WORKING WITH DIGITAL INFORMATION USING WWWISIS ON LINUX

By

A.R.D. Prasad*

ABSTRACT

The paper attempts to discuss the various issues in using wwwisis software on a Linux platform, where the database can include digitized information like photographs, audio and video files.

* Documentation Research and Training Centre, Indian Statistical Institute, Bangalore.
E-mail : ardprasad@hotmail.com

0. Introduction

The web technology originally was meant to display static web pages. With the advent of database connectivity, it possible to generate web pages dynamically. That is, the data is stored in a database and on the fly an HTML file is generated with data retrieved from the database. Any changes to the database will automatically be reflected in the output of the web pages. The CDS/ISIS package is the most commonly used information retrieval software and there have been many attempts to achieve CDS/ISIS database connectivity to the web servers. The outcome of one such attempt is ‘wwwisis’. This possibility can give rise to the use of many features of web technology. The browsers, as they are client side programmes can display data from a web server without bothering about the machine on which the web server is hosted. In a way, platform independence is achieved. Though some multimedia formats require extra plug-ins, many browsers have built-in capability to display digital information appearing in various multimedia formats. The present paper attempts to describe a system using CDS/ISIS for Linux, Apache Web server and wwwisis software.

1. Setting up CDS/ISIS on Linux

Originally the CDS/ISIS for Unix platform was developed for SCO-Unix. However, on the Intel machines with Linux operating system, the same software can be used with a little tinkering. Linux is provided with Intel Binary Compatibility Specification (iBCS). It means that the binary files produced on Intel platforms under any flavour of Unix can be run under Linux. However, as this is not by default, one should run the following command:

```
Insmod iBCS
```

With the above command the CDS/ISIS does not run once the system is rebooted. To make the above command run at the boot time, it may be inserted in the following file

```
/etc/rc.d/rc.local
```

so that whenever the system is on, iBCS runs and consequently the CDS/ISIS programme works. Run 'isis' to test whether it is working or not. Also make sure that the 'isis' programme is in your PATH and is executable. You may require to change the contents of SYSPAR.PAR file.

NOTE: All the CDS/ISIS files are in capital letters, all Unix flavours are case sensitive unlike DOS, WINDOWS-95.

2. Multimedia Files

The multimedia files can appear in various file formats. The audio files may be in 'wav', 'mp3' etc formats, whereas the photographs may be in 'gif' or 'jpg' format and the video files may be in 'mpg' or 'avi' format. These multimedia files are normally converted from analogue format to digitized format. Though in the past the database management systems were not tuned to processes digital information, now a days most of the major database software can handle digital information. However, it should be noted that the digital information is not normally stored in the database itself. On the other hand, they will be kept as separate files with usual filenames and a reference is made in the database record. The latest version of CDS/ISIS i.e WINISIS (It appears that there is no further development on plain text based CDS/ISIS after 3.08) includes some features that can handle digital information. However, the present approach is taken independent of such capability. In fact, the Unix version of CDS/ISIS belongs to plain text mode. The solution is through web interface. As most of the browser software like Netscape Navigator and Internet Explorer can handle multimedia files, we can use the strengths of the browser to handle digital information.

3. Steps in Entering Digital Information

Step –1: Presume that we want to display the cover page of the documents in our database (I am sure nobody wants to do this). To capture the cover pages, we normally use a scanner to capture the image and store it as either 'gif' file or 'jpg' file.

Step – 2: Modify the FDT file of an existing database to add a field to hold the name of the file containing digital data. In the present paper, tag number '10' is used to hold this information. We can also add a subfield to hold the caption of the digital information

In the input worksheet, the relevant entry may look like the following

```
Image (v10): ^abook-1.gif^bCover page
```

The subfield ^a will contain the name of the image file, whereas the subfield ^b contains the caption that need to be displayed along with the image.

NOTE: The image files are normally kept in '/var/www/icons' directory or any other directory under '/var/www'. The field '10' can be a repeatable field, so that one can display any number of images if one wishes to.

4. Apache Server

Apacahe is the most widely used and robust web server on the Internet. Fortunately, it is free of cost and is bundled with Linux. The HTTP daemon should be run in order to activate Apache server. The following procedure helps to do that:

1) Run "setup" command as super user (i.e. login as root)
2) In the "setup" choose the 'system services'
3) In the 'system services' go to "httpd" and mark it. To mark, press 'space bar' and a '*' appears to show that http daemon starts at boot time. Next time the system is booted automatically and the web server is activated
4) Quit setup

To test your web server -- Open a browser (in Linux, you may use Netscape Navigator) and enter something similar (not same) as the following

1) "httpd://drtc.isibang.ac.in"
2) Or: httpd://202.54.37.89
3) Or: httpd://127.0.0.1

The 1st option should be used if you know your alphabetical IP address, the second if you do not have an alphabetical address. The third option works if you do not have either. If everything works well, one should see the welcome page of the Apache Server, which is nothing but the following file in RedHat Linux 7.0 (earlier version of RedHat Linux had this file is in /home/www)

```
/var/www/index.html
```

The above file should be replaced by your home page.

5. Installing WWWISIS

The WWWISIS and its related programmes can be downloaded from the following site:

ftp.bireme.br/wwwisis/pc/linux

The programmes include:

wwwisis: the main programme to run the web interface
loadiso.sh: the shell script to generate master and cross reference file form the ISO format file of the given data base.
fullinv.sh: the shell script to generate inverted files
ifload used by the fullinv.sh file
mx used by loadiso.sh and fullinv.sh

All the above programmes may be kept in /var/www/cgi-bin directory, the only essential one is ‘wwwisis’.

Steps to generate database files:

Use CDS/ISIS to export data to be captured in a file called ‘MST.ISO’ and issue the following commands.

```
cp MST.ISO /var/www/cgi-bin/mst.iso
cp CDS.FST /var/www/cgi-bin/cds.fst
cp CDS.STW /var/www/cgi-bin/cds.stw
cp CDS.PFT /var/www/cgi-bin/cds.pft
cd /var/www/cgi-bin
./loadiso mst cds        ( this generates cds.mst and cds.xrf from mst.iso)
./fullinv.sh cds         ( this should generate the inverted files, but it does not)
```

NOTE: There is a problem with the sort programme of RedHat Linux 7.0, as it attempts to sort character by character instead of word by word. This creates problem and the inverted file generation is aborted. If you have RedHat Linux 6.0, copy the sort programme and use it. In addition, also note that the file names are generally in small letters, although Unix CDS/ISIS uses all the file names in capital letters.

To test whether loadiso.sh has generated the master and cross-reference files, give the following command

```
wwwisis db=cds from=1 to=5 pft=@cds.pft
```

The above command should display first 5 records. This should work even without the generation of inverted index files.

To test that the inverted files are properly generated, give the following command

```
wwwisis db=cds bool=”plant” pft=@cds.pft
```

The above command should display the records having the key word ‘plant’.

6. Testing WWWISIS

To test that the 'WWWISIS' programme is working, run the following commands at shell prompt

```
wwwisis hello
wwwisis menu=1
```

However, to test the CGI (Common Gateway Interface) to web server

Run Netsape Navigator
Enter the URL as: http://127.0.0.1/cgi-bin/wwwisis [hello].

NOTE: you can use server IP address in numeric or alphabetical form instead of 127.0.0.1

The above procedure should result in displaying a web page with 'hello'. You can also try with '[menu=1]' to see that menu '1' is displayed.

If everything works well, we are ready to develop CGI programmes.

7. Basic Concepts

The Common Gateway Interface (CGI) works as an intermediary between the browser and the web server (in our case Apache web server). The requests from the browser to the server may be sent through CGI scripts. There are many CGI scripting languages like Perl, Tcl etc. and even the Unix shell. Not only these CGI scripts forward the requests from the browser to the web server, they retrieve data from the server and send them to the browser in HTML format. These CGI programmes are normally kept in 'cgi-bin' directory.

Briefly, the CGI programmes

1. Should collect request from the browser
2. And send data back to the browser in HTML format

The most common method of collecting data from the browser is to use the HTML tag 'FORM'. The FORM tag contains another element called 'ACTION' where we can specify the action (i.e. the programme) to be invoked. A brief syntax of the 'ACTION' tag is given below

```
<html>
<body>
   …
   …
<FORM ACTION=http://127.0.0.1/cgi-bin/search.sh METHOD=”POST”>
   …
   …
</body>
</html>
```

NOTE: For a complete explanation of HTML tags, one should refer any book on HTML.

8. Steps in Setting Up WWWISIS Interface

1. Create an HTML file which serves as the first interface. If we call this file as ‘index.html’, it should be placed in ‘/var/www/html/index.html’. (Refer Appendix –1)
2. Create a programme which is invoked from the above file. If we call this programme as ‘search.sh’, it should be used with ‘ACTION’ of ‘FORM’ tag in the index.html file and should be in the ‘/var/www/cgi-bin’ directory. (Refer Appendix – 2)
3. Create ‘search.cgi’ file to present the various options for the ‘wwwisis’ command used in ‘search.sh’ file. The ‘search.cgi’ in turn refers to various other files. This file also should be in the /var/www/cgi-bin directory. (Refer Appendix – 3)
4. Copy ‘cds.fst’ file to /var/www/cgi-bin’ (Refer Appendix – 3)
5. Create cds.txt file (Refer Appendix – 3)
6. Create head.pft file (Refer Appendix – 3)
7. Create cds.pft file (Refer Appendix – 3)
8. Create tail.pft file (Refer Appendix – 3)

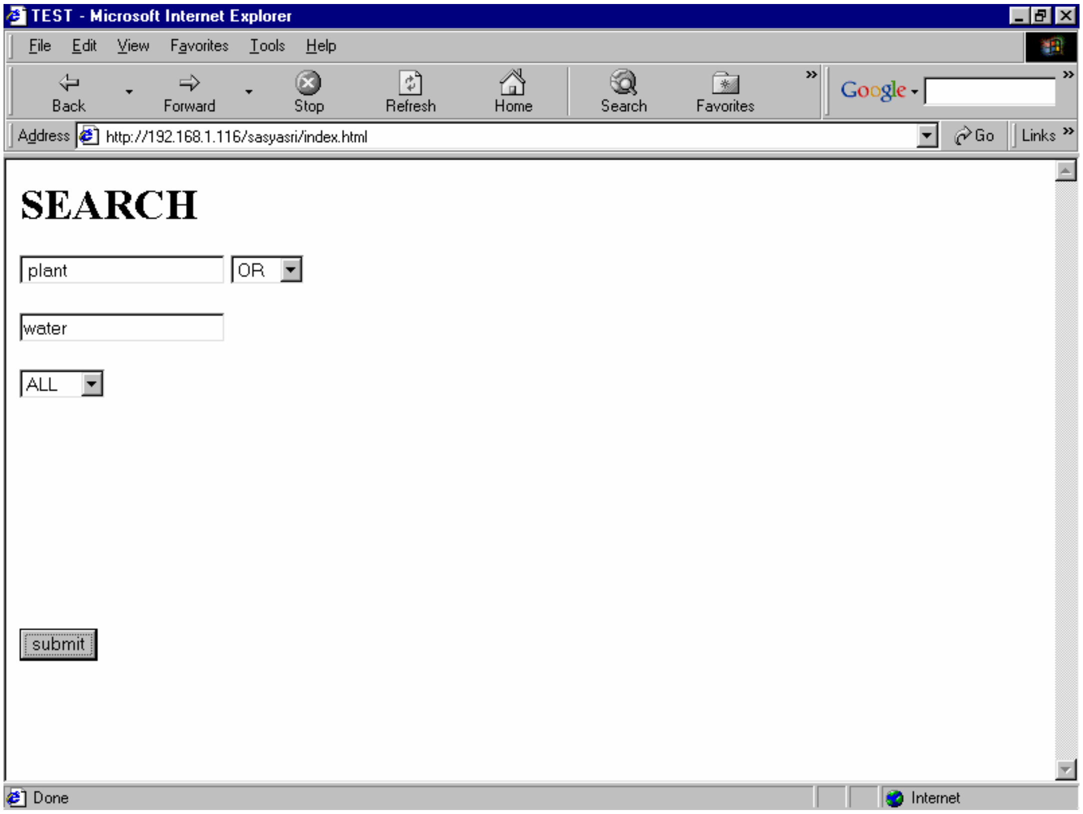
All the above files should be in /var/www/cgi-bin, except the ‘index.html’ file.

Now open the browser like Netscape Navigator and the URL as

http://127.0.0.1

NOTE: you cannot simply open the same file by entering the URL as /var/www/html/index.html as this cannot establish connection to the web server. This approach can only be used to view the disk files in html format. Please also note that 127.0.0.1 can be used only if you want to access from the same machine where the web server is installed. However, to access the web page from some other machine, you should give the IP address of the machine having the web server, either in alphabetical or numeric form such as ‘http://www.drtc.isibang.ac.in’ or ‘http://202.54.37.89’.

If everything is alright the following first page should be displayed.



TEST - Microsoft Internet Explorer

http://192.168.1.116/sasyasri/index.html

SEARCH

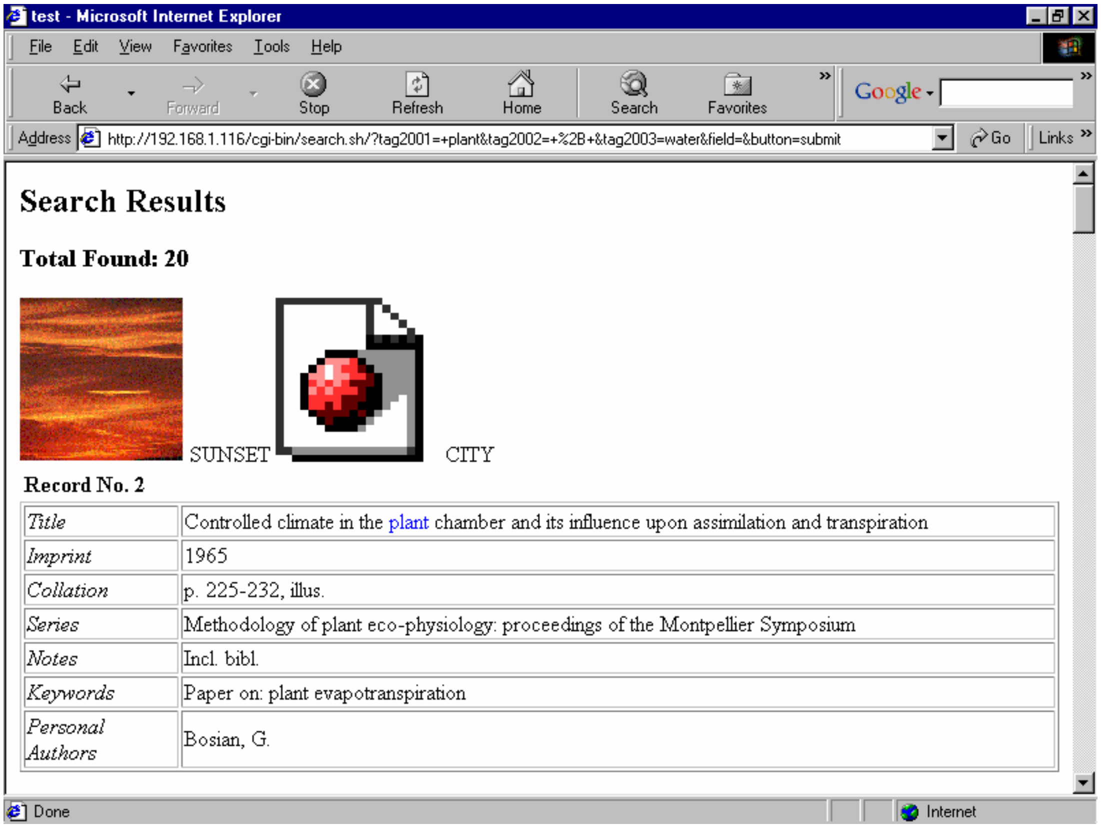
plant OR

water

ALL

Fig. 1 Search Interface

In the above screen one can enter the keyword or keywords. The second entry in the above page displays only ‘OR’, but by pulling down the menu one can find the other Boolean operators like ‘AND’ and ‘NOT’. The fourth entry displays only ‘ALL’, but by pulling down the menu, items like ‘TITLE’ and ‘ATHOR’ will appear. This is to limit the search either to AUTHOR field or to TITLE field. By default, the system will search in all the indexed fields. Once the search terms are filled in, if the user presses ‘submit’, the system displays the next web page with all the bell and whistles. The following output contains some irrelevant images, as they are used only for demonstration purpose.


test - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
Google
Address
http://192.168.1.116/cgi-bin/search.sh/?tag2001=+plant&tag2002=+%2B+&tag2003=water&field=&button=submit
Go
Links
SUNSET
CITY
Done
Internet

Search Results

Total Found: 20

Record No. 2

Title	Controlled climate in the plant chamber and its influence upon assimilation and transpiration
Imprint	1965
Collation	p. 225-232, illus.
Series	Methodology of plant eco-physiology: proceedings of the Montpellier Symposium
Notes	Incl. bibl.
Keywords	Paper on: plant evapotranspiration
Personal Authors	Bosian, G.

Fig. 2 Search Results

9. Conclusion

The WWWISIS has many more facilities. The complete commands can be obtained from the manual which includes the complete listing of the menus, commands and tags of the virtual records. The software along with the documentation can be downloaded from the ftp site: ftp.bireme.br/wwwisis/pc/linux

10. References

1. wwwisis: A World-Wide Web Server for ISIS Database Version 3.0, (ftp://ftp.bireme.br/winisis/cds.zip)

2. CDS/ISIS for windows: Reference Manual: Version 1. (ftp://ftp.bireme.br/wwwisis/doc/wwwisis12.doc)

APPENDIX – 1

Step 1: Creation of the first page

This step is involved in the creation of the first file i.e. index.html. If this file is in /var/www/html directory, one can invoke this file from the browser by entering the following URL

http://127.0.0.1 (if you know the IP address you can enter here)

The complete listing of the 'index.html' file is given below.

Listing of index.html

```
<html>

<head>
<title>TEST</title>
</head>

<body>

<h1>OPAC OF THE LIBRARY</h1>

<form method="get" action="/cgi-bin/search.sh/">
<p><input type="text" name="tag2001" value=" " size="20">

  <select name="tag2002" size="1">
    <option value=" * ">AND</option>
    <option value=" + ">OR</option>
    <option value="~">NOT</option>
  </select></p>
  <p><input type="text" name="tag2003" value=" " size="20">
</p>
  <p><select name="field" size="1">
    <option value="">ALL</option>
    <option value="24">Title</option>
    <option value="70">Author</option>
  </select></p>
  <p>&nbsp;</p>
  <p>&nbsp;</p>
  <p>&nbsp;</p>
  <p>&nbsp;</p>
  <p><input type="submit" name="button" value="submit">
</p>

</form>
```

```
</body>
</html>
```

In the above HTML file, we have mentioned the ‘search.sh’ file as the action to be taken once the “submit” button is pressed. We press ‘submit’ button once and we enter the search elements. The “input” tag of HTML describes that the input is “text” TYPE, and the “VALUE” is presently blank and once the value is entered, it will be captured in the variable called ‘tag2001’. Similarly, the next lines in the file describe the input for the tag2002, tag2003 and ‘field’. That is, the tag2001 captures the input of the first key word, tag2002 captures the choice of Boolean operator like ‘and’, ‘or’, ‘not’; tag2003 captures the next keyword and ‘field’ captures the information whether the keywords should be searched in all fields or only in title or author fields.

In other words, the FORM tag of HTML passes the information you have entered once the ‘submit’ button is pressed.

APPENDIX – 2

STEP – 2: Creation of Search.sh File

This is a Unix shell script and should contain the CGI scripting. The content of the file may look like the following:

LISTING OF search.sh

```
#!/bin/sh
echo "Content-type: text/html"
echo ""
./wwwisis pfxtag=tag cgi=@search.cgi
exit 0
```

In the above programme we have used various parameters to wwwisis programme. The ‘pfxtag’ parameter is meant that the variables ‘tag2001’, ‘tag2002’, ‘tag2003’ should be treated as tags v2001, v2002, v2003 in the virtual record of CDS/ISIS. Here, it should be noted that the most interesting part of wwwisis is that it generates virtual records out of each record generated from CDS/ISIS database. The present paper uses the CDS database for the sake of convenience. We know that the CDS database contains fields like 24 for Author; 70 for Title etc. In addition, each record will contain fields 2001, 2002, 2003 with the information/data, we have entered. The ‘pfxtag’ indicates that we are using the word ‘tag’ as prefix to the actual tag. It is a good idea to use tags greater than 2000 for the simple reason that they will not conflict with the database tags as the data base tags mostly contain 3 digits and the tags greater than 1000 and less than 2000 are used by wwwisis to hold various other kinds of information, like number of records retrieved (tag 1002), the Boolean expression (tag 1021) etc.

APENDIX – 3

STEP – 3: Creation of Search.cgi

The above programme ‘search.sh’ in turn refers to another CGI file called ‘search.cgi’. This is only a matter of convenience and for the sake of clarity. The contents of ‘search.cgi’ are given below

LISTING OF search.cgi

```
'db=cds'/
'bool=',
   (
   if p(v2001) then v2001, fi
   if p(v2002) then v2002, fi
   if p(v2003) then v2003, fi
   )/,
'hlbool=@cds.fst'/
'hltext=@cds.txt'/
'prolog=@head.pft'/
'pft=@cds.pft'/
'epilog=@tail.pft'/
```

In the above programme:

The ‘db=cds’ states that the database to be used is ‘cds’.

The ‘bool=’ states that the Boolean search expression is to be prepared from the contents of tag v2001, v2002, v2003.

The ‘hlbool=’ and ‘hltext=’ are meant to highlight the search terms in the output.

The ‘prolog=’ states that the output HTML file should contain the contents of ‘head.pft’ as the first few statements of HTML file.

The ‘pft=’ states that ‘cds.pft’ file is to used for the actual display of the records

The ‘epilog=’ states that the ‘tail.pft’ should be appended to ‘cds.pft’ file

LISTING OF cds.fst

The is the ‘fst’ file taken from /isis/data direcotry

```
70 0 MHU,(V70/)
24 4 MHU,V24
69 2 V69
```

LISTING OF cds.txt

This file tells that in the retrieved records, the key terms entered as search terms should be displayed in blue colour

```
70 0 '^1<FONT COLOR=BLUE>^2</FONT>'
24 0 '^1<FONT COLOR=BLUE>^2</FONT>'
69 0 '^1<FONT COLOR=BLUE>^2</FONT>'
```

LISTING OF head.pft

The following lines will be prefixing the cds.pft

```
'<html>'/
'<title>test</title>'/
'<body>'/
'<h1>'/
'Search Results'/
'</h1>'/
```

Notes on cds.pft

This is the main display format with the necessary HTML tags. In the following file,

if val(v1001) = 1 then 'Total Found: ', v1002,fi/

displays the number of records retrieved for the present query by displaying the tag v1002.

Whereas, the following code displays the image file along with its captions. The image files are kept in /var/www/icons. The subfield '^a' contains the name of the image file and the subfield contains the caption of the image. This code is meant for the first occurrence of the field v10, hence v10[1].

```
if p(v10[1]) then
   '<TR><TD> <img src="/icons/',v10[1]^a,'" width="200"
height="300" alt="city.gif (13864 bytes)>"</TD>',
   '<TD> v10[1]^b</TD></TR>',/
fi,/
```

Listing of cds.pft

```
'<p></p>'
```

```
if val(v1001) = 1 then 'Total Found: ', v1002,fi/

if p(v10[1]) then
   '<TR><TD> <img src="/icons/',v10[1]^a,'" width="200"
height="300" alt="city.gif (13864 bytes)>"</TD>',
   '<TD> v10[1]^b</TD></TR>',/
fi,/
if p(v10[2]) then
   '<TR><TD> <img src="/icons/',v10[2]^a,'" width="200"
height="300" alt="city.gif (13864 bytes)"></TD>',
   '<TD> v10[2]^b </TD></TR>',/
fi,/

mhl,'<TABLE WIDTH="100%" BORDER=0>'
'<TR><TD WIDTH="100%"><strong>Record No.
',mfn(1),'</strong></TD></TR></TABLE>'
'<TABLE WIDTH="100%" BORDER>'
if p(v12) then '<TR><TD WIDTH="15%"><I>Conference main
entry </I></TD><TD>',v12,'</TD></TR>'/fi,
if p(v24) then '<TR><TD WIDTH="15%"><I>Title
</I></TD><TD>',mpl,v24,mhl'</TD></TR>'/fi,
if p(v25) then '<TR><TD WIDTH="15%"><I>Edition
</I></TD><TD>',v25,'</TD></TR>'/fi,
if p(v26) then '<TR><TD WIDTH="15%"><I>Imprint
</I></TD><TD>',v26,'</TD></TR>'/fi,
if p(v30) then '<TR><TD WIDTH="15%"><I>Collation
</I></TD><TD>',v30,'</TD></TR>'/fi,
if p(v44) then '<TR><TD WIDTH="15%"><I>Series
</I></TD><TD>',v44+|; |'</TD></TR>',fi,/
if p(v50) then '<TR><TD WIDTH="15%"><I>Notes
</I></TD><TD>',v50,'</TD></TR>'/fi,
if p(v69) then '<TR><TD WIDTH="15%"><I>Keywords
</I></TD><TD>',v69,'</TD></TR>'/fi,
if p(v70) then '<TR><TD WIDTH="15%"><I>Personal Authors
</I></TD><TD>',v70+|; |'</TD></TR>',fi,/
if p(v71) then '<TR><TD WIDTH="15%"><I>Corporate Bodies
</I></TD><TD>',v71+|; |'</TD></TR>',fi,/
if p(v72) then '<TR><TD WIDTH="15%"><I>Meetings
</I></TD><TD>',v72+|; |'</TD></TR>',fi,/
if p(v74) then '<TR><TD WIDTH="15%"><I>Added Title
</I></TD><TD>',v74+|; |'</TD></TR>',fi,/
if p(v76) then '<TR><TD WIDTH="15%"><I>Other language
titles </I></TD><TD>',v76+|; |'</TD></TR>',fi,/
'</TABLE><P>'
```

LISTING of tail.pft

```
if (v1091 = '7') then 'No ----Records retrieved'fi,/
if (v1091 = '0') then 'END OF SEARCH RESULTS', fi,/
```