Integrating CDS/ISIS Databases with Greenstone Digital Library Software (GSDL)

Francis Jayakanth, B.S. Shivaram, K Venkatalakshmi[#] and Sukhdev Singh[*]
(franc, shivaram, lakshmivk)@ncsi.iisc.ernet.in)
sukhi@hub.nic.in
[#] National Centre for Science Information
Indian Institute of Science
Bangalore – 560 012
[*]Bibliographic Informatics Division
(Indian Medlars Division)
National Informatics Centre
A-Block, CGO Complex, Lodhi Road
New Delhi – 110 003

Abstract

CDS/ISIS is an advanced non-numerical information storage and retrieval software developed by UNESCO since 1985 to satisfy the need expressed by many institutions, especially in developing countries, to be able to streamline their information processing activities by using modern (and relatively inexpensive) technologies [1]. CDS/ISIS is available for MS-DOS, Windows and Unix operating system platforms. The formatting language of CDS/ISIS is one of its several strengths. It is not only used for formatting records for display but is also used for creating customized indexes. CDS/ISIS by itself does not facilitate in publishing its databases on the Internet nor does it facilitate in publishing on CD-ROMs. However, numbers of open source tools are now available, which enables in publishing CDS/ISIS databases on the Internet and also on CD-ROMs.

In this paper, we have discussed the ways and means of integrating CDS/ISIS databases with GSDL, an open source digital library (DL) software.

Introduction: Almost all information related activities in an enterprise/organization are increasingly driven by database, Internet and web technologies. These technologies offer both challenges and opportunities for developing effective mechanism for information dissemination. The database technology has been in existence for almost 4 to 5 decades now. For example, CDS/ISIS, a database management system, specially designed for handling bibliographical information, was developed in the 1960s. It has many powerful and unique features. Many organizations across the world are using this package for decades now. However, the search interface of CDS/ISIS though powerful, is not very user friendly. Also, CDS/ISIS does not facilitate in publishing its databases on the web nor does it facilitate in publishing its databases on CD-ROMs. With the availability more and more open source software tools for building DLs, it is now possible to integrate such legacy database systems with DL software. GSDL is one such open source software suite.

About CDS/ISIS: CDS/ISIS is a database management system specially developed for building and managing bibliographic databases. It is available for different operating system platforms like MS-DOS, Windows, and Unix. The formatting language of CDS/ISIS facilitates in defining precise formatting requirements for the purpose of displaying or printing of records onto a file. For example, the format can specify, which of the fields are to be displayed, the required indentation and the literals to be prefixed for the field values for the sake of clarity. It is quite possible to display or print the records in any desired format. This feature is particularly useful if the CDS/ISIS databases are to be ported to other RDBMS or software packages. The formatting language features are also used for generating customized indexes.

About GSDL: GSDL is an open source, multilingual software suite meant for building, managing, distributing, and provide access to digital collections. It was developed by the New Zealand Digital Library Project (NZDL) at the university of waikato[2]. It is distributed in cooperation with UNESCO and Human info NGO [3]. It is available for different operating system platforms, and is easy to install and use. The input collection can be in varied file formats like HTML, PDF, Word (Doc and RTF), PS, e-mail, bibliographic, etc. GSDL uses the concept of plugins to parse the input collection. Plugins for popular file formats like HTML, PDF, Word (Doc and RTF), e-mail etc are bundled with the software package. GSDL supports both full text and field level searching apart from flexible browsing facilities. Greenstone Librarian Interface (GLI), available from version 2.40 onwards, makes the collection building process much easier.

Why the CDS/ISIS – GSDL Integration? CDS/ISIS is an excellent tool for building and managing bibliographic databases. The databases, depending on the targeted users, may reside on standalone systems or in a networked environment. However, if the CDS/ISIS databases are to be published on the web or if they are to be distributed on CD-ROMs, CDS/ISIS facilitates neither. Other software tools like WWWISIS, WWW-ISIS, IQUERY etc. (some are freeware and some are shareware) do facilitate in publishing CDS/ISIS databases on the web. GENESIS, a WWWISIS based software tool, facilitates integration of CDS/ISIS database on the web. A version of GENISIS facilitates in publishing CDS/ISIS databases on CD-ROMs also. WWWISIS is by far the most popular utility for publishing CDS/ISIS databases on the web [3]. However, WWWISIS is not a free software beyond version 3.x.

GSDL, an open source software suite meant for building and distributing digital library collections, can serve the dual purpose of publishing CDS/ISIS databases on the web as well as on CD-ROMs. The user interfaces of the web and the CDROM versions are same. So, the main reason behind porting of CDS/ISIS databases onto GSDL is to publish CDS/ISIS databases on the web [4}. Also, if you need to distribute your CDS/ISIS databases on the CD-ROMs, it can be done so using GSDL.

GSDL uses the concept of plugins to work with the documents for which DL collection has to be built. Plugins are the programs, which parses the documents and extracts the metadata from them. For example, the HTML plugin (HTMLplug.pm) converts HTML documents into Greenstone Archive Format (GAF) and also extracts metadata, which are explicit in the documents, such as title, enclosed in <title> tag. Similarly, plugins are available for other document formats like PDF, MS-Word, BibTex, etc. Depending on the heterogeneity of input collection, different plugins are used to build a collection. The input documents, which comprise the collection, should conform to the formats, which the plugins expect. Failing to adhere to this requirement will result in unexpected hits while browsing or searching.

At least two approaches are possible in porting CDS/ISIS databases onto GSDL. Both these approaches are based on exporting CDS/ISIS database records in specific formats. The exported records are then used to build the GSDL collection. In the first approach the CDS/ISIS records are exported in HTML format. A sample record in HTML format is
given below.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html><head>
<meta name=Title content=Techniques for the measurement of transpiration of individual
plants></meta>
<meta name="Creator" content="Magalhaes, A.C."</meta>
<meta name="Creator" content=" Franco, C.M.>"</meta>
<meta name="Keywords" content=" plant physiology"</meta>
<meta name="Keywords" content=" plant transpiration"</meta>
<meta name="Keywords" content=" measurement and instruments>"</meta>
<title> Techniques for the measurement of transpiration of individual plants </title></head>
<body>
Title: Techniques for the measurement of transpiration of individual plants
<br>Creator: Magalhaes, A.C.; Franco, C.M.
Keywords: Paper on: plant physiology; plant transpiration; measurement and instruments
</body></html>
ENDOFRECORD
```

To build a GSDL collection for the records exported in HTML format, following two steps are involved. Please note that we have used records from 'cds' database, a sample CDS/ISIS database, which comes along with CDS/ISIS package, to illustrate our work.

Step1: HTMLPlug is used by GSDL to parse and extract required data from the HTML documents. This plugin by default builds an index for the <title> tag and creates a fulltext index for rest of the content. If additional indexes are required for other fields then all such fields should be included in the <meta> tag in the <head> section of HTML document. In the sample record shown above Title, Creator, and Keywords are included in the <meta> tag, which implies that indexes will be built for

Title, Creator and Keywords. Having indexes for different fields will facilitate in performing field specific searching as well as searching across the fields, which is a very desirable feature, especially when the number of records in a database is substantially high.

CDS/ISIS database records printed in the above format will all be contained in a single file. Since HTMLPlugin works at file level, each CDS/ISIS record should be written onto a separate file. In order to do this, in every record the literal 'ENDOFRECORD' is introduced through the display format so that it becomes easier in the next step to extract each record and be written as a separate file. Any other technique could as well be used to differentiate one record from the other. The main idea is that the program, which parses the file containing CDS/ISIS databases records, should be able to distinguish individual records.

Step 2: Write a program in any high-level programming language to parse the file containing CDS/ISIS database records in HTML format. The program should extract individual records and write them in separate files. We have written a short PERL program for this purpose. The program creates the files, file0.htm, file1.htm, ..., filen.htm, depending on the total number of records contained in the input file. These individual files are then used by GSDL to build the GSDL collection of CDS/ISIS database records. The name of this collection is 'cds'.

The second and the easier approach of integrating CDS/ISIS database with GSDL is to write a CDS/ISIS display format, which extracts the database records in a format, which conforms to the format the 'RefPlug' expects. This plugin is almost same as the 'ReferPlug' except for minor changes have been incorporated for the purpose of proper display. ReferPlug is used for processing bibliographic files in 'refer' format. A sample record for ReferPlugin is given below.

```
%A Magalhaes, A.C.; Franco, C.M.
%T Techniques for the measurement of transpiration of individual plants
%K Paper on: plant physiology; plant transpiration; measurement and instruments
```

In the above record:

- %A corresponds to Author field,
- %T corresponds to Title field, and
- %K corresponds to Keywords field

Similarly, other fields are mapped to the corresponding alphabets preceded by % sign.
To use the RefPlugin to process CDS/ISIS records, write a CDS/ISIS display format to generate the records in the above format and print the records onto a file. You should ensure that the display format inserts a blank line between the records in order to differentiate one record from the other. The output file thus generated can be straight away used to build the GSDL collection. There is no need to split the records into separate files because RefPlug works at record level. A couple of screen shots of collection built using this approach are given below. The name of this collection is 'bib'.

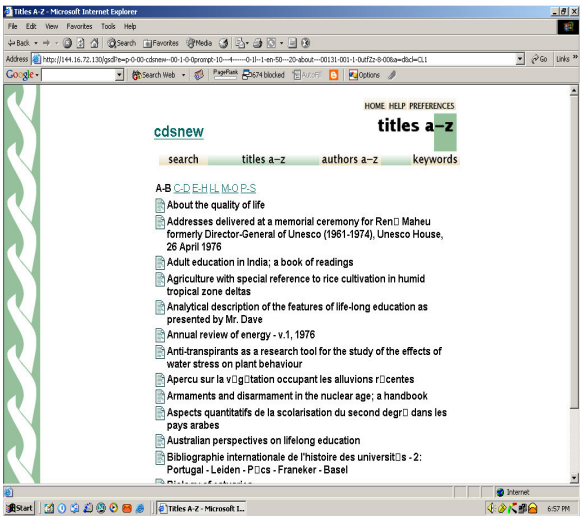
GSDL Collection Building Process: GSDL, by default, uses MG (Managing Gigabytes) as the compressing and indexing system. MG provides document level index and compression of source documents. In other words, MG does not facilitate field level searching, i.e. limiting the search to say 'Title' field only, is not possible using MG. Hence searching across the fields is also not possible. MG++ or MGPP, is a re-implementation of MG [5]. It enhances the functionality of MG by facilitating proximity, field level and across the fields searching. It comes bundled with the GSDL package. Since GSDL uses MG as the default indexing system, it should be configured through its collection configuration file to use MGPP as the indexing system instead of MG. Every collection that is being built will have its collection configuration file namely, 'collect.cfg' It is through this file that all the desired features for a collection can be expressed and achieved. The 'collect.cfg' file used for building the 'cdsnew' collection is given below.

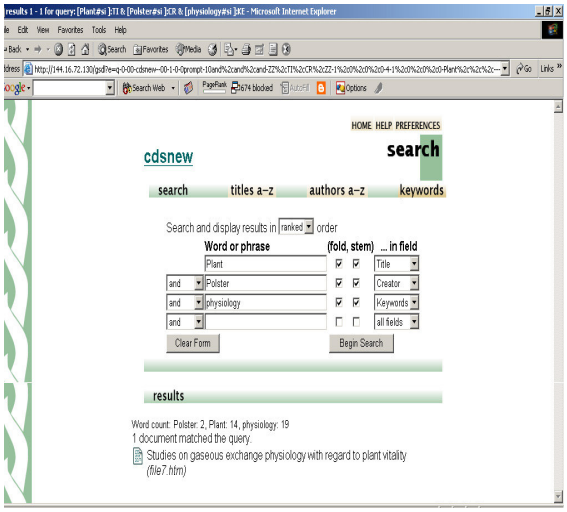
```
creator       franc@ncsi
maintainer    franc@ncsi
public        true
buildtype     mgpp
searchtype    form plain
 indexes      "allfields,Title,Creator,Keywords"
 plugin       ZIPPlug
plugin        GAPlug
plugin        TEXTPlug
plugin        HTMLPlug -metadata_fields Title,Creator,Keywords
plugin        ArcPlug
classify      AZList -metadata Title
classify      AZList -metadata Creator
Classify      AZList -metadata Keywords
collectionmeta collectionname    "cdsnew"
collectionmeta iconcollection    ""
collectionmeta collectionextra   ""

collectionmeta ".allfields,Title,Creator,Keywords" "documents"
```

Collection Configuration File Used for 'cdsnew' Collection

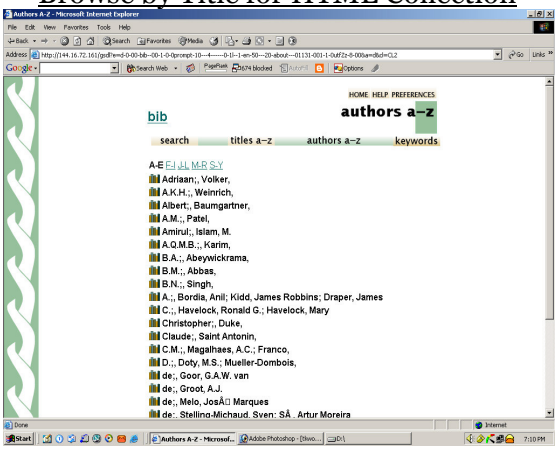
The collection thus built using MGPP will facilitate in field level searching and searching across the fields. A few screen shots are given below to illustrate this.

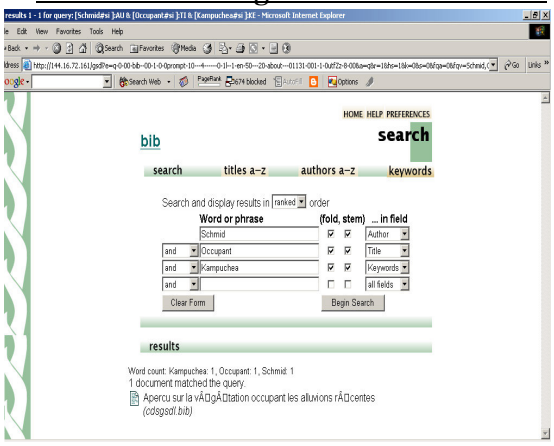




Browse by Title for HTML Collection

Fielded Searching for HTML Collection





Browse by Author for 'bib' collection

Fielded Search for 'bib' Collection

Conclusion: Though CDS/ISIS is an excellent tool for managing bibliographic databases there is no provision in the system either to publish the database on the web or on CD-ROM. One has to depend on other software tools to achieve either or both these features. Using GSDL it is quite possible to achieve the objectives of publishing a collection on web and/or on CD-ROM. It doesn't require great deal of effort in porting CDS/ISIS databases onto GSDL. An easier approach in porting the CDS/ISIS database onto the GSDL is by having a plugin for CDS/ISIS databases. In fact the latest release of GSDL, version 2.41 is bundled with one such plugin. Due to want of time we haven't been able to test this feature yet. Hence there is no mention about the usage of CDS/ISIS plugin in this paper.

References

1. http://portal.unesco.org/
2. http://www.greenstone.org/cgi-bin/library
3. A Tutorial on Integrating CDS/ISIS Databases with the World Wide Web, Francis Jayakanth, S Jayashree, Information Today and Tomorrow, Vol. 18, Iss. No. 4, Oct – Dec 1999 (http://dsir.nic.in/division/nissat/nisnat/news/itt9904.html)
4. http://dlist.sir.arizona.edu/archive/00000187/
5. MGPP: A Search Engine for XML Documents, Katherine Don, Dept. of Computer Science, University Waikato, Hamilton, New Zealand (http://www.greenstone.org/docs/mgpp_user.pdf)