8. CDSWARE

8.1 INTRODUZIONE

Il presente capitolo ha per oggetto la descrizione dell'applicazione CDSware, software di interesse della sezione sperimentale del presente elaborato di tesi.
Per prima viene data una descrizione di tipo generale di tale prodotto, ideato, sviluppato ed utilizzato al CERN di Ginevra e delle sue caratteristiche; gran parte dei concetti qui espressi derivano da un lavoro della dott.ssa De Robbio [S19].
Successivamente vengono presi in considerazione gli utilizzi dell'applicazione al sistema bibliotecario del CERN, appresi anche durante lo stage formativo svolto presso tale struttura, nel periodo dal 28 novembre al 5 dicembre del 2003.
Infine, utilizzando la documentazione interna al programma di istallazione di CDSware v.0.0.9, vengono descritti i moduli costituenti tale prodotto software.

In particolare, il modulo BibConvert, costituisce la parte fondamentale della sezione sperimentale, presente al capitolo successivo.

8.2 DESCRIZIONE DI CDSWARE

L'applicazione CDSware (CERN Document Server Software) permette lo sviluppo e la gestione del proprio sistema documentario di tipo open archive sul web, conforme al protocollo OAI-PMH [B1] e che utilizza il formato MARC 21, trattato al capitolo 4, come standard bibliografico per i suoi metadati.

E’ possibile realizzare in tal modo un server di e-print, un catalogo bibliotecario online, ecc..
Le sue caratteristiche sono, al tempo stesso, quelle di Data e Service Provider, in quanto permette sia la disseminazione dei metadati in esso contenuti, sia la fornitura di servizi a valore aggiunto di varia natura.

E’ stato creato dagli sviluppatori del CERN di Ginevra prendendo in considerazione i bisogni e le richieste del comitato di biblioteca afferente al servizio di informazione scientifica, che è una sezione facente parte dell’organigramma strutturale del CERN.
L’applicazione è tutt’ora in corso di sviluppo al CERN Document Server, che è il sito dove risiede l’istallazione principale di CDSware.

Dato che CDSware è un software free, esso viene distribuito sotto licenza GNU General Public Licence (GPL) il cui testo è consultabile presso il sito Internet http://www.gnu.org/licenses/gpl.html.

Il supporto, offerto agli utenti che intendono istallare e gestire la propria applicazione, dallo staff di CDSware, è individuato da due tipi distinti:

- *Supporto gratuito*
- *Supporto a pagamento*

Il primo consistente nella possibilità di inviare, attraverso e-mail (cds.support@cern.ch) o mailing list (http://cdsware.cern.ch/lists/) i propri commenti, eventuali bug riscontrati, e qualunque altra idea o problema sul prodotto, ai quali verrà fornita una risposta nei tempi necessari.
Il secondo consistente in un supporto di tipo commerciale che prevede un contratto speciale di gestione che può essere accordato con il gruppo di sviluppo del CDS su base annua, affinché gli sviluppatori possano aiutare fisicamente ad

istallare, configurare e gestire il sistema. Il contatto è possibile attraverso l'indirizzo email cds.support@cern.ch.

Al CERN, CDSware gestisce più di 400 collezioni di dati, consistenti di oltre 650.000 record bibliografici, includendo 320.000 documenti fulltext, tra cui "preprints", "articles", "books", "journals", "photographs", ed altri ancora.
Ogni giorno oltre 200 documenti vengono sottomessi all'applicazione e l'8% di essi, proviene proprio dal CERN, mentre il restante è importato da circa 80 sorgenti nel mondo, che non sono OAI compatibili. Ciò è reso possibile grazie all'utilizzo di un software esterno a CDSware, chiamato Uploader, che si occupa di risolvere l'importazione convertendo il formato esterno di tali sorgenti in un formato che può essere immagazzinato in Aleph, sistema di gestione di dati bibliografici (vedi paragrafo 8.3). Tale applicazione è stata scritta da uno dei membri dello staff di sviluppo di CDSware: Martin Vesely.

Altre istallazioni di CDSware sono le seguenti:

- *Atlantis Institute of Scienze*: è il sito dimostrativo di CDSware che contiene semplicemente pochi record bibliografici di esempio e utilizza l'unica versione stabile del prodotto, la 0.0.9.
- *Atlantis Institute of Fictive Scienze*: è il sito dimostrativo utilizzato per testare versioni in via di sviluppo e non riconosciute stabili del prodotto. Nuove funzionalità in fase di miglioramento sono infatti: ricerca basata su combinazioni di metadati, fulltext e citazioni, document basket personalizzabile dall'utente, email alert, ecc..

8.2.1 Caratteristiche di CDSware

Le caratteristiche principali di CDSware sono:

- *Interfacce tipo portale* – sono configurabili via web per la gestione di vari tipi di collezioni ed offrono:
    - servizi di ricerca per tutti i possibili campi indicizzabili
    - browsing nelle collezioni
    - meccanismo di sottomissione dei propri documenti
    - personalizzazione degli accessi
- *Motore di ricerca potente* – adotta una sintassi tipo Google che consente la ricerca nei campi informativi dei metadati e nei fulltext depositati
- *Sottomissione e aggiornamento guidati* – disponibili per vari tipi di documenti, articoli, libri, preprint, ecc.
- *Meccanismo di approvazione* – consente ai responsabili di un eventuale comitato di approvazione dei documenti sottoposti, di accettare o rifiutare i lavori sottomessi dai vari utenti
- *Funzione di data e service provider OAI compatibili* - consente lo scambio di metadati tra repository eterogenei

Da un punto di vista tecnico, CDSware "gira" sotto un sistema Linux compatibile e prevede l'istallazione di diverse componenti software di base:

- MySQL server e client – utilizzato per gestire le tabelle costituenti il repository di CDSware (<http://mysql.com/>)
- Server Apache, compilato con supporto PHP e MySQL – avente la funzionalità di webserver per risolvere le richieste inoltrate al sistema tramite protocollo HTTP (<http://httpd.apache.org/> <http://www.php.net/>)
- PHP compilato come eseguibile da linea di comando – nuova funzionalità

disponibile a partire dalla versione 4.0.15 che permette di eseguire script PHP direttamente da linea di comando come per qualsiasi altro script di tipo bash, PERL, ecc. (<http://www.php.net/manual/en/install.commandline.php>)

- Python, con modulo MySQL DB di python (<http://python.org/>)
- GNU Autoconf (http://www.gnu.org/software/autoconf/)
- WML (Website META Language) – permette di generare pagine HTML personalizzabili a partire da template di pagine WML, viene usato anche per configurare l'email dell'amministratore del sistema prima che l'applicazione venga effettivamente installata (<http://www.engelschall.com/sw/wml/>)

Per l'istallazione del prodotto è possibile consultare il relativo documento, presente all'indirizzo web http://cdsware.cern.ch/download/INSTALL.

8.3 UTILIZZO DI CDSWARE AL CERN

Il gruppo di sviluppo di CDSware ha collaborato con i bibliotecari per realizzare l'interfacciamento tra Aleph e CDSware stesso.
Questo perché la biblioteca del CERN di Ginevra utilizza come sistema di gestione delle proprie risorse bibliografiche l'applicazione Aleph500, successore di Alph300, versione più evoluta dal punto di vista dell'interfaccia utente, che è grafica a differenza del suo predecessore.
In tal modo, è stato possibile far utilizzare ad Aleph l'OPAC specializzato di CDSware, dove per OPAC si intende il database catalografico; ciò nonostante Aleph disponga del suo OPAC personale.
Questo consente di far sì che i bibliotecari del CERN continuino ad utilizzare

Aleph per le proprie catalogazioni, e allo stesso tempo per le consultazioni su CDSware.
La figura seguente mostra a grandi linee la struttura logica adottata al CERN.

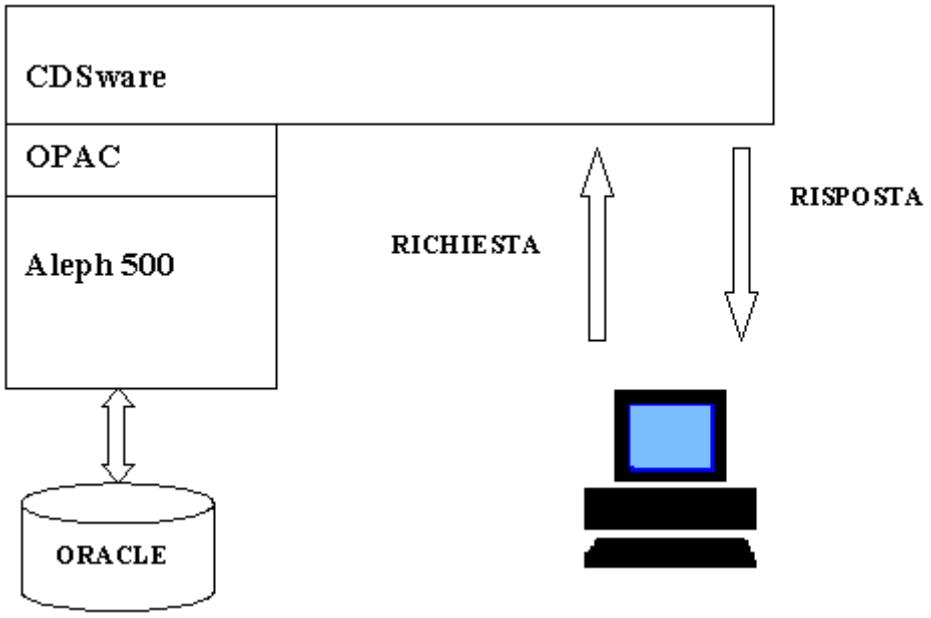

Figura 1 – Sistema Aleph 500 – CDSware adottato al CERN

Il tipo di contenuto che caratterizza le capacità di CDSware è la cosiddetta letteratura grigia, con la quale si intendono:

- Preprint – lavori di ricerca in forma preliminare la cui pubblicazione non è certa, ma che comunque, in tal modo, sono messi a disposizione della comunità
- Tesi – le tesi dei laureandi in forma elettronica

- Conferenze – testi di conferenze, interventi dei partecipanti, ecc.

Tutti questi documenti trattano le seguenti materie, presentate in ordine di impatto decrescente, per l'ambito del CERN:

- fisica delle particelle
- matematica
- astrofisica e astronomia
- informatica

In aggiunta a tali applicazioni, CDSware viene impiegato per la catalogazione dei periodici scientifici sia a stampa che elettronici, tra cui, degni di nota, sono HEP Libraries Webzine ed Helsevier.
In linea generale, l'applicazione CDSware è indipendente dalla copertura disciplinare sulla quale opera.

8.4 MODULI CDSWARE

CDSware è composto da un insieme di moduli software che realizzano le diverse funzionalità che l'applicazione offre all'utente.
Alcuni di questi offrono un interfaccia grafica oltre alla possibilità di essere "configurati" da linea di comando, mentre altri attualmente offrono solo quest'ultima possibilità.
Un elenco completo dei moduli di CDSware è il seguente:

- BibData
- BibHarvest
- BibConvert
- BibFormat

- BibUpload
- BibWords

Mentre le interfacce grafiche sono:

- WebAccess
- WebSubmit
- WebSearch
- WebPerso

Tali moduli, che verranno trattati nei paragrafi successivi, fanno riferimento all'unica versione stabile 0.0.9 finora realizzata, liberamente scaricabile all'indirizzo web http://cdsware.cern.ch/download/cdsware-0.0.9.tar.gz

8.4.1 BibData

E' il modulo funzionale di CDSware che permette di manipolare direttamente le tabelle dei dati bibliografici.
Esempi di tali manipolazioni sono le seguenti: editazione di record, modifiche di campi, semplici catalogazioni, modifiche globali, ecc..

Attualmente le funzionalità del modulo sono le seguenti:

- Edit Record
- Global Replace

La prima permette di accedere direttamente ai record del database e di editarli, ad esempio per correggere un errore di digitazione nell'URL del fulltext, oppure

aggiungere informazioni riguardanti la pubblicazione di un documento.

La seconda permette di effettuare la modifica globale di un valore, come ad esempio la correzione di un errore di digitazione nel campo nome dell'autore dei record di metadati. Tale funzionalità va utilizzata con attenzione poiché di tipo globale e quindi con effetto coinvolgente tutti i record presenti nel database.

E' da notare che tali funzionalità sono incomplete nella versione in esame e qualora la lista delle acquisizioni più recenti necessiti di essere aggiornata, bisognerà lanciare da linea di comando lo script *create_collection_pages* presente nella directory cgi-bin del sistema.

8.4.2 BibHarvest

E' il modulo funzionale di CDSware, non ancora disponibile per la versione 0.0.9, che permette di configurare l'harvester OAI per eventuali raccolte di dati periodiche in maniera automatica, specificando ad esempio cosa raccogliere, come trasformare i dati, come immagazzinarli in CDSware, ecc..

BibHarvest rappresenta uno strumento OAI compatibile che conferisce a CDSware la caratteristica di service provider in aggiunta a quella propria di data-provider, come definito nel protocollo OAI-PMH [B1].

Il suo compito principale è quello di assicurare trasferimenti di dati dai repository di metadati al database locale. Il modulo può essere configurato per raccolte periodiche di metadati in modo incrementale o può realizzare raccolte di tutti i record presenti all'esterno.

Da notare che le funzionalità previste dal modulo non sono, nella versione in esame, disponibili.

8.4.3 BibConvert

E' il modulo funzionale di CDSware che permette la conversione dei record di metadati da vari formati in un altro supportato dal database locale.
Esso è progettato per elaborare i record di metadati raccolti precedentemente, convertendoli nel formato XML del MARC 21 prima che vengano caricati nel database.

In ogni caso, BibConvert è un applicazione flessibile in quanto permette la conversione in generale da un qualsiasi formato di ingresso ad un altro XML in uscita.

Gli utilizzi più comuni per tale modulo possono essere: migrazione di dati da un sistema ad un altro, conversione di record ricevuti da diverse sorgenti di dati, acquisizioni automatiche di dati non OAI compatibili e di dati non ben strutturati, ecc..

BibConvert deve essere opportunamente configurato in accordo alle necessità dell'utente. Tale configurazione viene effettuata attraverso la creazione di un file di conversione per un particolare insieme di record di metadati espressi in uno specifico formato d'ingresso. Bisognerà quindi creare file di conversione diversi per differenti formati di record.
Tale file è di testo e la sua struttura viene analizzata in maggior dettaglio nel capitolo successivo.

Una volta preparato il file di configurazione, BibConvert convertirà il file di record di metadati di input in accordo alle corrispondenze presenti in maniera automatica.

Attualmente BibConvert può essere configurato ed utilizzato solo attraverso linea di comando. Per effettuare la conversione del file di input bisognerà usare il seguente comando:

```
# bibconvert -cconvertSBA.cfg < sampleSBA.txt> outSBA.xml
```

ottenendo infine il corrispondente file in formato XML MARC 21 di output.

Generalmente a questa fase di trasformazione segue la procedura di caricamento attraverso i moduli BibFormat, BibUpload e BibWords.

Tale modulo, oggetto della parte sperimentale di questo elaborato, verrà dettagliatamente descritto e analizzato più avanti al cap 9.

8.4.4 BibFormat

E' il modulo funzionale di CDSware che permette di specificare come presentare i dati bibliografici all'utente finale nell'interfaccia di ricerca e nelle pagine dei risultati della stessa interfaccia.
Per esempio si può decidere di presentare i titoli con un font grassetto, l'abstract in corsivo, ecc.

BibFormat non possiede solo la funzionalità di "formattatore di output", ma è anche un "costruttore automatico di link". Per esempio:

- si possono creare automaticamente dei link al sito dell'editore in funzione del nome e delle informazioni presenti nelle pagine di un giornale, basandosi su alcune regole di configurazione prestabilite
- si può creare automaticamente un link alla home page dell'autore per il campo author di un record di metadati, basandosi su alcune regole prestabilite
- si può stabilire che per certi numeri di report di documento, vengano svolte determinate azioni di formattazione

Per default l'applicazione definisce un semplice formato HTML di presentazione che mostra i campi informativi più comuni: titolo, autore, abstract, keywords , link al fulltext del documento, ecc.. In ogni caso l'utente può definire i propri formati di output per la presentazione di specifiche strutture di metadati.

BibFormat può essere eseguito in entrambe le seguenti modalità:

- Interfaccia web
- Linea di comando

L'esecuzione da interfaccia web normalmente viene preceduta dall'esecuzione dello strumento applicativo *Reformat Records*. Esso permette di aggiornare i formati dei record bibliografici immagazzinati. Normalmente prima di lanciare tale strumento è necessario configurare i *comportamenti* e i *formati* del modulo BibFormat.
Una volta effettuati questi passi preliminari, è possibile scegliere se ricostruire i formati dei record in base a collezioni selezionate oppure inserire manualmente una query di ricerca avvalendosi dell'interfaccia web che effettuerà tutti i passi di formattazione necessari.
Ad esempio è possibile effettuare una richiesta per ricostruire i formati di

collezioni di foto in HTML, o riformattare tutti i record presenti nel sistema.

L'esecuzione da linea di comando invece, avendo un file di dati in formato XML MARC 21 che deve essere caricato in CDSware, prevede di configurare BibFormat e i suoi comportamenti di output di default e successivamente di lanciare l'applicazione nel modo seguente:

```
# bibformat < outSBA.xml > outSBA_fmt.xml
```

saranno così creati i formati HTML di default che arricchiscono il file XML di ingresso.
Generalmente tale fase viene seguita dalla procedura di caricamento nel database attraverso i moduli BibUpload e BibWords.

Un'altra situazione possibile può essere quella di aggiungere nuovi formati personalizzati, ad esempio chiamati "HTML portfolio" e "HTML caption" per formattare diverse fotografie in una pagina. Consideriamo che tali formati siano già caricati nella tabella *collection_format* e che siano chiamati hp e hc.
Come caricare tali o altri formati all'interno della tabella collection_format, richiede una procedura di accesso al database particolare che attualmente, nella documentazione interna di CDSware, non è ancora specificata.

Successivamente bisogna preparare i comportamenti di output corrispondenti, HP e HC rispettivamente, che produrranno semplicemente un file XML contenente solo i tag 001 e FMT. Tale operazione non coinvolge l'aggiornamento dei record bibliografici ma solo dei loro formati. Contemporaneamente bisognerà preparare anche i formati corrispondenti e alla fine bisognerà eseguire la formattazione nel modo seguente:

```
# bibformat otype=HP,HC < outSBA.xml > outSBA_fmt.xml
```

che restituisce un file XML contenente i soli tag 001 e FMT.

Successivamente può essere effettuato il caricamento dei formati, tramite il modulo BibUpload.
A questo punto i nuovi formati dovrebbero apparire nell'interfaccia WebSearch.

Tale modulo verrà dettagliatamente descritto e analizzato nella tesi correlata di Amelotti Ercole [B1].

8.4.5 BibUpload

E' il modulo funzionale di CDSware che permette di caricare i dati bibliografici in formato XML MARC 21 nel database bibliografico di CDSware.

E' previsto, in linea teorica, che tutte le trasformazioni di dati siano realizzate attraverso i moduli BibHarvest e BibConvert attivati in maniera automatica e sequenziale. In realtà, la versione in esame di CDSware non contiene il modulo BibHarvest completo e dunque il caricamento sul database deve essere effettuato attraverso BibUpload, chiamando separatamente l'applicazione.
Il modulo attualmente non necessità di essere configurato, in quanto tutte le configurazioni sono effettuate nel modulo BibConvert.
Esso considera il tag 037 $a di ogni record, se presente, come "numero di report primario" (primary report number) che è l'identificatore unico del record nel sistema. Di conseguenza, se sono presenti due record con lo stesso valore per il campo 037 $a, il record esistente verrà sovrascritto dall'ultimo caricato.
E' previsto dagli implementatori, che ulteriori funzionalità configurabili saranno successivamente incluse.

Una volta che è stato creato un file in formato XML MARC 21 (tramite BibConvert o BibFormat) contenete i record di metadati, è possibile caricare i dati sul database bibliografico di CDSware chiamando lo script BibUpload da linea di comando, nel modo seguente:

```
# bibupload outSBA.xml
# bibupload outSBA_fmt.xml
```

Dopo che in tal modo tutti i record sono stati caricati su CDSware, è buona norma indicizzarli attraverso il modulo BibWords.

8.4.6 BibWords

E' il modulo funzionale di CDSware che permette di configurare i cosiddetti *indici di parole*, cioè definire quali campi bibliografici sono indicizzati. Tali indici vengono così usati dall'interfaccia di ricerca dell'applicazione.

Questa parte dell'applicazione è particolarmente utile nei casi in cui vengono utilizzate strutture dati non standard, per configurare i propri indici. Ad esempio si può dire al programma che l'indice autore è costruito a partire dai tag bibliografici 100 $a e 700 $a.
CDSware, comunque, definisce i file di indice più comunemente utilizzati, quali ad esempio: autore, titolo, abstract, keywords, ecc..

Nella seguente tabella sono presenti alcune corrispondenze tra le parole utilizzate per l'indicizzazione e i tag MARC 21 corrispondenti, definite automaticamente da CDSware:

Parole d'indice	Provenienza parole	Tag MARC 21 indicizzati
Global	Tutti i campi nel record	All
Collection	Indicatori di collezione	980 $a
Abstract	Abstract	520
Author	Nomi di autore	100, 700
Keyword	Parole chiave	6531 $a
Reference	Riferimenti	909C5
Reportnumber	Numeri di report	037 $a, 088 $a, 909C0 $r
Title	Titoli	245, 246
Fulltext	Tutte le parole dai fulltext	[*.pdf,*.ps,*.doc,*.ppt]

Tale funzionalità di indicizzazione, nella versione in esame, non è ancora inclusa come interfaccia web, è comunque possibile utilizzarla tramite linea di comando editando manualmente le tabelle *wordsindex* e *wordsindex_field*.

Per l'utilizzo dell'applicazione, possibile solo dall'interfaccia di linea di comando, la sintassi generale per indicizzare una o più parole, di uno o più record, è la seguente:

```
bibwords <add|del> <lim_inf> [-lim_sup] [indice1,indice2,...]
```

dove gli argomenti obbligatori sono:

- **<add|del> -** aggiunge o elimina uno o più indici dal/i record specificato/i
- **<lim_inf>** - specifica il limite inferiore, ossia il numero del record, dal quale iniziare l'indicizzazione (eventualmente l'unico)

mentre gli argomenti opzionali sono:

- **[-lim_sup]** – specifica il limite superiore, ossia il numero del record, dell'intervallo interessato dall'indicizzazione
- **[indice1,indice2,…]** – specifica la/e parola/e da utilizzare come indice/i per il o i record specificato/i

Per esempio, se si desidera indicizzare le parole dal record numero 12, bisognerà lanciare:

```
# bibwords add 12
```

in tal modo tutti campi del record, corrispondenti a quelli specificati nelle tabelle del database, saranno rintracciabili attraverso l'interfaccia web di ricerca.

Altri esempi sono i seguenti:

```
# bibwords add 234-250
# bibwords add 1-25000 author,keyword
# bibwords del 340
```

Da notare che BibWords lavora direttamente sui dati precedentemente

immagazzinati nei database bibliografici di CDSware.

8.4.7 WebAccess

E' l'interfaccia di CDSware che permette di definire chi possiede i diritti di accesso e/o di amministrazione ai vari moduli dell'applicazione.
Ad esempio si può stabilire che Aldo è il gestore dei dati bibliografici, Giovanni può modificare le pagine di ricerca e Giacomo è il redattore per l'approvazione delle sottomissioni.

8.4.8 WebSearch

E' l'interfaccia di CDSware, che permette di configurare le impostazioni di ricerca per le varie collezioni di dati.
Ad esempio è possibile definire l'intestazione, il piè di pagina e i box di portale (sinistra, destra, sopra, sotto) per la pagina di ricerca, per descrivere le diverse collezioni di dati. Si possono anche definire opzioni di ricerca, campi di ricerca, esempi di ricerca, ecc..

8.4.9 WebSubmit

E' l'interfaccia di CDSware che permette le impostazioni di sottomissione e la logica per vari tipi di dati.
Ad esempio è possibile definire i campi di metadati per i vari tipi di dati, stabilire cosa fare dei valori inseriti prima che siano effettivamente caricati, effettuare il peer review, stabilire strategie di approvazione, ecc..

8.4.10 WebPerso

E' l'interfaccia di CDSware, non ancora inclusa, che permetterà di gestire gli account degli utenti e le loro differenti caratteristiche di personalizzazione, come basket definiti dall'utente e query di alert.
Ad esempio, sarà possibile abilitare o disabilitare la funzionalità di alerting, cancellare tabelle utente e di sessione, modificare i basket utente, ecc..

BIBLIOGRAFIA

[B1] Amelotti Ercole, *Protocollo OAI-PMH negli Open Archive e applicazione CDSware per la rappresentazione dei relativi dati bibliografici*, tesi di laurea in informatica, Università degli Studi di Messina, A.A. 2003-2004 (relatore Puccio L., correlatore De Robbio A.).

SITOGRAFIA

[S19] http://eprints.rclis.org/archive/00000031/