Publishing a CDS/ISIS Database in GSDL

(Revised/Corrected Version)

[By SUKHDEV SINGH (http://www.geocities.com/esukhdev) Email: esukhdev@hotmail.com dated 21st November 2002.]

WHY:

ICMR-NIC Centre for Biomedical Information has developed a Bibliographic database of Indian Biomedical Journals known as IndMED. It is available on Internet at http://indmed.nic.in. It is served from HTTPD server using PERL scripts from ISIS files (Singh, 2001).

We normally demonstrate this database during various User Awareness Programmes at users' sites. For this we had to ensure that Internet connectivity is available at such locations. In case, connectivity is not available or unreliable, we needed some alternative to demonstrate the database. Carrying a Laptop with HTTPD and PERL configured is one alternative. Other better and convenient option is to have the database on a CDROM. However, such a CDROM should be self-contained and should not require any setup or special software at the users' machines. Moreover it should protect the database against download in its native file format from the CDROM. Greenstone Digital Library software has the ability to produce such a CDROM.

HOW:

STEP 1.

Create Print format in ISIS to generate complete valid HTML document for each record.

a) Each record should start with '<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">' string to define the Doc Type.

b) It should produce other required html tags e.g. /'<html>'/'<head>'/

c) GSDL does full indexing. If you want to generate additional indexes to support queries that can be qualified i.e. restricted to various elements of the records (say title, keywords etc.) then you need to create HTML META TAGs for these elements. These META TAGs would be used by GSDL to create indexes. I preferred to duplicate these elements. So that full text as well restricted searching could be possible e.g.

```
'<meta name="Title" content="',if p(v200) then v200 else 'No Title' fi,'">'/
'<meta name="Creator" content="',(v300+|, |),'">'/
'<meta name="JSource" content="',v201,". "v440"; ",v490,"("v491")",":
"v492,'">'/
'<meta name="Date" content="',if val(v493)>100 then v493 else v440*0.4 fi,'">'/
'<meta name="Subject" content="',(v620+|, |),'">'/'<title>',if p(v200) then v200
else 'No Title' fi,
```

d) Create a Unique End of Record Marker by any unique string that your database is not having as data.
e.g.
'ENDOFRECORDENDOFRECORDENDOFRECORD'/

The format used by me is given ahead.

```
'<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">'/'<html>'/
'<head>'/
'<meta name="Title" content="',if p(v200) then v200 else 'No Title' fi,'">'/
'<meta name="Creator" content="',(v300+|, |),'">'/
'<meta name="JSource" content="',v201,". "v440"; ",v490,"("v491")",": "v492,'">'/
'<meta name="Date" content="',if val(v493)>100 then v493 else v440*0.4 fi,'">'/
'<meta name="Subject" content="',(v620+|, |),'">'/'<title>',
if p(v200) then v200 else 'No Title' fi,
'</title>'/'</head>'/'<body>'/'<table>'/
'<tr><td>'/
'<font color="#FF00FF">'/(v300+|; |)'</font>
</td></tr>'/
'<tr><td>'/v330/'</tr></td>'/
'<tr bgcolor="#FFFFCC"><td>'/v200/'</tr></td>'/
'<tr><td><i>'/v201,". "v440"; ",v490,"("v491")",": "v492/'</i></td></tr>'/
'<tr bgcolor="#c0c0c0"><td><HR></td></tr>'/
'<tr><td>'/"ABSTRACT: "v600/'</td></tr>'/
if p(v620) then '<tr><td>'/'KEYWORDS:</td></tr>'/
'<tr><td>'/(v620+|; |)/'</td></tr>'/,fi,/
if p(v621) then '<tr><td>'/'OTHER KEYWORDS:</td></tr>'/
'<tr><td>'/(v621+|; |)/'</td></tr>'/,fi,
if p(v703) then
'<tr><td>'/'References: ',v703,'</td></tr>'/,fi,
if p(v1) then '<tr><td>'/'Record Identifier: ',v1/'</td></tr>'/,fi,/
'</table>'/'</body></html>'/
'ENDOFRECORDENDOFRECORDENDOFRECORD'/
```

STEP 2.

Take a printout using this "HTML GENERATING" format to a text file. I took printout of entire IndMED database having about 22,000 records.

Here is the Sample:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<head>
<meta name="Title" content="Association of vitiligo with Lichen planus">
<meta name="Creator" content="Martis J, Bhat MR">
<meta name="Source" content="Indian Journal of Dermatology. 2002 Apr - Jun; 47(
2): 129">
<meta name="Date" content="2002">
<meta name="Subject" content="Vitiligo/DI, Lichen Planus/DI, Skin Diseases,
Autoimmunity, Face/PA, Case Report, Human, Adult: Female">
<title>Association of vitiligo with Lichen planus</title>
</head>
<body>
<table>
<tr><td>
<font color="#FF00FF">
Martis J; Bhat MR</font></td></tr>
<tr><td>
Department of Dermatology, Venereology and Leprosy, Father Muller's Medical
Collage, Mangalore - 575 002, Karnataka.
</tr></td>
<tr bgcolor="#FFFFCC"><td>
Association of vitiligo with Lichen planus
</tr></td>
<tr><td><i>
Indian Journal of Dermatology. 2002 Apr - Jun; 47(2): 129
</i></td></tr>
<tr bgcolor="#c0c0c0"><td><HR></td></tr>
<tr><td>
</td></tr>
<tr><td>
KEYWORDS:</td></tr>
<tr><td>
Vitiligo/DI; Lichen Planus/DI; Skin Diseases; Autoimmunity; Face/PA; Case
Report; Human; Adult: Female
</td></tr>
```

```
<tr><td>
References: 4</td></tr>
<tr><td>
Record Identifier: NI208097
</td></tr>
</table>
</body></html>
ENDOFRECORDENDOFRECORDENDOFRECORD
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<head>
<meta name="Title" content="Air pollution and health [editorial]">
<meta name="Creator" content="Chhabra SK.">
<meta name="Source" content="The Indian Journal of Chest Diseases Allied
Science. 2002 Jan-Mar; 44(1): 9-11">
<meta name="Date" content="2002">
<meta name="Subject" content="Air Pollution, coronary Disease, Smoking,
Regression Analysis, Risk Factors, Questionnaires, Spirometry, Human">
<title>Air pollution and health [editorial]</title>
</head>
<body>
<table>
<tr><td>
<font color="#FF00FF">
Chhabra SK.</font></td></tr>
<tr><td>
Department of Cardiorespiratory Physiology, clinical Research Centre,
Vallabhbhai Patel Chest Institute University of Delhi-110007
</tr></td>
<tr bgcolor="#FFFFCC"><td>
Air pollution and health [editorial]
</tr></td>
<tr><td><i>
The Indian Journal of Chest Diseases Allied Science. 2002 Jan-Mar; 44(1): 9-11
</i></td></tr>
<tr bgcolor="#c0c0c0"><td><HR></td></tr>
<tr><td>
</td></tr>
<tr><td>
KEYWORDS:</td></tr>
<tr><td>
Air Pollution; coronary Disease; Smoking; Regression Analysis; Risk Factors;
Questionnaires; Spirometry; Human
</td></tr>
<tr><td>
References: 9</td></tr>
<tr><td>
Record Identifier: NI205715
</td></tr>
</table>
</body></html>
ENDOFRECORDENDOFRECORDENDOFRECORD
```

STEP 3.

Now write a simple PERL script (or any other language you like) that would "CHOP" the printout file into html files using the "End of Record" string. I used the following script to create html files each having a single cds/isis record.

```
#! /usr/bin/perl
#print "Hello\n"; #To Test if the srcipt is working.
open (FH,dmt) ; #dmt is the file output from ISIS
@lines=<FH>;
$alines="@lines";
@records = split (/ENDOFRECORDENDOFRECORDENDOFRECORD/,$alines);
$total=@records;
#$total=20; #To Test whether it works by creating only 20 files.
print "Total Records: $total\n\n";
for ($xx=0; $xx < $total ; $xx++)
	{
		print "File $xx being generated\n";
		open (WFH, ">ind".$xx.'.html');
		$precord = shift @records;
		print WFH "$precord\n";
		close (WFH);
	};
```

IMPORTANT TIP: This technique can be used to publish your entire database over Internet and made searchable by search engines like GOOGLE!!!!!!.

STEP 4.

Now leave aside ISIS and come to GSDL. Create a new collection. I created INDMED by using the following command:
mkcol.pl -creator dukhi@hotmail.com indmed.

STEP 5.

Transfer the html files created by you in the "import" directory.

STEP 6.
Now edit your "collect.cfg" in the "etc" directory. Give names of the indexes you want to generate. Moreover remember to give suitable arguments to the HTMLPlug to extract the meta data for the indexes from the HTML META TAGs. I have used the following collect.cfg.

```
creator      dukhi@hotmail.com
maintainer   dukhi@hotmail.com
public       true

indexes      document:text document:Title document:JSource document:Creator document:Date document:Subject
defaultindex  document:text

plugin       ZIPPlug
plugin       GAPlug
plugin       HTMLPlug -metadata_fields Title,JSource,Creator,Date,Subject -input_encoding iso_8859_1

plugin       ArcPlug
plugin       RecPlug

collectionmeta collectionname    "indmed"
collectionmeta iconcollection    "_httpprefix_/collect/indmed/images/indmed.jpg"
collectionmeta collectionextra   "IndMED: A Bibliographic Database of Indian Biomedical Research"
collectionmeta .document:text    "All Text"
collectionmeta .document:Title   "Titles"
collectionmeta .document:JSource  "Journal Sources"
collectionmeta .document:Creator  "Authors"
collectionmeta .document:Date  "Years"
collectionmeta .document:Subject  "Keywords"
format       DocumentHeading ""
format       VList   "<td>[link][icon][/link] - [link][Title][/link]</td>"
format       SearchVList "<td valign=top> [link][icon][/link][Title][/link]</td>"
format       DocumentButtons ""
format       DocumentText    "[Text]"
```

STEP 7.

Now import the documents into the GSDL archive format using the import.pl. You may first test before import large number files using the argument "-maxdocs". I used the following command.

import.pl -OIDtype incremental -removeold -groupsize 100 indmed

STEP 8.

Simple, Build the collect now using the buildcol.pl.
I used the following:

buildcol.pl indmed

STEP 9.

Rest of the things you know. I used the following:

rm index/*
mv building/* index