DRAFT COPY - FIRST ITERATION
WILL BE MODIFIED AFTER COMMENTS RECEIVED
DO NOT USE FOR PRODUCTION WORK

Relational Integrity/Test Ingestion Procedure, v0.01 13May2003

(based on Valerie Henderson's Relational Integrity Procedure v5 2/4/2003)

Introduction

Relational Integrity and Test Ingestion is a process whereby a data engineer takes a previously validated volume supplied to him or her and perform a variety of tests to ensure database compliance in preparation for the volume being added to the Planetary Data Systems (PDS) online search database. The information to be added to the database consist of catalog files only. Specifically, the root level VOLDESC.CAT file and all of the *.CAT files in the /CATALOG/ root level sub-directory. Neither actual data, documentation, nor any other type of file, other than *.CAT files are ingested into the database.

The data engineer will not actually add the catalog files to the database and instead add the files to their own personal copy of the latest, official database. John Ho, the PDS database engineer will perform the actual database update based on the results of the data engineers test ingestion. By performing test on the data engineers personal database, the official search database will not be corrupted. Only after a successful test ingestion and validation of relational database integrity within the data engineer's personal database, will the catalog files be added to the official database by John Ho.

The procedure is not overly complicated, once the data engineer understands the process and the reasons for performing each step. There will be frustrations in determining the cause of errors at each step. There are many PDS data engineers more than willing to help in determining causes of errors. Among these are Valerie Henderson, Colleen Schroeder, Ron Joyner, Tyler Brown, and Steven Adams. If none of these can help or determine the error, they will direct the data engineer to the best appropriate person. As a last resort, John Ho may be consulted.

The general procedure to follow in this process is as follows. First, as explained in Step 1, below is to return the data engineer's own personal database to the last known valid official database. In the process of test ingestion and as a data engineer manipulates his or her own copy of the official database, errors could have been introduced into the database. Hence, it is imperative that a test ingestion be commenced with a valid, known copy of the database. Explained below is how to determine the latest version of the database and two different procedures to refresh the data engineer's database. Neither procedure is better than the other and is a matter of convenience to the data engineer as the results of both procedures are the same.

After the data engineer's copy of the database mirrors the official database, a test ingestion of the catalog files are performed using a web interface. This procedure is explained in Step 2 below. The PDS computer, PDSPROTO, is used for this test ingestion step and it is necessary for the catalog files to reside on the PDSPROTO disk drives. These disk drives can be network mounted on your desktop personal computer. It will most likely be necessary to perform editing upon the catalog files before they will be successfully ingested. Having the PDSPROTO disk drive network mounted on your computer will allow a simple "drag and drop" of the files between PDSPROTO and your local computer editing software.

The next step in the process, explained in Steps 3-7 below, is to perform a relational integrity check between your modified database and the official online database. A relational integrity check invokes database rules which validate the entity and referential integrity of the catalog files the data engineer is ingesting. These steps will be performed on the PDS computer CADENZA for steps 3-5 and on the data engineer's desktop computer for steps 6 and 7.

Finally, upon successful test ingestion and validation of relational integrity, an e-mail is sent to John Ho detailing the location of the files to be ingested into the official database as well as a brief explanatory text of what is being ingested. Also included should be the location of report files generated by the test ingestion and relational integrity checks.

Useful programs to have for this procedure.

The following programs have proven their usefulness by other data engineers in performing the following procedures. These are Intel CPU based programs. There are equivalent programs avail for Macintosh computers.

- SQL Advantage (database program)
- WS-FTP (file transfer program utilizing network connections)
- UltraEdit-32 (extremely powerful text editor, (not a word processor), useful for computer platform file conversion, general text editing, file comparison, etc.)
- ExamDiff Pro (file comparison utility)

(DISCLAIMER: The Jet Propulsion Laboratory, The California Institute of Technology, and The National Aeronautics and Space Administration does not recommend or endorse any commercial product mentioned in this document.)

Step 0. Verifying catalog files to process have not already been ingested.

(Note: This step has been labeled "Step 0" so that subsequent steps below follow the same numbering sequence as the original "cookbook" step-by-step list developed by Valerie Henderson and attached as a supplement at the end of this document.)

It is first necessary to determine if the catalog files supplied to the data engineer have already been ingested into the master database, or to determine if they have been updated or revised and need to be re-ingested.

All new catalog files now contain a keyword at the top of the catalog files named "LABEL_REVISION_NOTE". This keywork will most likely be the second element in a catalog label and will include the author name and a date indicating the version of the file. By comparing the dates within the LABEL_REVISION_NOTE between the newly supplied catalog file and the database catalog file, it will be easy to determine if the new file is of a later version. If the new catalog file is newer, it will be ingested and overwrite the existing database file. The steps below detail how to view a catalog files that has already been ingested.

A. Go to the following internet web site using the browser of your choice.

http://pdsproto.jpl.nasa.gov/onlinecatalog/top.cfm

(This is an important web page in your PDS work and will be useful to bookmark as the contents of these web pages will be referred to often.)

B. In the left-hand frame displayed, there is a "Category" list of items that a user can perform a search upon. Depending on the type of catalog file to be ingested, select the appropriate radio button and then left mouse click on the "Show List" button at the bottom of the list.

The large right-hand frame will display the results of Step B above. The user can either scroll down the list or enter a text string in the "Element (wildcard) Search" box below the list of previously ingested items. If using the search box, either press the computer "return" key or left mouse click on the "Submit" button below the fill-in box to execute the search. (It is not necessary to enter the traditional DOS wildcard character "*" within the search string. It is assumed.)

Select the item being review by highlighting the item using the left mouse button. Then left mouse click on the "EMAIL template" button above the displayed list. There are nine buttons in a 3x3 grid, the "EMAIL Template" button is the second button in the left-hand column.

The ingested catalog file will be displayed within the same frame. The data engineer can then visually inspect the file, or highlight the file with the computer mouse and cut-and-paste the selection into a text editor and perform a file comparison between the existing ingested file and the new, supplied catalog file.

If the file already exists in the database, and has not been updated or modified, it is unnecessary to re-ingest the file and may be deleted from consideration.

Step 1. Refresh your personal database.

The first step is to bring your personal database to the last known valid and stable state, i.e. to mirror the current catalog database generated and maintained by John Ho. You will need to know the current version number of the database. The database name takes the form of "pdscat1rxx", where "xx" is the version number. As of 13 May 2003, the current version is 39 (pdscat1r39). The three steps below explain how to find out what the current database version is.

A. Go to the following internet web site using the browser of your choice.

http://pdsproto.jpl.nasa.gov/onlinecatalog/top.cfm

(This is an important web page in your PDS work and will be useful to bookmark as the contents of these web pages will be referred to often.)

B. Select the "Data Dictionary" radio button (second option from the bottom in the "Category" list) in the left frame and then click the "Show List" button at the bottom of that same section.

C. After the right-hand frame updates, the current database version name will be displayed at the bottom of the frame below the horizontal rule. The filename will be "pdscat1rxx_ri.rpt", where "xx" represents the version number.

There are two different ways to refresh your personal database. One is to log onto CADENZA and enter the database commands detailed below. The second method can only be used if you have SQL Advantage installed on your personal desktop computer.

1. Using CADENZA

Log onto CADENZA.JPL.NASA.GOV computer using your username and password. Type the following commands at the user prompt exactly as shown, (except for italicized words which indicate your own personal information.

```
> isql –U your_userid –P your_password
1> use master
2> go
1> load database your_database_name from “/sybase/pdscat1r39_dump”
2> go
          (there will be quite a bit of program output to your CRT at this step and will run
          for a few minutes)
1> online database your_database_name
2> go
Database ‘your_database_name’ is now online
1> use your_database_name
2> go
1> exit
```

You can stay logged onto CADENZA since it will be used in subsequent steps detailed below.

2. Using SQL Advantage installed on your personal desktop computer

   Start execution of SQL Advantage by clicking the desktop icon or selecting the program from your computer’s Start Menu.

   a. Select “Server” from the pulldown menu bar and select the highlighted “Connect…” option.

   b. Enter the appropriate information in the “Connect” popup window.

Server:	enter “schema”
Login:	enter your userid name
Password:	enter your userid password
Client Host Name:	enter your desktop computer name

   c. After successful logon, enter the following two commands in a session window. (“xx” in the first command below indicates the latest version number of the official database as determined in Step 1C above.)

```
load database your_database_name from “/Sybase/pdscat1rxx_dump”
online database your_database_name
```

   d. Exit from SQL Advantage using standard Windows procedures.

Step 2. Test Ingestion

This step is performed using a web browser pointing to a html page residing on the PDS computer, PDSPROTO. The catalog files to be ingested need to be moved to a subdirectory on PDSPROTO. It is possible to have the PDSPROTO disk drive network mounted on the data engineer's desktop computer. This makes it a relatively simple task to move the catalog files from a supplied CD-ROM, DVD-ROM, or other device to the PDSPROTO disk drive. Alex Leung or Francisco Loaiza, the PDS system administrators, can help get the PDSPROTO disk drive network mounted on the data engineer's desktop computer.

A. Move catalog files to be ingested to data engineer's sub-directory on PDSPROTO.

B. Execute browser of choice and navigate to URL detailed above in Step 1A, (http://pdsproto.jpl.nasa.gov/onlinecatalog/top.cfm).

C. The largest frame displayed shows nine buttons in a 3x3 grid. Select the lower right button, "IMPORT Template".

D. A new web page will be displayed in the same frame. Enter the appropriate information (first name, last name, PDSPROTO login password) in the fill-in boxes displayed and left click with a mouse on the "Login" button displayed below the fill-in boxes.

Two new frames will now be displayed where the large frame was. The new left frame allows the data engineer to enter appropriate information, as explained below and the right frame will display output results of processing steps.

Click your browsers "reload" button. In the left-hand frame, above the "Category" list will appear "Database nnn DB", where nnn is the name of your own personal database. This indicates that your personal database is loaded and available for updating, as explained in the steps below. When you selected your brower's "reload" button, the two new frame pages were replaced by the previous single frame. Click the "IMPORT Template" button again to reload the two new frames. The "Database nnn DB" will still appear at the head of the first frame.

E. Several environment variables need to be changed before processing catalog files. In the new left frame, the top fill-in box is the path to the catalog files to be ingested. Since the box is not wide enough to display the fill pathname, the data engineer needs to left click anywhere within the box, to make it active, and use the left arrow key to move the cursor to the appropriate place within the path where changes will be made. The default path displayed will be:

D:\WWW\ONLINETEMPLATE\data_engineer's_last_name

where "data_engineer's_last_name" is the appropriate name used in step 2D above. Move the cursor to the end of the path and type in the appropriate sub-directory where the catalog files to be ingested reside. After entering the appropriate pathname, either press "Enter" or left mouse click on the "Update Path" button to the left. The path entered will be reflected in the second-from-bottom cell below where the path name was entered. Also, the second cell below the path name entered will change and reflect the number of catalog files within that sub-directory.

The cells within this frame are as below:

Button:	Cell Content:
Update Path	Fully qualified pathname of files to process.
Update Ext	Name of files to process, may be wildcard or full filename.
GetDir Stats	Number of catalog files within sub-directory.
Process Files	Status, press "Process Files" to execute tool.
Delete Files	Don't use this option!
Generate ALLREFS	Not used!
Update Results File	Filename output report is written to.
File Path & Extension:	Fully qualified pathname and filename(s) as entered in "Update Path" and "Update Ext" above.
Results File:	Fully qualified pathname and filename for output report as entered in "Update Results File" above.

F. Click "Process Files" button in new left frame to submit files for processing. Depending on the number of files and machine load, this step could take a few seconds to minutes. The results will be display in the new right-hand frame as well as written to the "Results File" as detailed above. Upon completion of processing, the "Process Files" Status cell will change from "waiting" to "x files processed", where "x" is the number of files processed.

It is at this stage where the output result file need to be analyzed and any existing errors need to be eliminated before proceeding further. This is a iterative process where the output report, hopefully, reveals the error, or most likely, give a hint as to the location of any errors. The text editor of choice is utilized in changing the appropriate catalog file before another attempt at a test ingestion.

After the initial ingestion attempt using all catalog files in the directory, it is easier and faster to process each catalog file containing errors individually by changing the "Update Ext" fill-in box to reflect the single catalog file being ingested. This will allow the computer to run much faster as it will not re-process valid catalog files nor re-process files with known errors, which were revealed in the first ingestion attempt. Before re-running another "Process Files" iteration, it would be wise to either change the filename of the results output file (as to not overwrite the existing results file which contains errors of all files processed) or rename the original results output file. One suggestion would to be to rename "results.out" to, perhaps, "results_all.out". Then the "results_all.out" file could be referred to as each catalog file was modified and reprocessed. After successful ingestion of all catalog files, re-run the test ingestion with all catalog files to create a new, "clean" "results.out" file, which will need to be sent to John Ho upon completion of a successful test ingestion and relational integrity check.

Step 3. Run ddri (relational integrity check)

The relational integrity check is performed utilizing both the PDS computer CADENZA and the data engineer's desktop computer. The next few steps will compare the system relational integrity report generated as part of the last systemwide database built with one the data engineer creates that includes the recently added catalog files performed in the steps above. The process of performing a relational integrity check and resolving any introduced errors are performed in this "Step 3" as well as steps 4-7 below. Step numbering has been maintained to conform to the same numbering sequence as the original "cookbook" step-by-step list developed by Valerie Henderson and attached as a supplement at the end of this document. Depending on the CPU load when the ddri command is submitted, execution will take approximately ten minutes.

A. Download the current systemwide relational integrity (ri) report (pdscat1rXX_ri.rpt, where "XX" is the current version number) to the data engineer's current working directory on pdsproto or local desktop computer. The file can be found at the same location as explained in Step 1A through Step 1C above. Depending on the configuration of the data engineer's web browser, it may be necessary to either left mouse click on the pdscat1rXX_ri.rpt which will display a standard Windows "Save As..." popup window or right mouse click on the file and select the "Save Link As..." or "Save Target As..." option. The pdscat1rXX_ri.rpt file is a simple ASCII text file.

B. Using the PDS computer, CADENZA, execute the ddri program with the following command:

```
>/usr/local/bin/ddri -s -r -d databasename -P password
```

where *databasename* is the name of the data engineer's personal copy of the systemwide, official database which was modified in the above steps and *password* is the data engineer's password.

Two new files will be created in the data engineer's working directory from which the ddri program was executed. One of these is *database*_ri.rpt (where *database* is the name of the data engineer's personal copy of the systemwide, official database) which is the data engineer's new relational integrity report. The other is *database*_stdval.dat (where *database* is the same name as explained above) and contains new standard values introduced as a result of the data engineer's test ingestion performed in the steps above. This new standard value file will usually not be a very large file. As a check when performing the steps below, it is advisable to perform a "ls -al" UNIX command on CADENZA to determine the actual size. It is also advisable to execute the UNIX "mv" command to create an additional "backup" copy of the standard value file as the file will be modified in the steps below.

Step 4. Execute block copy (bcp) command on CADENZA

The block copy command will load the data engineer's new standard values created in the previous step into the data engineer's copy of the system database. This step is in preparation to executing the relational integrity (ddri) check again, as detailed in Step 3 above. Execute the block copy by executing the following command:

```
>bcp "database..ddcolstdval" in database_stdval.dat -f ddcolstdval.fmt -P password
```

where *database* is the name of the data engineer's personal copy of the systemwide, official database and *password* is the data engineer's password. The file, "ddcolstdval.fmt" can be found in John Ho's directory on CADENZA. Either copy this file to a local directory or prepend the fully qualified pathname to the above command.

Since this command does not change variables from execution to execution and from ingestion to ingestion, it is helpful to create a UNIX script command which can be recalled as needed. Using a text editor of choice, create a one-line file with the above command, save as a filename of choice (i.e. block_copy.script), and change file property to become an executable command (i.e. chmod +x *filename*, where *filename* is the name of the file).

Step 5. Execute relational integrity check (ddri) again.

Execute the same command as detailed in Step 3B above. This time, the *database*_stdval.dat file that is created should have a size of zero bytes.

Step 6. Transfer *database*_ri.rpt to directory contianing nri2.exe

Using WS-FTP or file transfer program of choice, move *database*_ri.rpt created in the above steps from CADENZA to either the data engineer's working directory on PDSPROTO or the desktop computer's directory that contains the DOS program "nri2.exe".

Step 7. Execute nri2.exe

This step compares the relational integrity report created during the last systemwide database build with the relational integrity report created in the previous steps as a result of the test ingestion of the data engineer's catalog files. Four files are needed to perform this step, "nri2.exe", "nri-keys.inp", pdscat1rXX_ri.rpt, and *database*_ri.rpt. These first two files may be obtained from John Ho or any data engineer who has previously performed data ingestion. The third file can be found as detailed in Step 3A above and the fourth files was created in Step 5 above.

It is imperative to ensure that *database*_ri.rpt and pdscat1rXX_ri.rpt are DOS format files. If not, nri2.exe execute, but the output file generated will be invalid and unusable. The files can be converted before transfer from CADENZA using the UNIX utility "unix2dos" or by using the conversion command within UltraEdit-32. Execute the following command in either the working directory on PDSPROTO or the local desktop directory where the above files reside:

```
nri2 pdscat1rXX_ri.rpt database_ri.rpt -b+
```

Successful execution of this program will create a file named *database*_ri.lst which will look very similar to both of the input *_ri.rpt reports. The difference is a highlight of the differences between the two input reports and will be designated with either "(-)", "(b)", or "(+)". Relational integrity conflicts that have been removed by the ingestion of the data engineer's catalog files will be designated with the hyphen notation. Relational integrity conflicts that were pre-existing before the addition of the data engineer's catalog files ingestion are designated with the "b" notation. Conflicts introduced into the database as a result of the catalog files ingestion and designated with the plus sign and need to be resolved before completion of the test ingestion. If the conflict cannot be removed, and there is a valid reason for such conflict, inform John Ho at the time of submission of the catalog files to him and explain the reason for the conflict.

Step 8. Notification of location of files for inclusion into database build.

The last step in the process is to transfer the catalog files to be ingested, the RESULTS.OUT file created in step 2 above, the database_ri.rpt report created in step 5 above, and the database_ri.lst report created in step 7 above into a directory on PDSPROTO and inform John Ho as well as the appropriate personnel that the files are ready to be ingested.

This notification should be in the form of an e-mail indicating the location of the files as well as a brief description of project/mission/instrument/etc. that the catalog files describe as well as any residual errors that cannot be removed or are valid errors introduced as a part of the test ingestion. Personnel notified should include the data engineer's immediate project supervisor, the project data engineer, and any other central node personnel involved in the creation of the volume being ingested.