere lanciata su diversi server. Il protocollo permette di usare diversi tipi di query, ad esempio può essere usato il linguaggio nativo del server ("query type 0") se esso è compreso dal client. Questa situazione si verifica di norma quando client e server sono dello stesso produttore. Vi sono poi altri tipi, alcuni definiti nello standard (ad esempio il query type 1 e il 101, il tipo 1 è quello più diffuso). Altri tipi ancora sono definiti esternamente da altri standard. Ad esempio questa casistica riguarda il query type 2 e il query type 100. In totale vi sono 7 query type [2].

Una volta definito come comporre la struttura della query, bisogna accordarsi sulla grammatica dei singoli elementi. Lo Z39.50 supporta numerose grammatiche (il nome usato è "attribute sets"), ognuna con il suo campo di applicazione. Per le ricerche bibliografiche l'attribute set più usato è il bib-1[5]. Bib-1 è un attribute set molto complesso e diversificato. Copre sia i punti di accesso di un database basato su MARC (autori persona distinti da autori enti, titolo proprio distinto dai titoli paralleli, etc.) sia i punti di accesso di un database che contiene dati strutturati in Dublin Core (un solo punto d'accesso per tutti gli autori, uno solo per tutti i titoli, etc.). Per questo motivo difficilmente un

server avrà tutti i punti d'accesso possibili secondo bib-1. Oltre ai punti d'accesso e alla loro identificazione attraverso uno specifico codice numerico, bib-1 specifica una serie di parametri che informano come il server debba considerare le parole e gli operatori che compongono la query:

1. Use (1). Indica il punto d'accesso preso in considerazione (ad esempio autore persona, creator di Dublin Core, etc.). Alcuni codici sono: 4 (titolo), 7 (isbn), 8 (issn), 1097 (titolo DC), 31 (data pubblicazione)
2. Relation (2). Indica la relazione tra il punto d'accesso (sulla destra) con il termine della ricerca così come qualificato dalle specificazioni. (ad esempio data di pubblicazione => 1997). Alcuni codici sono: 1 (minore di), 3 (uguale), 6 (diverso da), 100 (uguaglianza fonetica).
3. Position (3). Indica dove è posto il termine di ricerca all'interno del campo o del sottocampo in cui è presente. Chiaramente pensata riferendosi a dati in formato MARC. I codici possibili sono: 1 (il termine si trova in prima posizione nel campo dove è presente); 2 (il termine si trova in prima posizione nel sottocampo dove è presente), 3 (il termine si trova in qualsiasi posizione).
4. Structure (4). Indica la tipologia del termine di ricerca (ad esempio se si tratta di una parola chiave, di una frase, di una data, etc.). Alcuni codici sono: 1 (frase), 2 (parola), 4 (anno).
5. Truncation (5). Indica come il server deve trattare i termini che riceve, ad esempio se considerarli troncati a destra, oppure se considerare i termini come non troncati. Alcuni codici sono: 1 (troncamento a destra), 2 (troncamento a sinistra), 100 (non troncato).
6. Completeness (6). Indica se i termini della ricerca occupano o no interamente il campo o il sottocampo dove sono presenti. Anche questa specificazione è chiaramente pensata riferendosi a dati in formato MARC. I codici possibili sono: 1 (i termini della ricerca non occupano tutto il campo o il sottocampo in cui sono presenti); 2 (i termini della ricerca occupano tutto il sottocampo in cui sono presenti, altre parole possono essere presenti nel resto del campo); 3 (i termini della ricerca occupano tutto il campo in cui sono presenti).

La semantica di bib-1 in gran parte è stata definita nel 1995 [3], poi ci sono state delle estensioni, introdotte attraverso il dibattito tra gli

implementatori di Z39.50 riuniti nell'iniziativa denominata ZING[13]. Poiché spesso bib-1 non è implementato in tutte le sue opzioni, sono stati sviluppati degli accordi tra gli utilizzatori per definire un minimo comun denominatore. Il termine usato per chiamare questi accordi è "profile". Nel campo dei dati bibliografici il profile più citato è il Bath Profile, giunto alla release 2.0 [4].

Lo Z39.50 è dunque uno standard per l'information retrieval, nato per gestire solo informazioni bibliografiche ma poi ampliato alle esigenze di altri enti come archivi, musei, etc. Presso il pubblico generico esso è stato ormai sostituito dalle interfacce via web degli Opac e degli altri strumenti di ricerca. I client per il web sono molteplici, vari e quasi sempre gratuiti. Invece i client Z39.50 sono quasi sempre a pagamento e molto meno diffusi. Le previsioni di Scolari [9] al riguardo si sono puntualmente avverate.

Ritengo che il futuro di Z39.50 non si trovi più in client usati direttamente dall'utente finale ma nell'essere un protocollo specializzato che verrà usato da diversi programmi per parlarsi tra loro. In questa direzione va l'iniziativa dello ZING di creare un forma di Web Service per Z39.50, iniziativa denominata SRW/SRU [10]

Il software selezionato, Cheshire II

Lo strumento software utilizzato per integrare l'OPAC nelle risorse di MetaLib è Cheshire II [6][7]. Esso è un Information Retrieval che colloquia con l'utenza sia attraverso lo Z39.50 che attraverso l'http. I dati vengono immessi in formato XML e possono uscire sia in formato XML che in formato MARC. Inoltre gli sviluppatori fanno parte dello ZING e sono particolarmente coinvolti nell'iniziativa SRW/SRU. La documentazione è buona e il codice è abbastanza ben documentato. Il codice è in C ed è disponibile con licenza GPL. Manca però una documentazione dell'organizzazione interna del codice, della sua struttura e della sua logica. Questo rende difficoltoso un intervento diretto sul codice sorgente. Tuttavia la mailing-list collegata al software è molto reattiva quando si segnala un problema.

L'implementazione del server Z39.50

Una volta selezionato Cheshire II come piattaforma di partenza, è iniziato il lavoro per caricare i dati. I dati di partenza provengono dallo scarico UNIMARC di SBN Unix C/S e da una base dati locale. Essi sono in formato UNIMARC con caratteri in ISO 646 e ISO 5426 (7 bit). Al

momento attuale Cheshire II può gestire al meglio solo dati MARC21. La gestione di dati UNIMARC è possibile, ma necessita di alcune personalizzazioni all'interno del codice C. Il client Z39.50 di MetaLib non riesce a interpretare l'ISO 5426 a 7 bit e la conversione da UNIMARC a MARC21 (il formato interno di MetaLib) non è particolarmente sofisticata. Per tutti questi motivi si è deciso di realizzare un convertitore personalizzato da UNIMARC a MARC21, convertendo anche i caratteri da ISO 5426 (7 bit) ad ANSEL¹. La scrittura di questo software è stata abbastanza impegnativa e probabilmente sarà materia di un prossimo articolo.

Successivamente si è configurato Cheshire II per ottenere le caratteristiche desiderate. In particolare bisogna gestire le esigenze, per lo stesso punto d'accesso, della ricerca per parole chiave da un lato e di scorrimento delle intestazioni dall'altro. Questa duplice esigenza ha portato a definire, in alcuni casi, due indici diversi per lo stesso punto d'accesso. In questo modo quando si chiedono delle keywords questo avviene dando all'attributo struttura il valore 6 e viene interrogato un indice per keywords. Quando invece si effettua lo scorrimento per le intestazioni, si usa l'attributo struttura con il valore 1. In questo caso viene usato un altro indice, che crea come termine di ricerca l'intera frase. Per ottenere questo risultato va opportunamente settato il file di configurazione usato da Cheshire II per generare gli indici.

Un file di configurazione con queste caratteristiche è stato inviato agli sviluppatori di Cheshire per essere inserito come esempio nella distribuzione del software.

I dati del server Z39.50

Il server Z39.50 attivato ha queste caratteristiche:

- Indirizzo: isis.cilea.it
- Porta: 2100
- Nome database: usmarc
- Formati di output: MARC21, XML, SUTRS
- I caratteri sono nel set ISO 646 e ANSEL
- Del Bath profile supporta: area A level 1, area B level 1, area D level 1.

Acknowledgments

Si fa presente che la configurazione di Cheshire II usata a Brescia deve molto a quanto scritto da Robert Sanderson e da Ray Larson, sia nella documentazione resa disponibile, sia in risposta a domande specifiche.

Bibliografia

- [1] ANSI/NISO Z39.50-2003:
<http://lcweb.loc.gov/z3950/agency/Z39-50-2003.pdf>
- [2] ANSI/NISO Z39.50-2003:
<http://lcweb.loc.gov/z3950/agency/Z39-50-2003.pdf>, p. 23-24
- [3] Attribute set Bib-1 (Z39.50-1995), semantics:
<ftp://ftp.loc.gov/pub/z3950/defs/bib1.txt>
- [4] The Bath Profile:
<http://www.collectionscanada.ca/bath/tp-bath2-e.htm>
- [5] Bib-1 Attribute Set:
<http://lcweb.loc.gov/z3950/agency/defns/bib1.html>
- [6] Cheshire II Project:
<http://cheshire.berkeley.edu/>
- [7] Cheshire Resources:
<http://cheshire.berkeley.edu/>
- [8] Opac Università di Brescia:
<http://opacunibs.cilea.it>
- [9] Scolari, *World Wide Web e Z39.50: standard per la ricerca a confronto* in Bollettino AIB, 1996 n.4, p. 397-407:
<http://www.aib.it/aib/boll/1996/96-4-397.htm>
- [10] SRW/U:
<http://lcweb.loc.gov/z3950/agency/zing/srw/>
- [11] UniBSirio:
<http://www.bci.unibs.it/biblio/sub144.html>
- [12] Z39.50 Maintenance Agency Page:
<http://www.loc.gov/z3950/agency/>
- [13] ZING Z39.50 International:
<http://www.loc.gov/z3950/agency/zing/zing-home.html>

¹ ANSEL è un set di caratteri a 8 bit. In MARC21 i caratteri sono in un formato chiamato MARC-8; esso è formato da ISO 646, da ANSEL e da altri set di caratteri. Solo di recente si è iniziato ad usare UTF-8.