21ST Century Tools For Archivists:

A Progress Report on Persistent Archives at the Stanford Linear Accelerator Center (SLAC)

Jean Marie Deken
Archivist

Overview

- SLAC Test Collection: SLD Records
- SLAC Metadata development
- PAWN Workflow: SLAC → NARA
- Some conclusions / issues for the future

SLAC Test Collection: SLD

- **S**LAC **L**arge **D**etector for the SLAC Linear Collider (SLC) 1983-1988
- Early and prolific user of world-wide web
- No further need to keep data confidential
- Many types of electronic documents
- Meet US Department of Energy/NARA criteria for retention

SLAC Test Collection: SLD

Initial Electronic Records Appraisal

- Surfed SLD web site – "manual crawl" http://www-sld.slac.stanford.edu/sldwww/sld-working.html
- Generated list of records series – technical appraisal http://www.slac.stanford.edu/history/sldpat.htm
- Interviewed collaboration's
  - Data Czar
  - Web manager
  - Spokesperson

SLAC Test Collection: SLD

Targeted Web Crawls

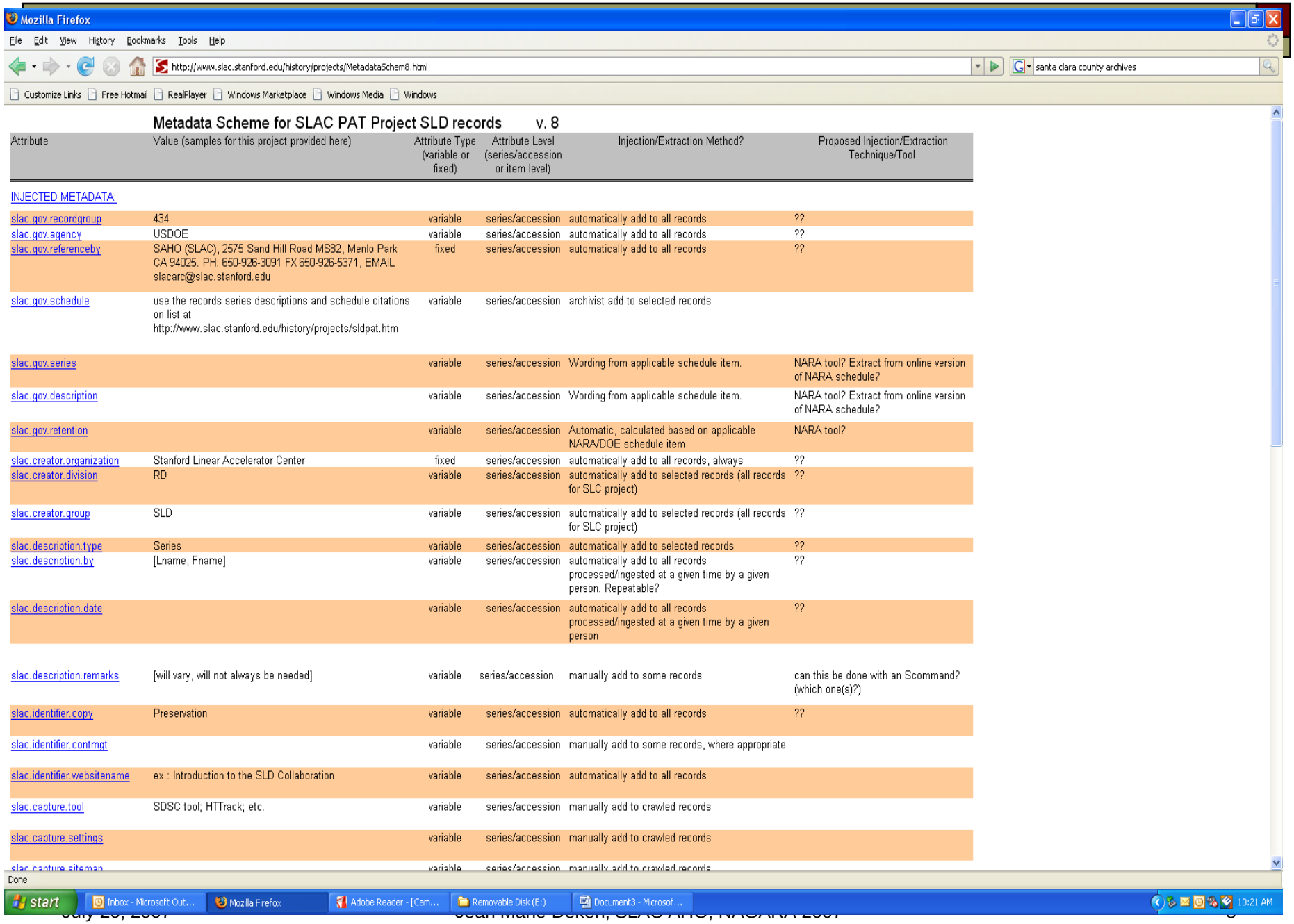
- SDSC crawl agent combs site, starting at http://www-sld.slac.stanford.edu and http://www-sldnt.slac.stanford.edu
- Crawl result reviewed by SLAC
  - Too broad horizontally
    - 1100 pages → 112 urls of interest
    - "No Crawl" messages stopped the agent

SLAC Metadata Development

- Began with the data elements that we currently use for our archives collections database, SLACARC
- Looked at
  - Dublin Core
  - METS
  - NARA metadata scheme

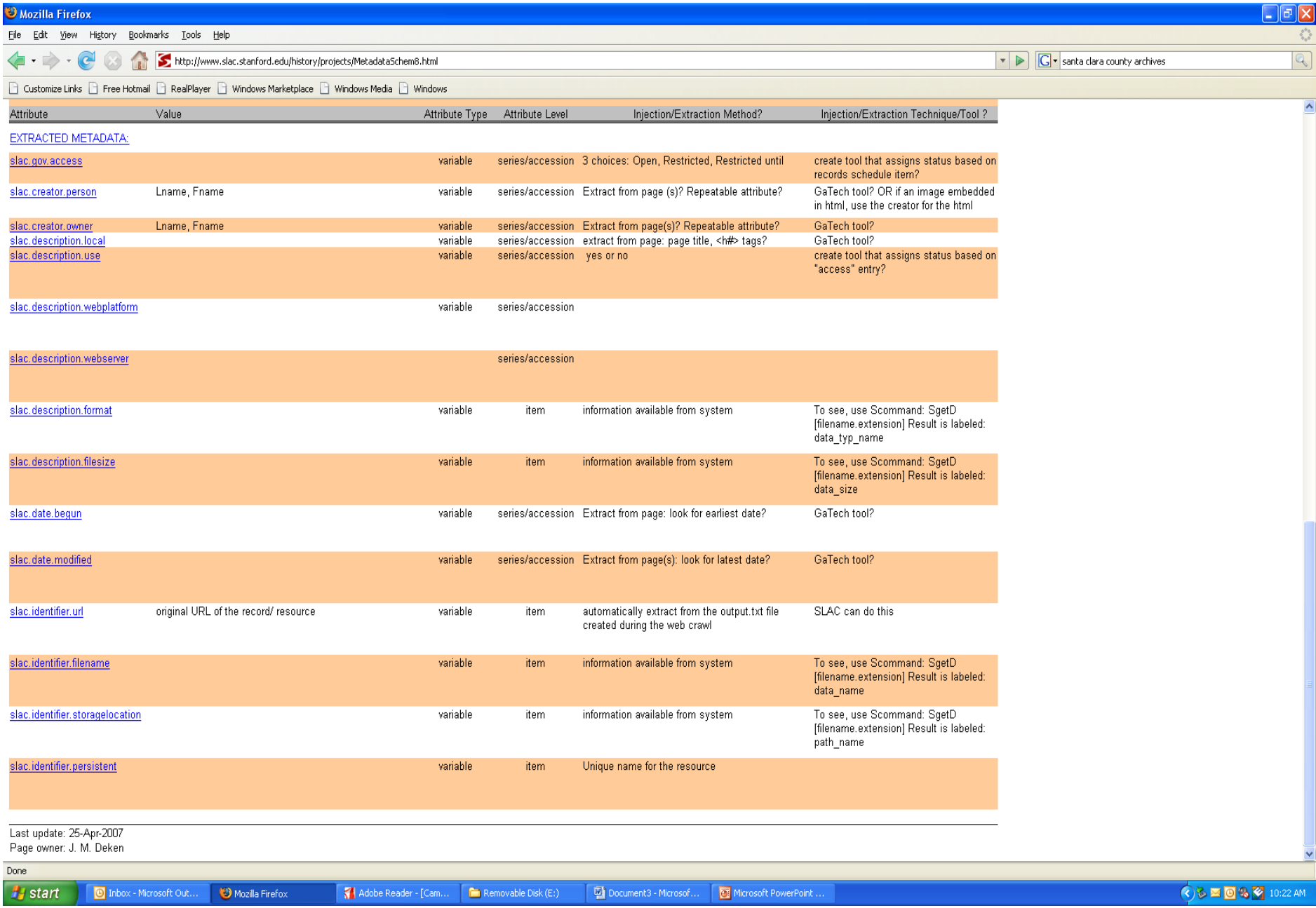
SLAC Metadata Development

- Now :
  - Application and revision
  - Some automated application of std. Metadata
  - http://www.slac.stanford.edu/history/projects/MetadataScheme8.html
- Future:
  - Automate application of std. metadata
  - Automate extraction of metadata



Metadata Scheme for SLAC PAT Project SLD records v. 8

Attribute	Value (samples for this project provided here)	Attribute Type (variable or fixed)	Attribute Level (series/accession or item level)	Injection/Extraction Method?	Proposed Injection/Extraction Technique/Tool
INJECTED METADATA:					
slac.gov.recordgroup	434	variable	series/accession	automatically add to all records	??
slac.gov.agency	USDOE	variable	series/accession	automatically add to all records	??
slac.gov.referenceby	SAHO (SLAC), 2575 Sand Hill Road MS82, Menlo Park CA 94025. PH: 650-926-3091 FX 650-926-5371, EMAIL slacarc@slac.stanford.edu	fixed	series/accession	automatically add to all records	??
slac.gov.schedule	use the records series descriptions and schedule citations on list at http://www.slac.stanford.edu/history/projects/sldpat.htm	variable	series/accession	archivist add to selected records	
slac.gov.series		variable	series/accession	Wording from applicable schedule item.	NARA tool? Extract from online version of NARA schedule?
slac.gov.description		variable	series/accession	Wording from applicable schedule item.	NARA tool? Extract from online version of NARA schedule?
slac.gov.retention		variable	series/accession	Automatic, calculated based on applicable NARA/DOE schedule item	NARA tool?
slac.creator.organization	Stanford Linear Accelerator Center	fixed	series/accession	automatically add to all records, always	??
slac.creator.division	RD	variable	series/accession	automatically add to selected records (all records for SLC project)	??
slac.creator.group	SLD	variable	series/accession	automatically add to selected records (all records for SLC project)	??
slac.description.type	Series	variable	series/accession	automatically add to selected records	??
slac.description.by	[Lname, Fname]	variable	series/accession	automatically add to all records processed/ingested at a given time by a given person. Repeatable?	??
slac.description.date		variable	series/accession	automatically add to all records processed/ingested at a given time by a given person	??
slac.description.remarks	[will vary, will not always be needed]	variable	series/accession	manually add to some records	can this be done with an Scommand? (which one(s)?)
slac.identifier.copy	Preservation	variable	series/accession	automatically add to all records	??
slac.identifier.contmgt		variable	series/accession	manually add to some records, where appropriate	
slac.identifier.websitename	ex.: Introduction to the SLD Collaboration	variable	series/accession	automatically add to all records	
slac.capture.tool	SDSC tool; HTTrack; etc.	variable	series/accession	manually add to crawled records	
slac.capture.settings		variable	series/accession	manually add to crawled records	
slac.capture.siteman		variable	series/accession	manually add to crawled records	

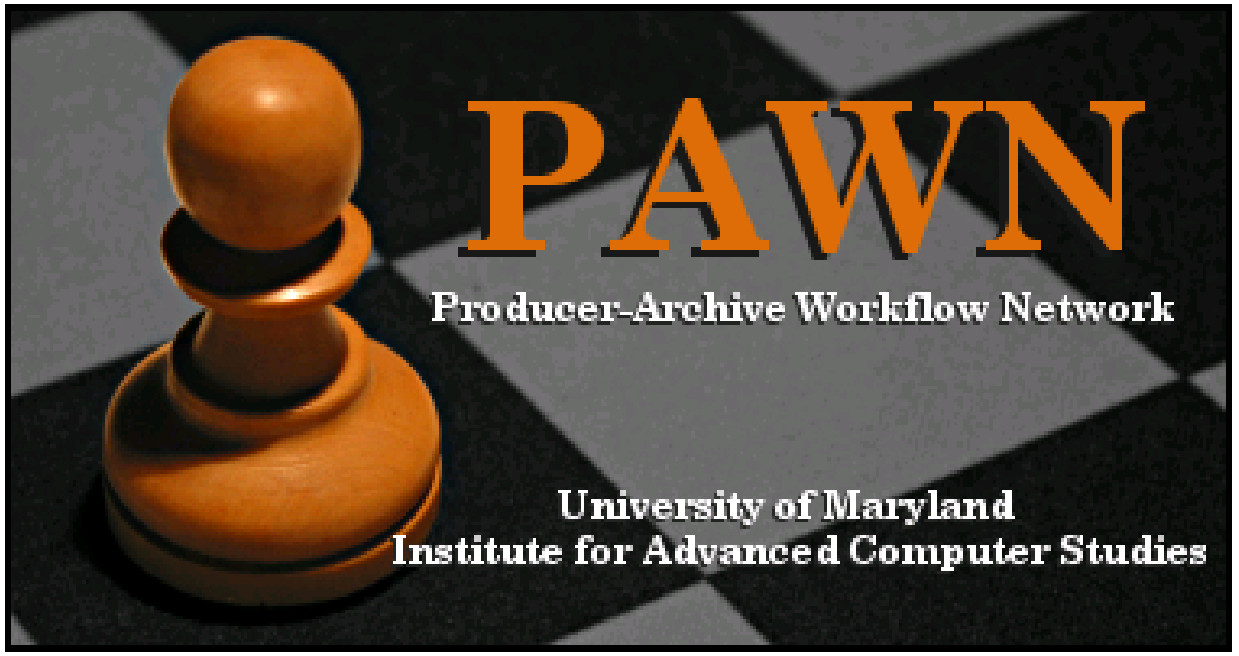


Attribute	Value	Attribute Type	Attribute Level	Injection/Extraction Method?	Injection/Extraction Technique/Tool ?
EXTRACTED METADATA:					
slac.gov.access		variable	series/accession	3 choices: Open, Restricted, Restricted until	create tool that assigns status based on records schedule item?
slac.creator.person	Lname, Fname	variable	series/accession	Extract from page (s)? Repeatable attribute?	GaTech tool? OR if an image embedded in html, use the creator for the html
slac.creator.owner	Lname, Fname	variable	series/accession	Extract from page(s)? Repeatable attribute?	GaTech tool?
slac.description.local		variable	series/accession	extract from page: page title, <h#> tags?	GaTech tool?
slac.description.use		variable	series/accession	yes or no	create tool that assigns status based on "access" entry?
slac.description.webplatform		variable	series/accession		
slac.description.webserver			series/accession		
slac.description.format		variable	item	information available from system	To see, use Scommand: SgetD [filename.extension] Result is labeled: data_typ_name
slac.description.filesize		variable	item	information available from system	To see, use Scommand: SgetD [filename.extension] Result is labeled: data_size
slac.date.begun		variable	series/accession	Extract from page: look for earliest date?	GaTech tool?
slac.date.modified		variable	series/accession	Extract from page(s): look for latest date?	GaTech tool?
slac.identifier.url	original URL of the record/ resource	variable	item	automatically extract from the output.txt file created during the web crawl	SLAC can do this
slac.identifier.filename		variable	item	information available from system	To see, use Scommand: SgetD [filename.extension] Result is labeled: data_name
slac.identifier.storagelocation		variable	item	information available from system	To see, use Scommand: SgetD [filename.extension] Result is labeled: path_name
slac.identifier.persistent		variable	item	Unique name for the resource	

Last update: 25-Apr-2007
Page owner: J. M. Deken

PAWN WORKFLOW:

SLAC → NARA



Version 0.5.2 (2006-05-12)

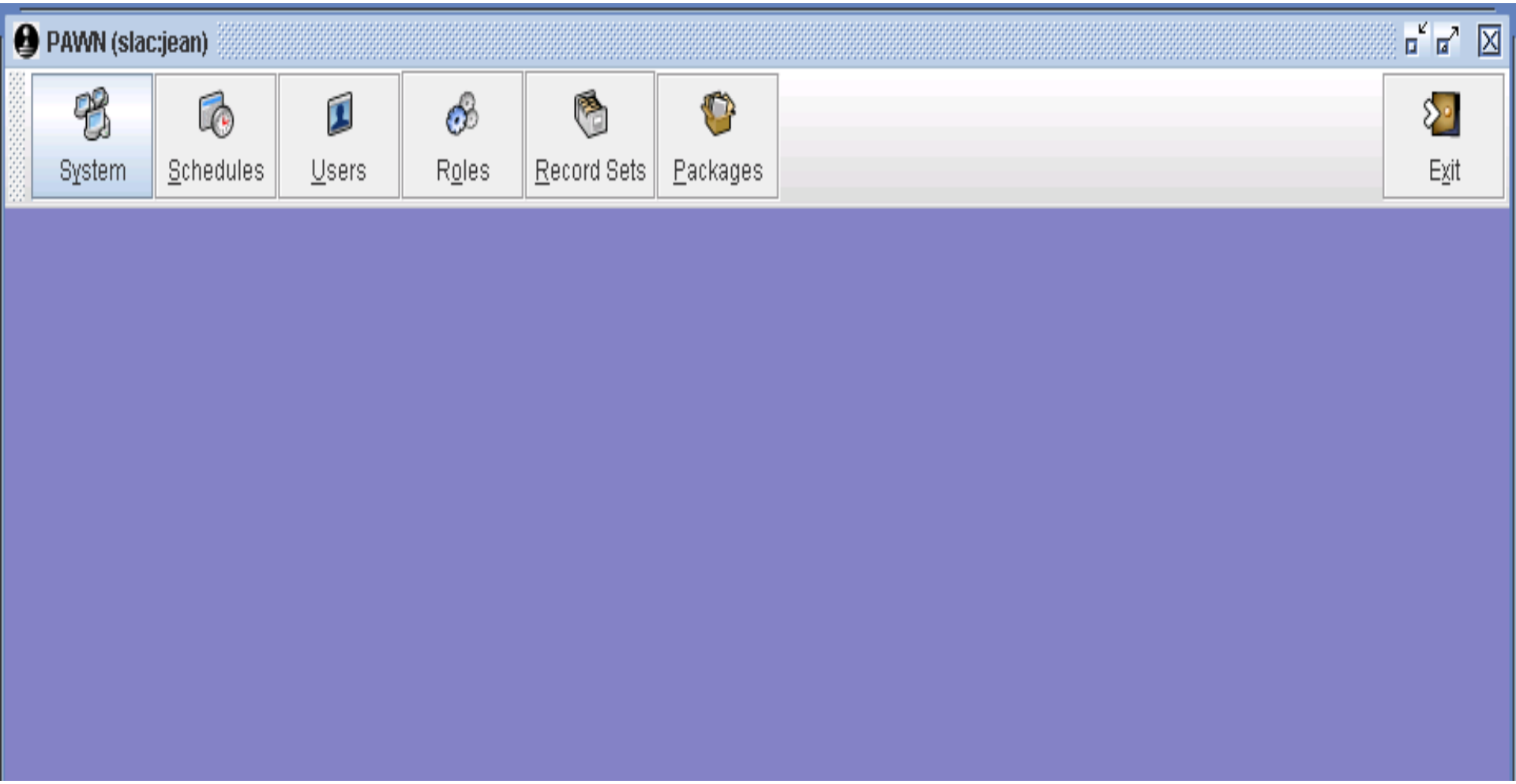
Hostname: narapawn.umiacs.umd.edu

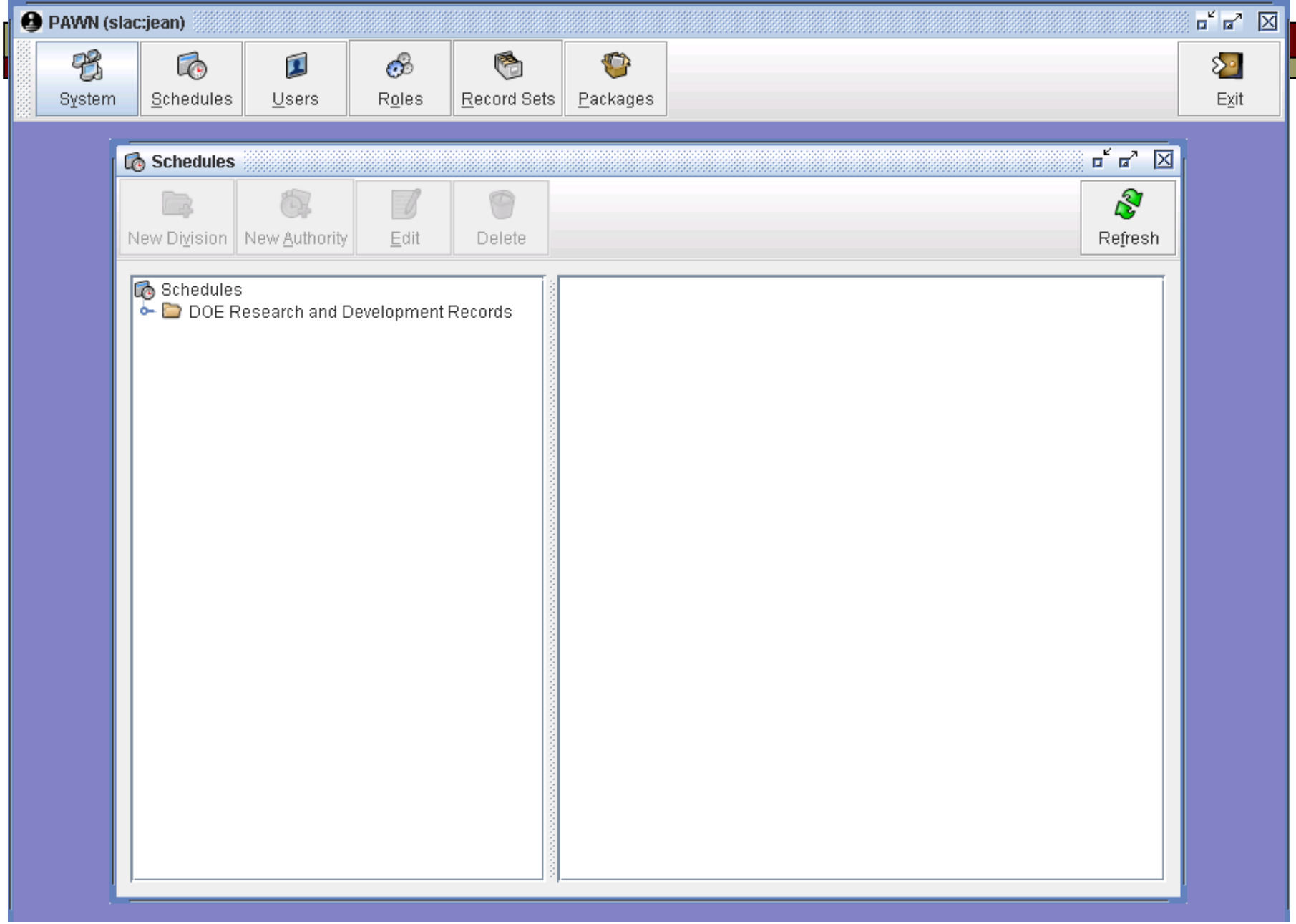
Username: slac:jean

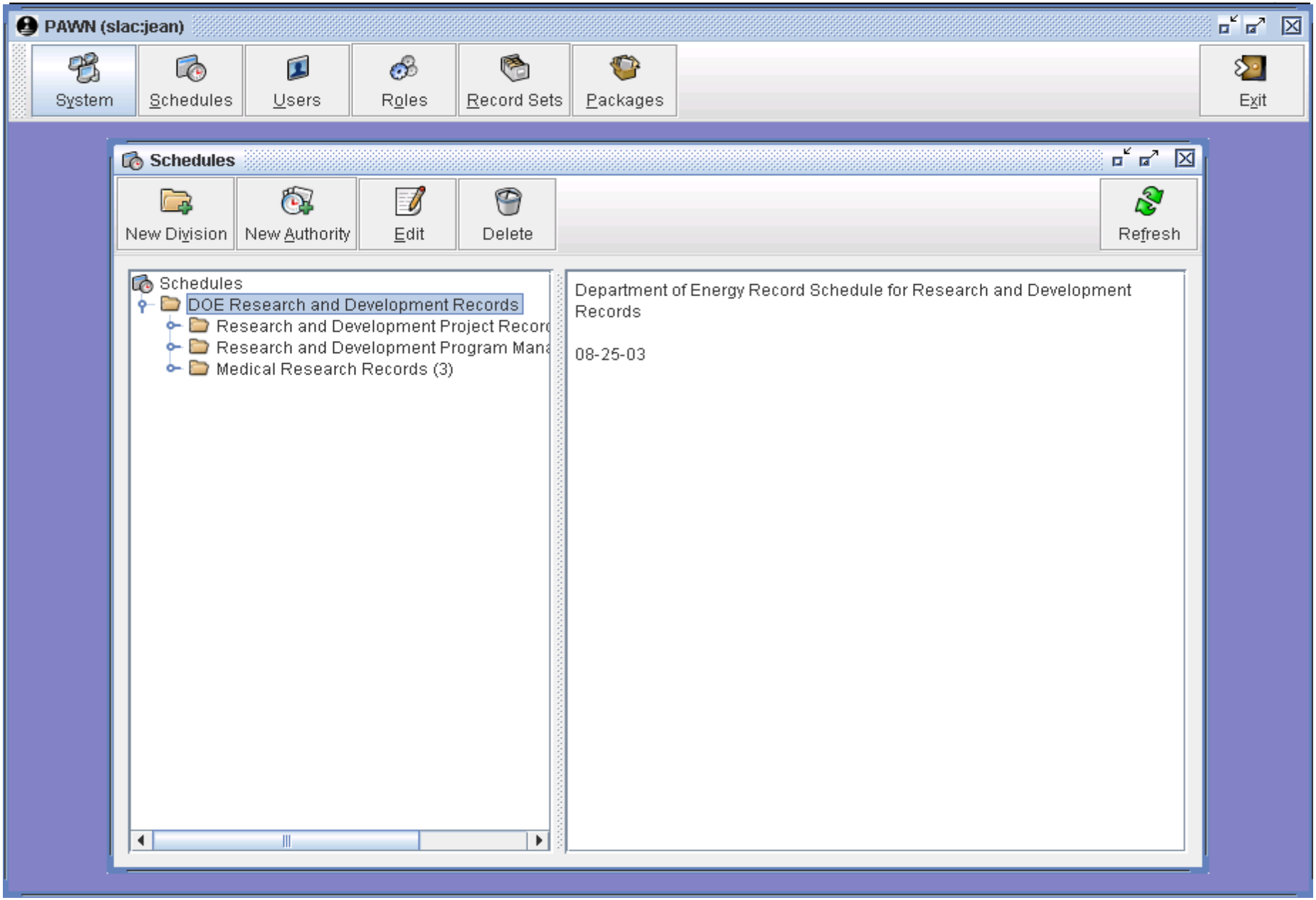
Password: ****

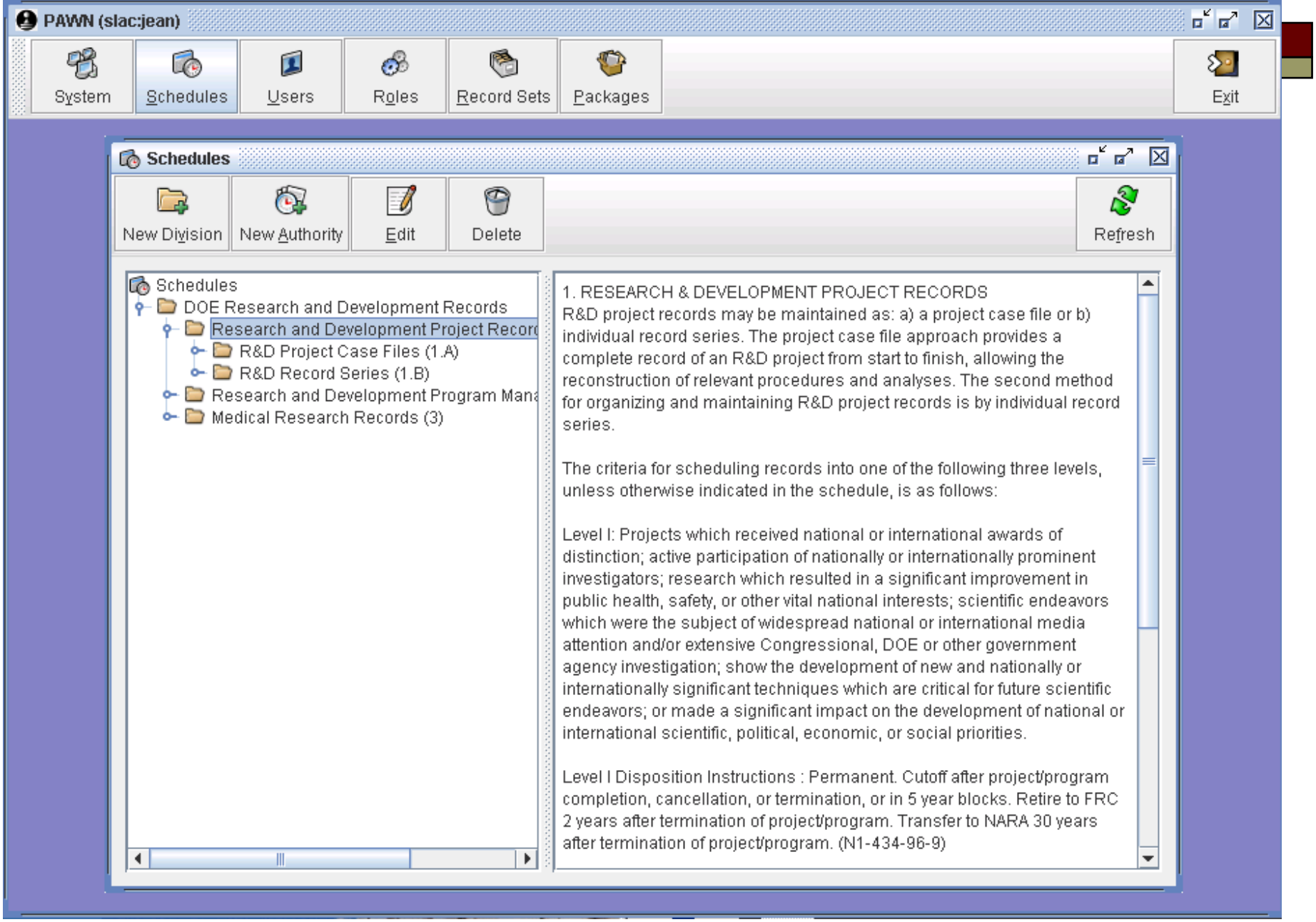
Keystore File: :ings\jmdeken\.edu.umiacs\keystore.pl2

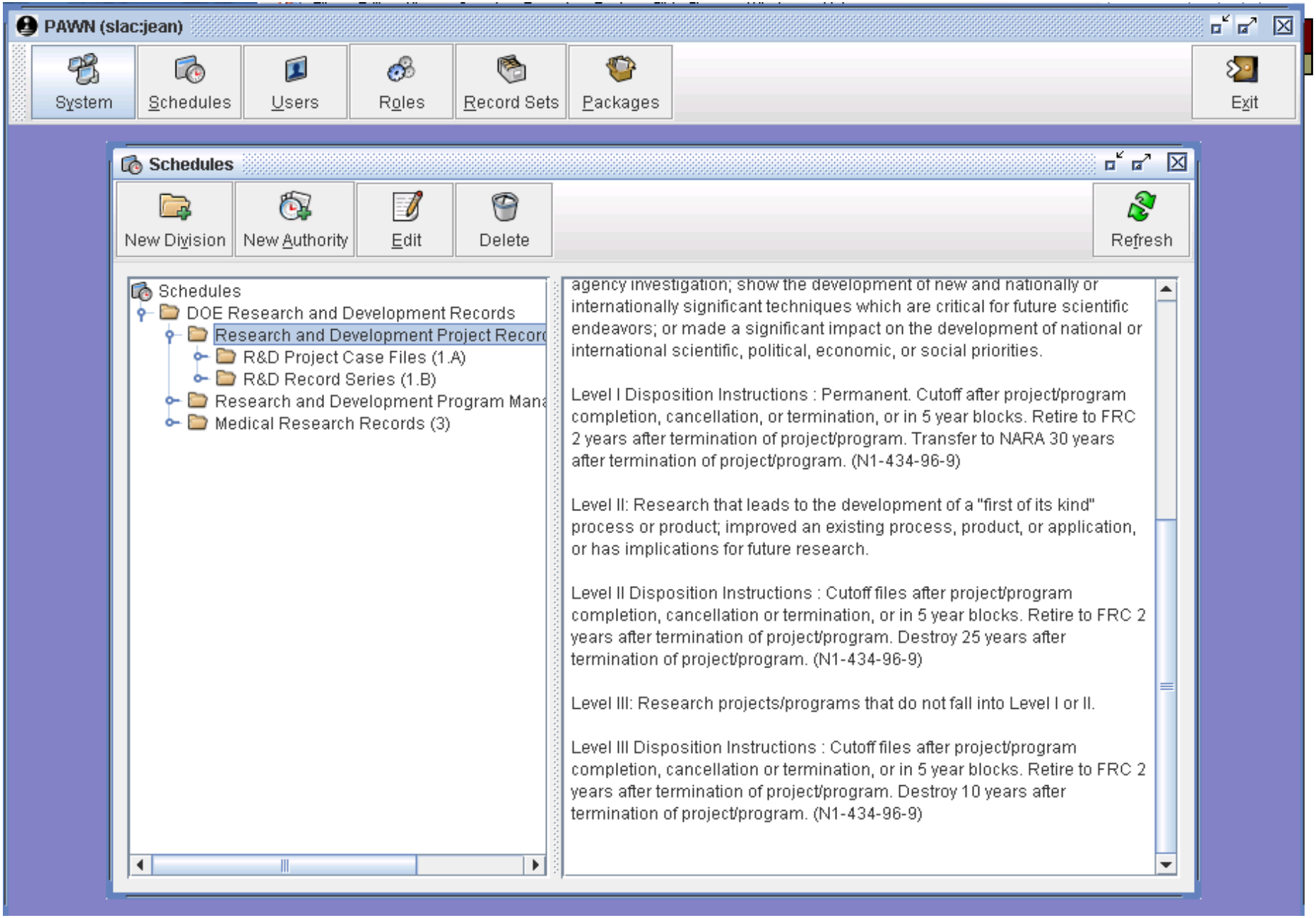
Exit Connect

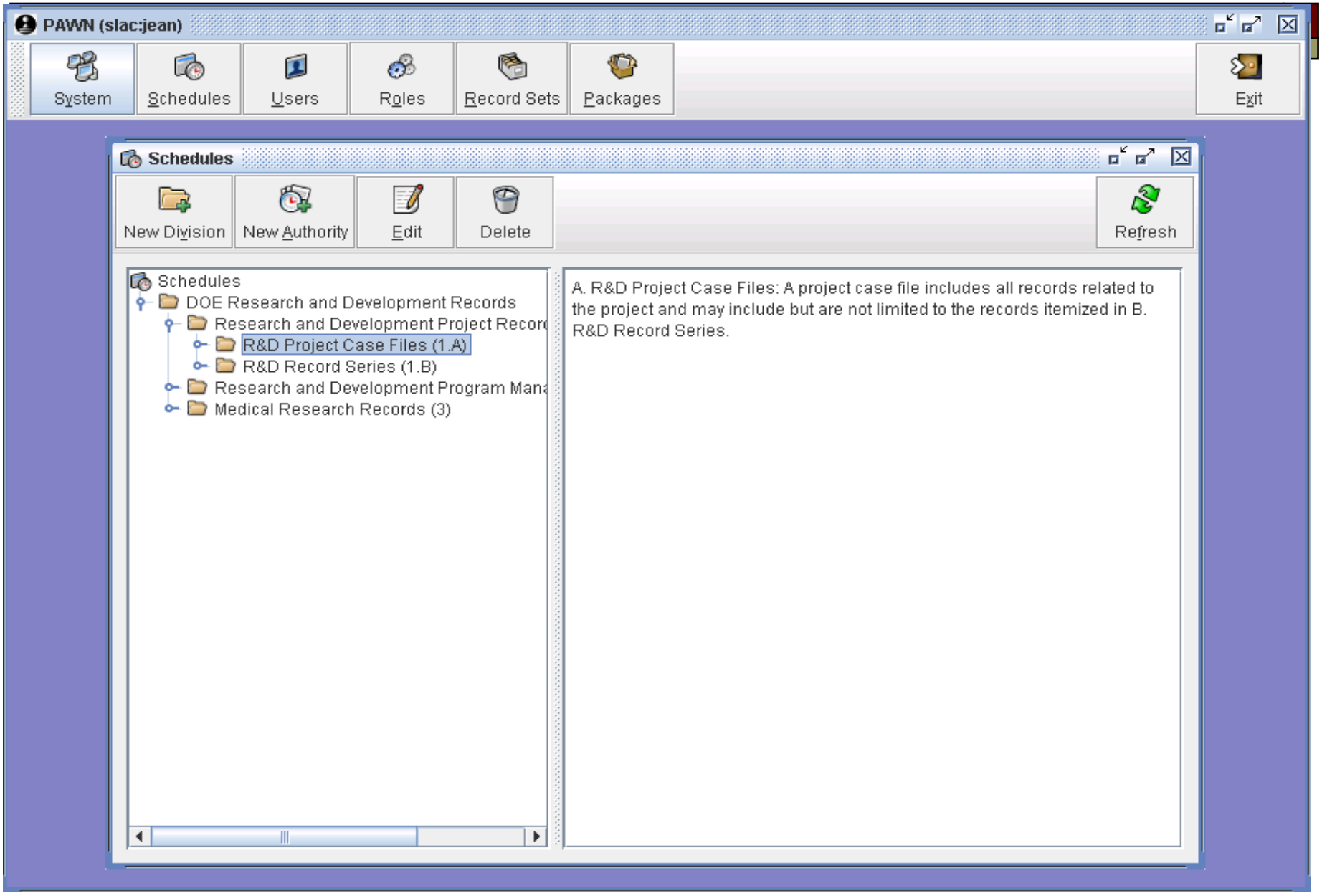

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit

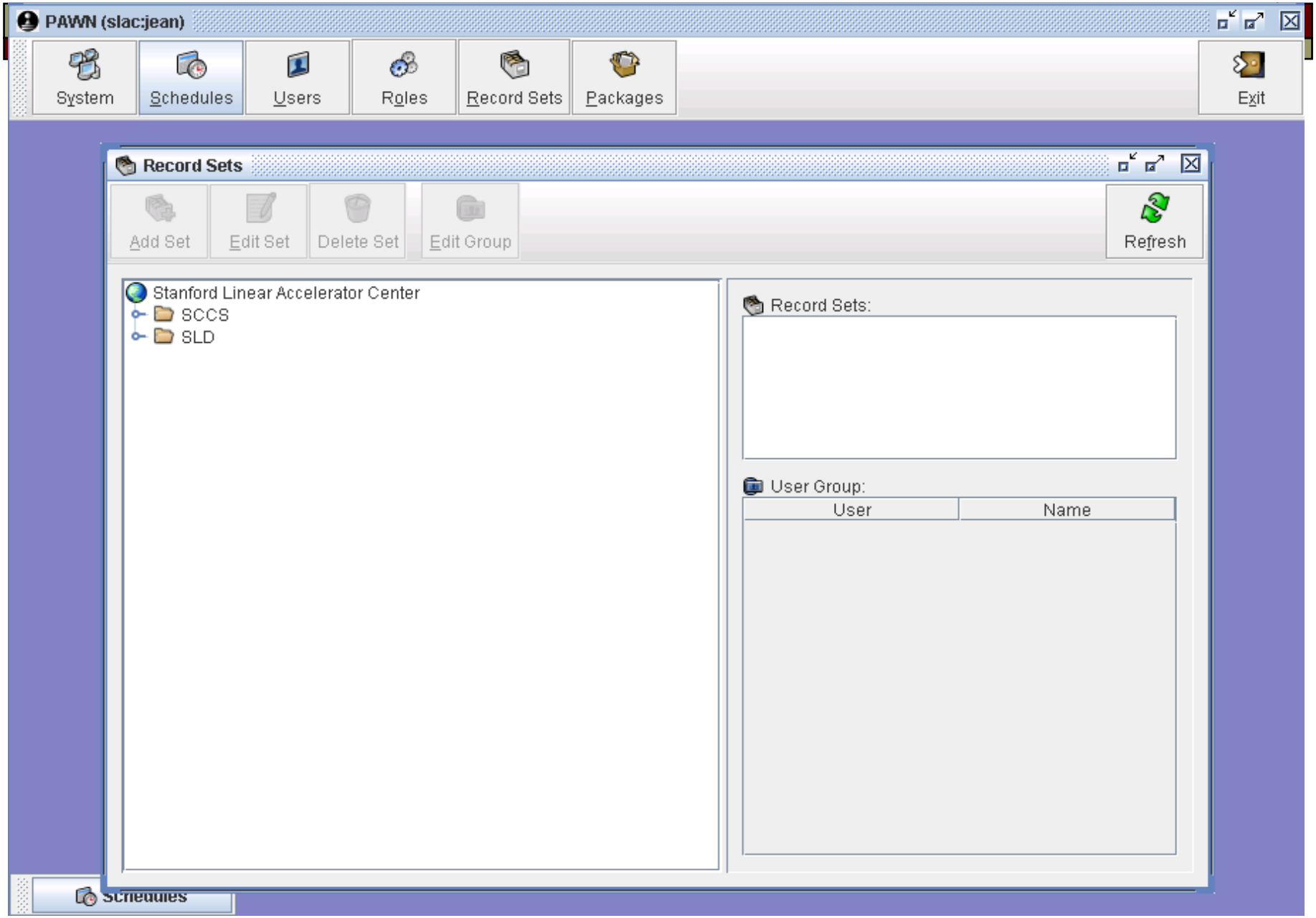

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Schedules
New Division
New Authority
Edit
Delete
Refresh
Schedules
DOE Research and Development Records

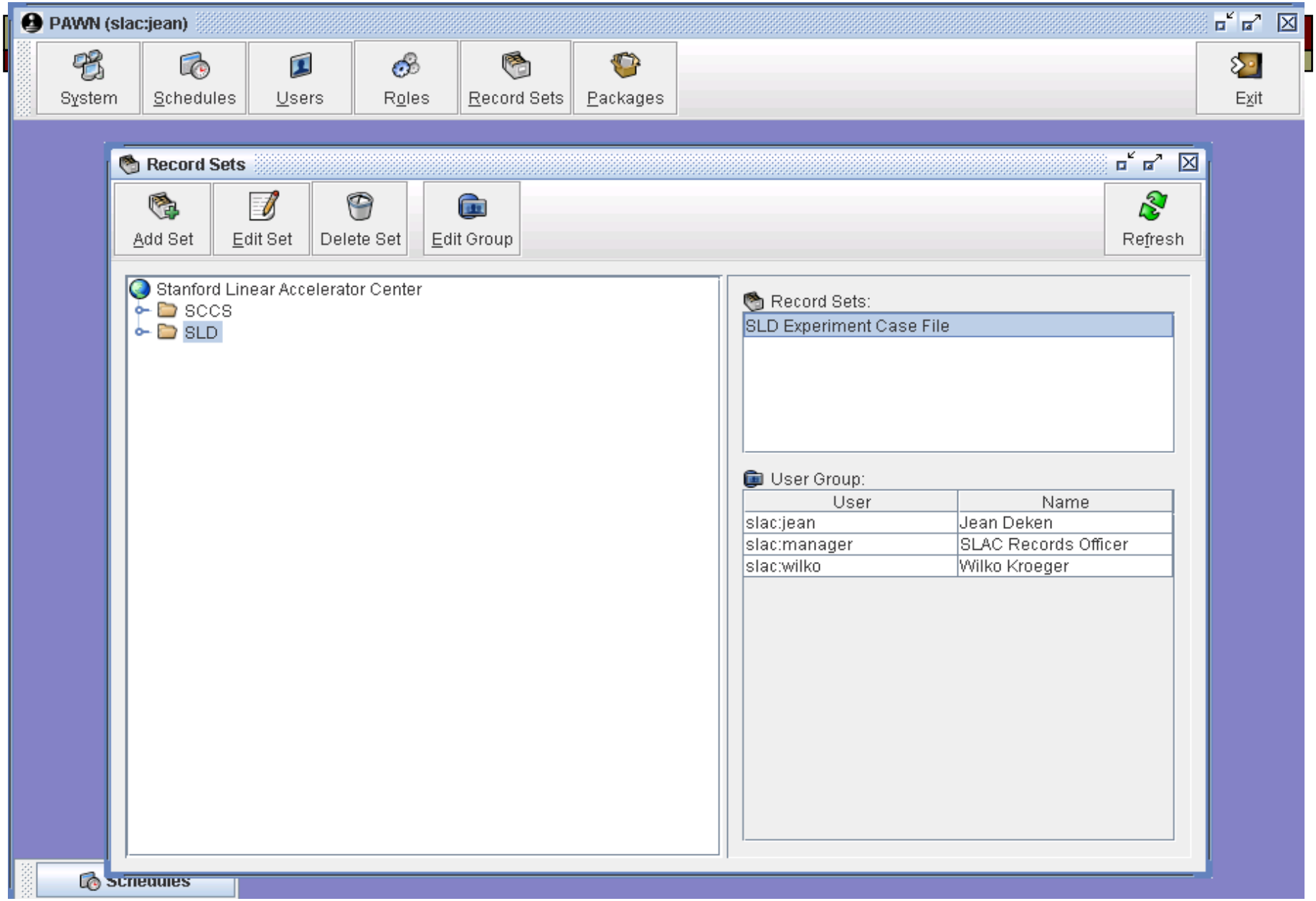

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Schedules
New Division
New Authority
Edit
Delete
Refresh
Schedules
DOE Research and Development Records
Research and Development Project Record
Research and Development Program Mana
Medical Research Records (3)
Department of Energy Record Schedule for Research and Development Records
08-25-03

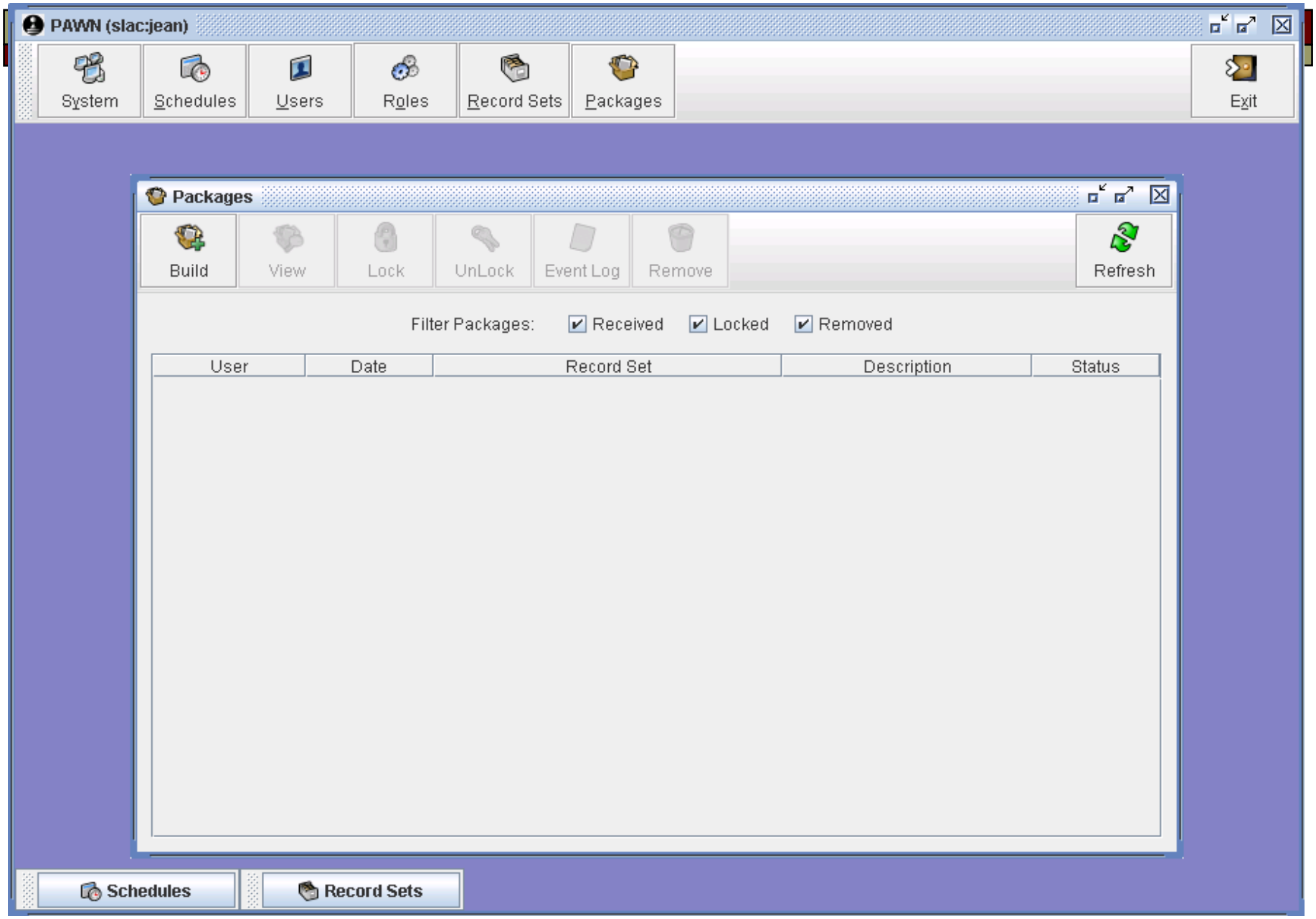

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Schedules
New Division
New Authority
Edit
Delete
Refresh
Schedules
DOE Research and Development Records
Research and Development Project Record
R&D Project Case Files (1.A)
R&D Record Series (1.B)
Research and Development Program Mana
Medical Research Records (3)
1. RESEARCH & DEVELOPMENT PROJECT RECORDS
R&D project records may be maintained as: a) a project case file or b) individual record series. The project case file approach provides a complete record of an R&D project from start to finish, allowing the reconstruction of relevant procedures and analyses. The second method for organizing and maintaining R&D project records is by individual record series.
The criteria for scheduling records into one of the following three levels, unless otherwise indicated in the schedule, is as follows:
Level I: Projects which received national or international awards of distinction; active participation of nationally or internationally prominent investigators; research which resulted in a significant improvement in public health, safety, or other vital national interests; scientific endeavors which were the subject of widespread national or international media attention and/or extensive Congressional, DOE or other government agency investigation; show the development of new and nationally or internationally significant techniques which are critical for future scientific endeavors; or made a significant impact on the development of national or international scientific, political, economic, or social priorities.
Level I Disposition Instructions : Permanent. Cutoff after project/program completion, cancellation, or termination, or in 5 year blocks. Retire to FRC 2 years after termination of project/program. Transfer to NARA 30 years after termination of project/program. (N1-434-96-9)

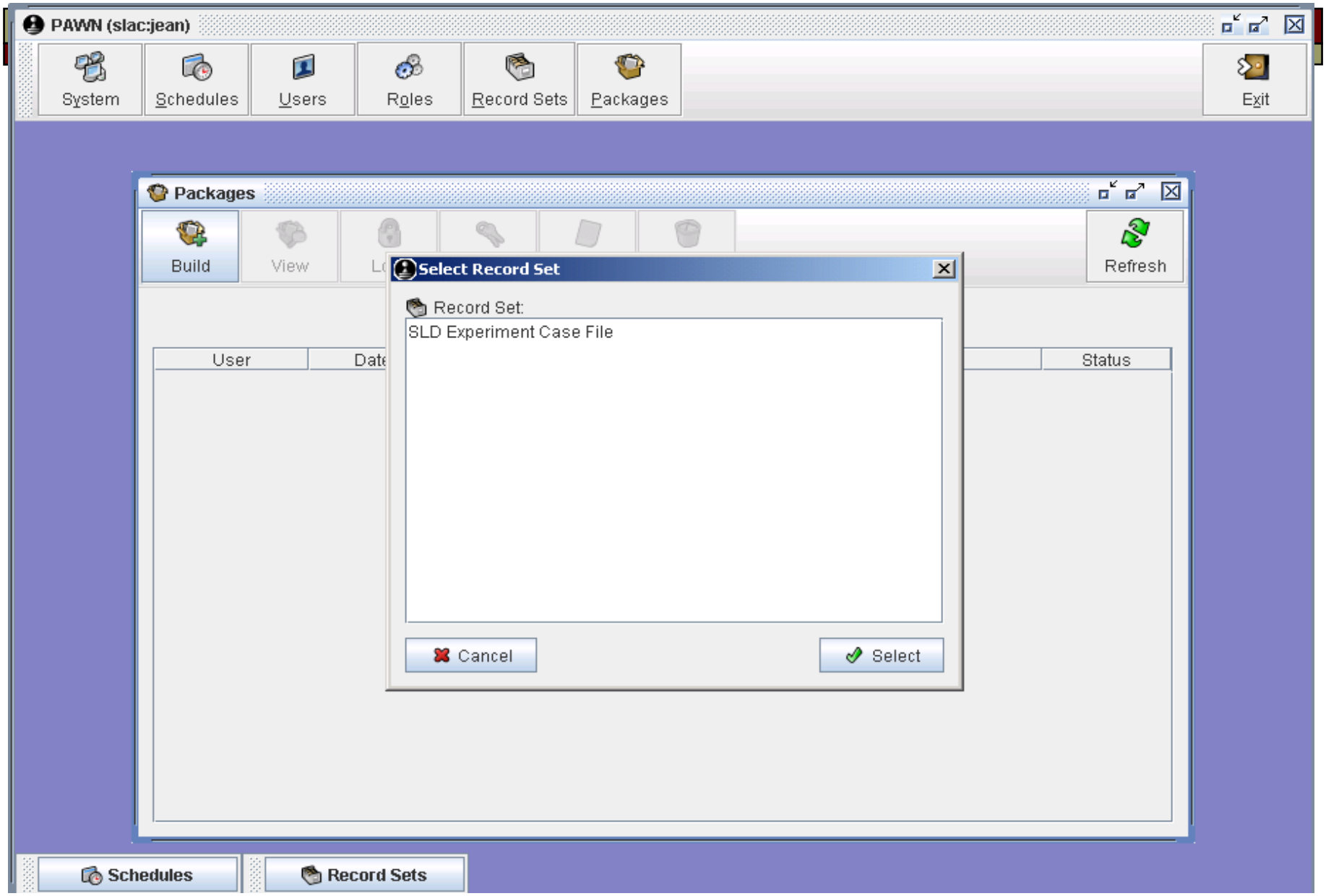

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Schedules
New Division
New Authority
Edit
Delete
Refresh
Schedules
DOE Research and Development Records
Research and Development Project Record
R&D Project Case Files (1.A)
R&D Record Series (1.B)
Research and Development Program Mana
Medical Research Records (3)
agency investigation; show the development of new and nationally or internationally significant techniques which are critical for future scientific endeavors; or made a significant impact on the development of national or international scientific, political, economic, or social priorities.
Level I Disposition Instructions : Permanent. Cutoff after project/program completion, cancellation, or termination, or in 5 year blocks. Retire to FRC 2 years after termination of project/program. Transfer to NARA 30 years after termination of project/program. (N1-434-96-9)
Level II: Research that leads to the development of a "first of its kind" process or product; improved an existing process, product, or application, or has implications for future research.
Level II Disposition Instructions : Cutoff files after project/program completion, cancellation or termination, or in 5 year blocks. Retire to FRC 2 years after termination of project/program. Destroy 25 years after termination of project/program. (N1-434-96-9)
Level III: Research projects/programs that do not fall into Level I or II.
Level III Disposition Instructions : Cutoff files after project/program completion, cancellation or termination, or in 5 year blocks. Retire to FRC 2 years after termination of project/program. Destroy 10 years after termination of project/program. (N1-434-96-9)

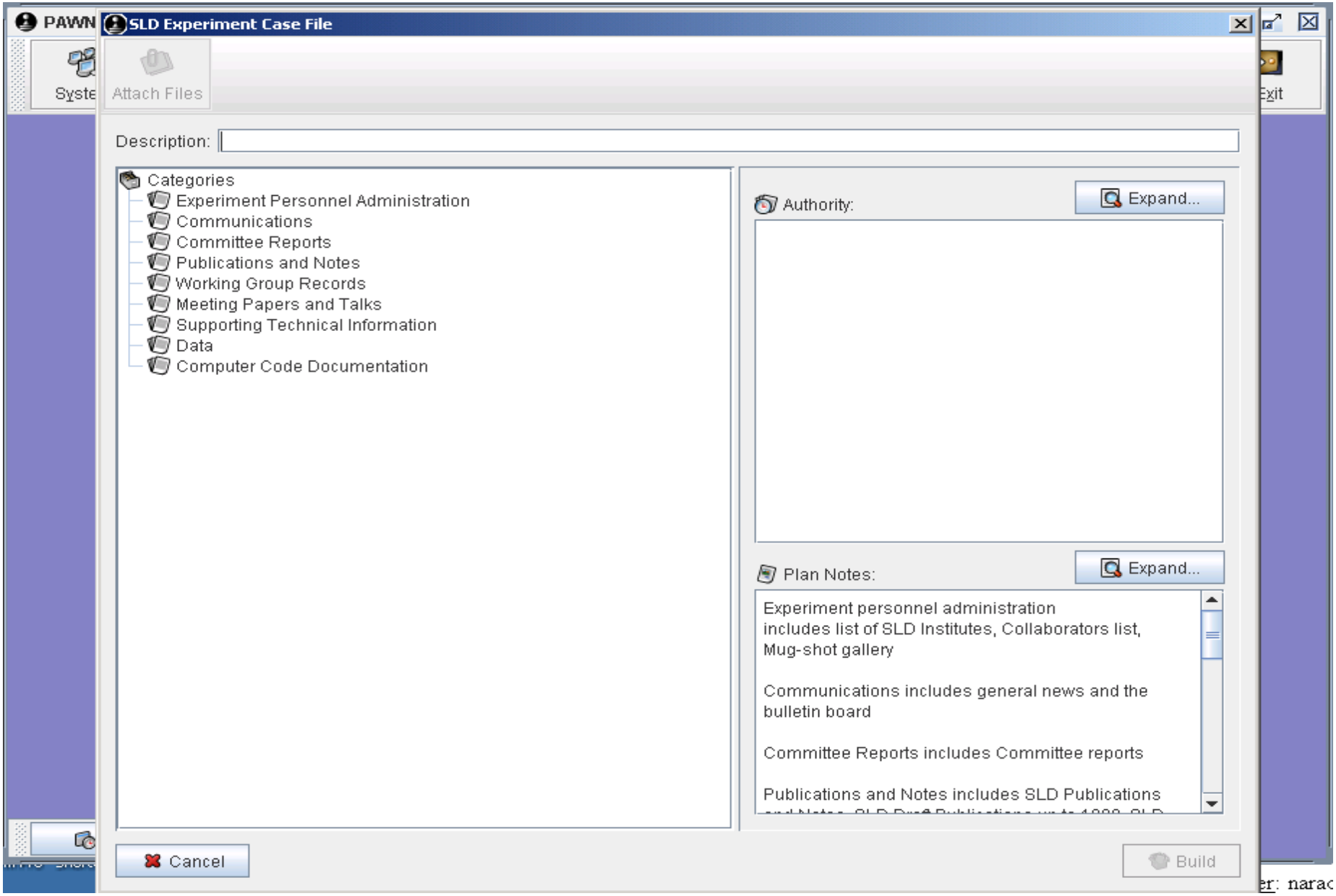

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Schedules
New Division
New Authority
Edit
Delete
Refresh
Schedules
DOE Research and Development Records
Research and Development Project Recor
R&D Project Case Files (1.A)
R&D Record Series (1.B)
Research and Development Program Mana
Medical Research Records (3)
A. R&D Project Case Files: A project case file includes all records related to the project and may include but are not limited to the records itemized in B. R&D Record Series.

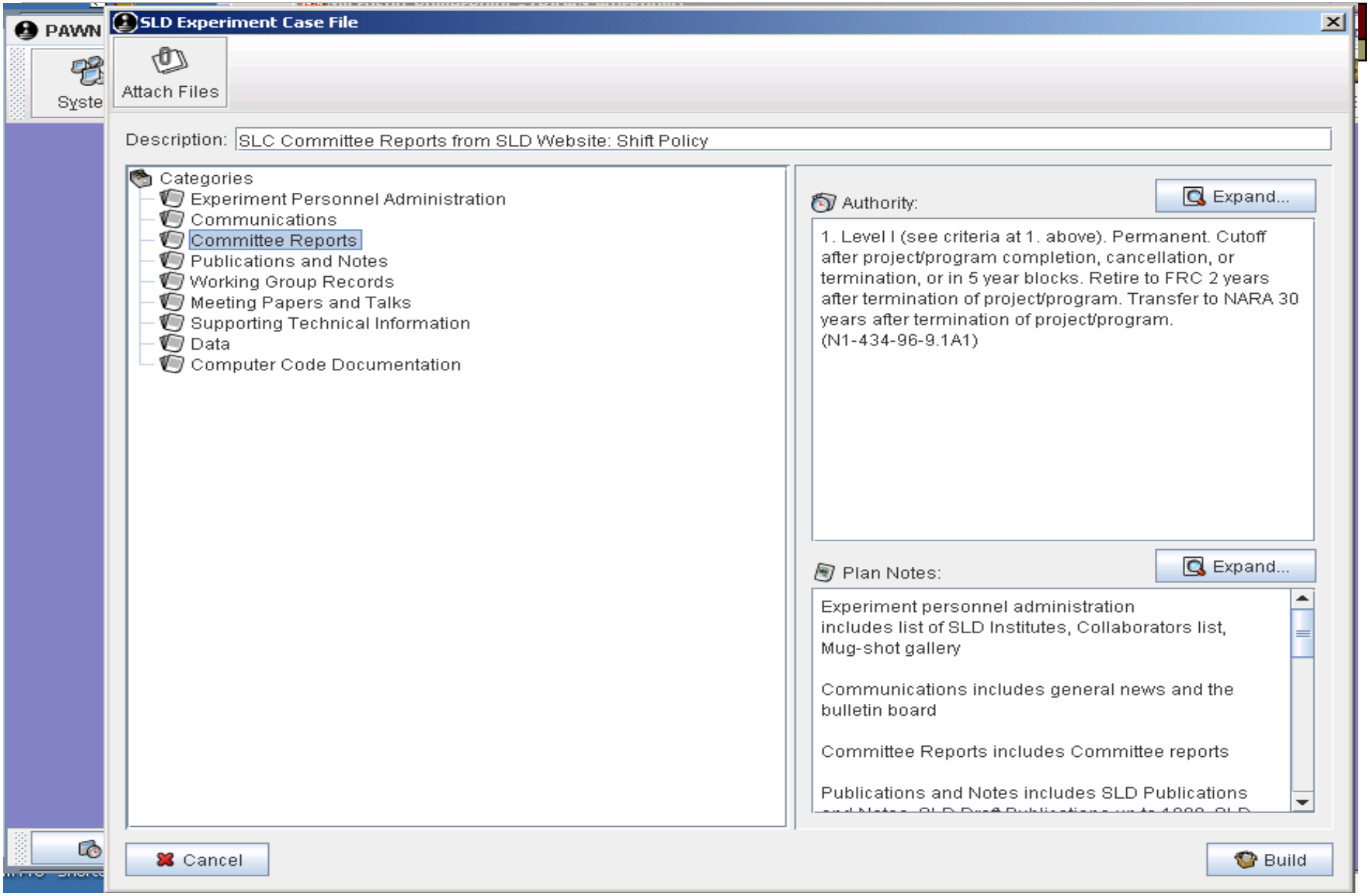

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Record Sets
Add Set
Edit Set
Delete Set
Edit Group
Refresh
Stanford Linear Accelerator Center
SCCS
SLD
Record Sets:
User Group:
User
Name
Schedules

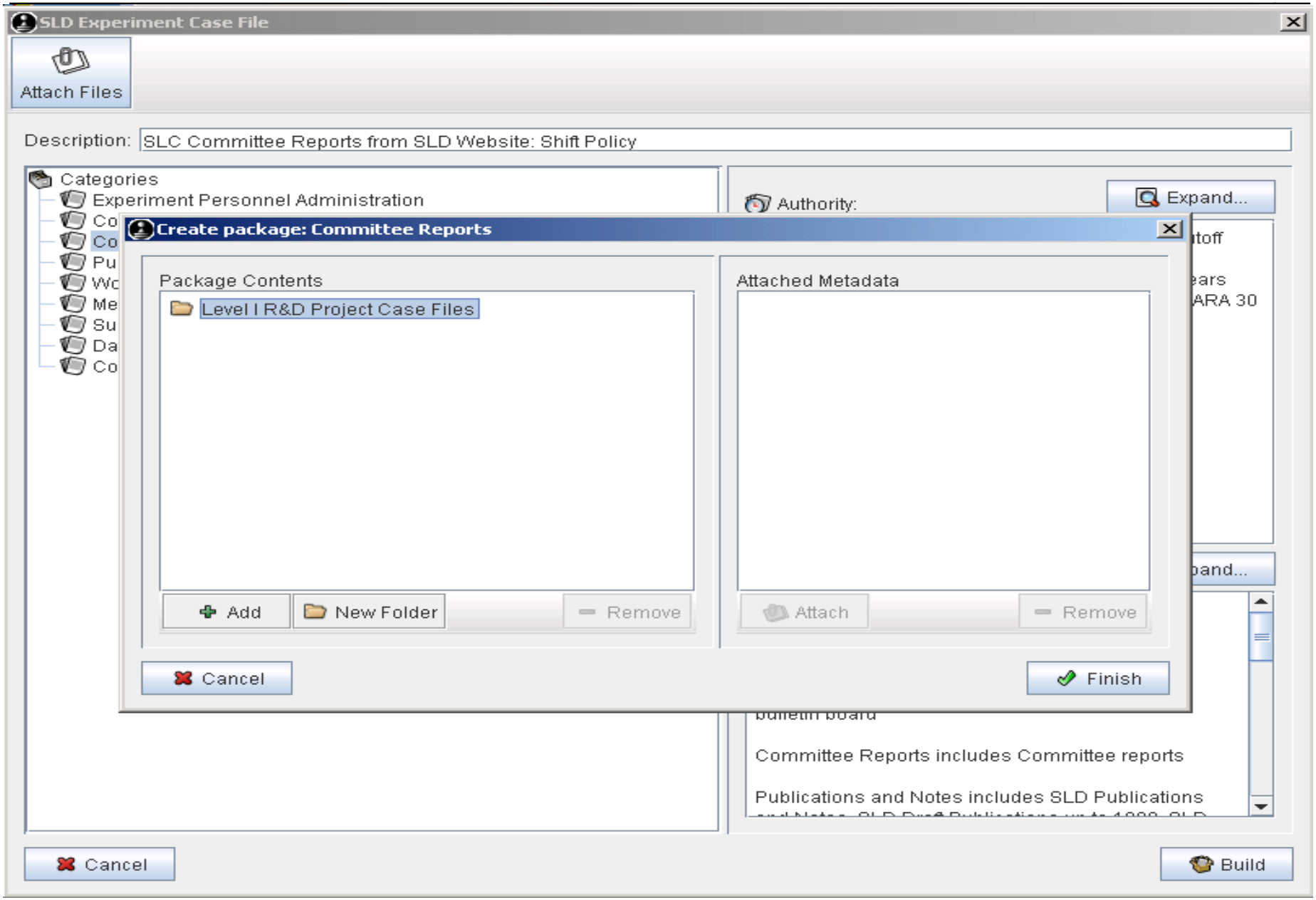

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Record Sets
Add Set
Edit Set
Delete Set
Edit Group
Refresh
Stanford Linear Accelerator Center
SCCS
SLD
Record Sets:
SLD Experiment Case File
User Group:
User
Name
slac:jean
Jean Deken
slac:manager
SLAC Records Officer
slac:wilko
Wilko Kroeger
Schedules

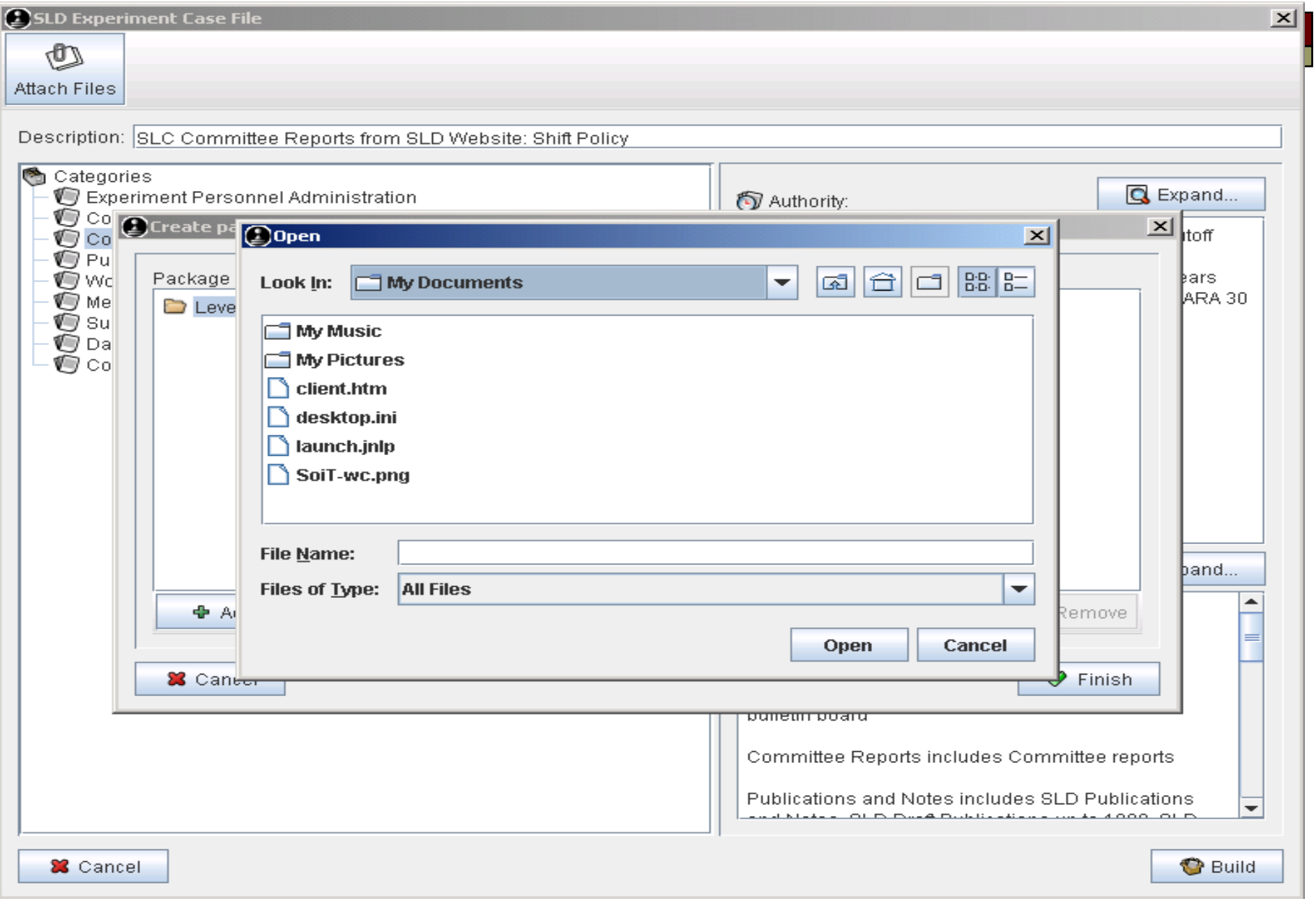

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Packages
Build
View
Lock
UnLock
Event Log
Remove
Refresh
Filter Packages:
Received
Locked
Removed
User
Date
Record Set
Description
Status
Schedules
Record Sets

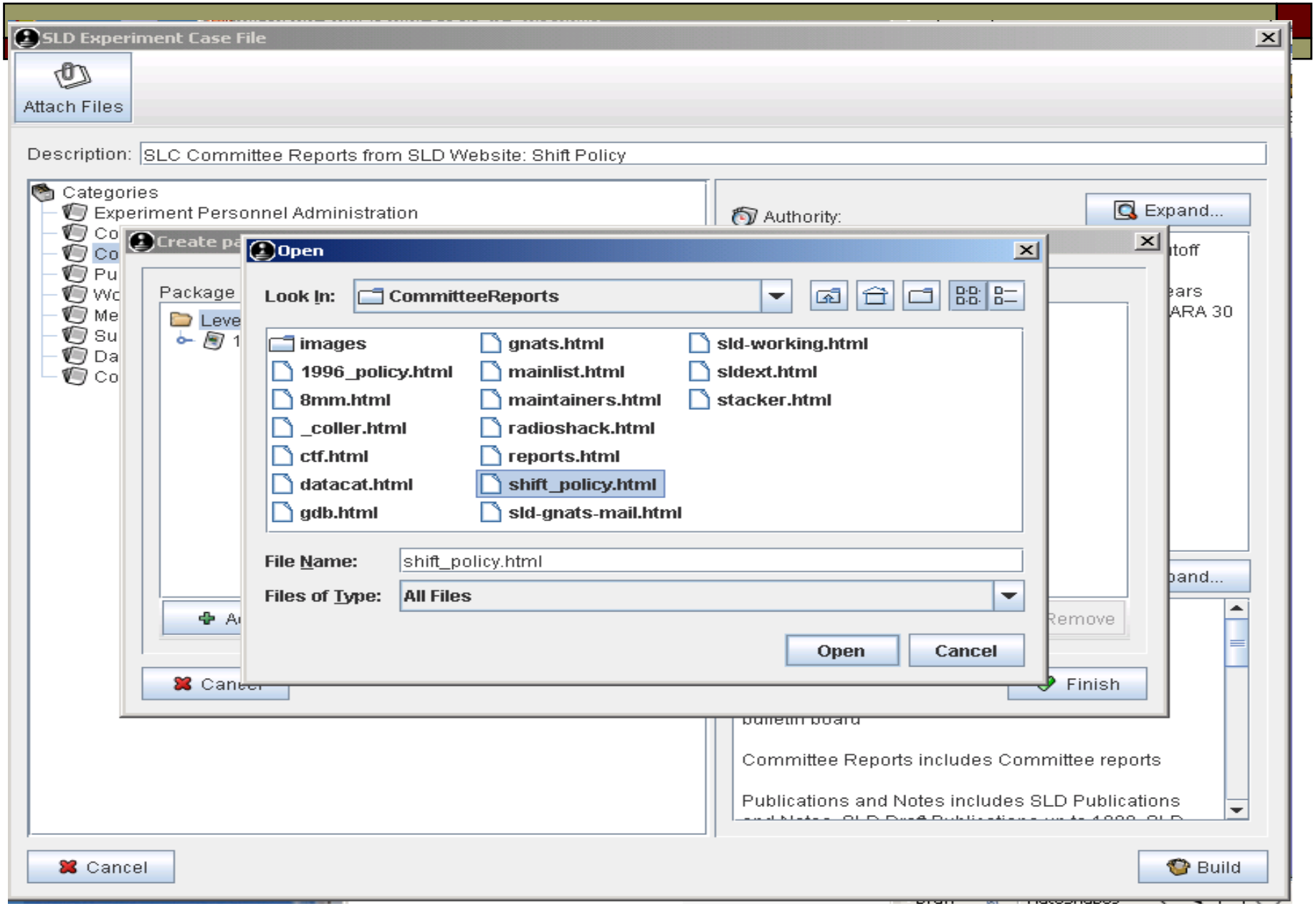

PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Packages
Build
View
Refresh
Select Record Set
Record Set:
SLD Experiment Case File
Cancel
Select
User
Status
Schedules
Record Sets

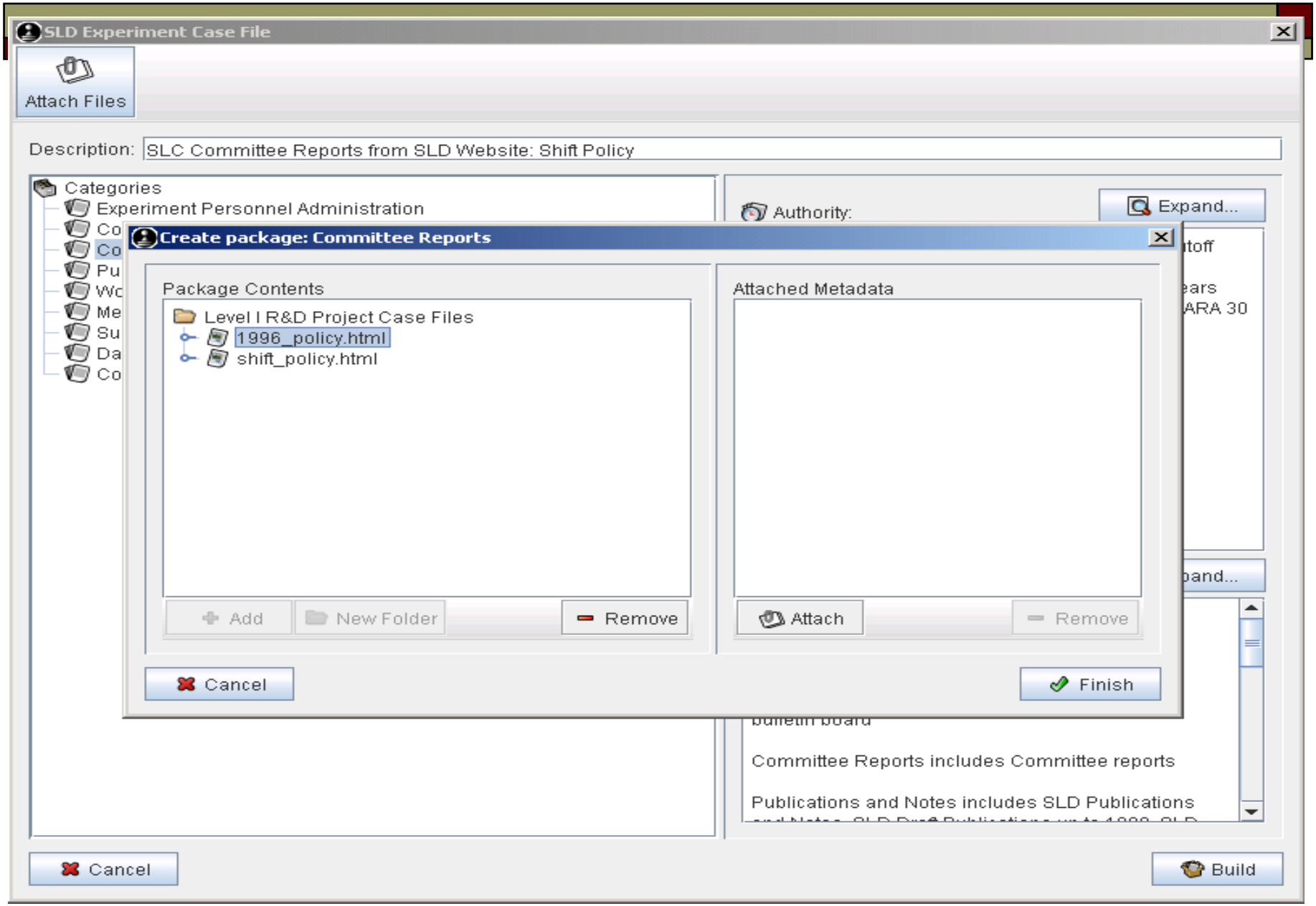

PAWN
SLD Experiment Case File
Attach Files
Description:
Categories
Experiment Personnel Administration
Communications
Committee Reports
Publications and Notes
Working Group Records
Meeting Papers and Talks
Supporting Technical Information
Data
Computer Code Documentation
Authority:
Expand...
Plan Notes:
Expand...
Experiment personnel administration includes list of SLD Institutes, Collaborators list, Mug-shot gallery
Communications includes general news and the bulletin board
Committee Reports includes Committee reports
Publications and Notes includes SLD Publications
Cancel
Build

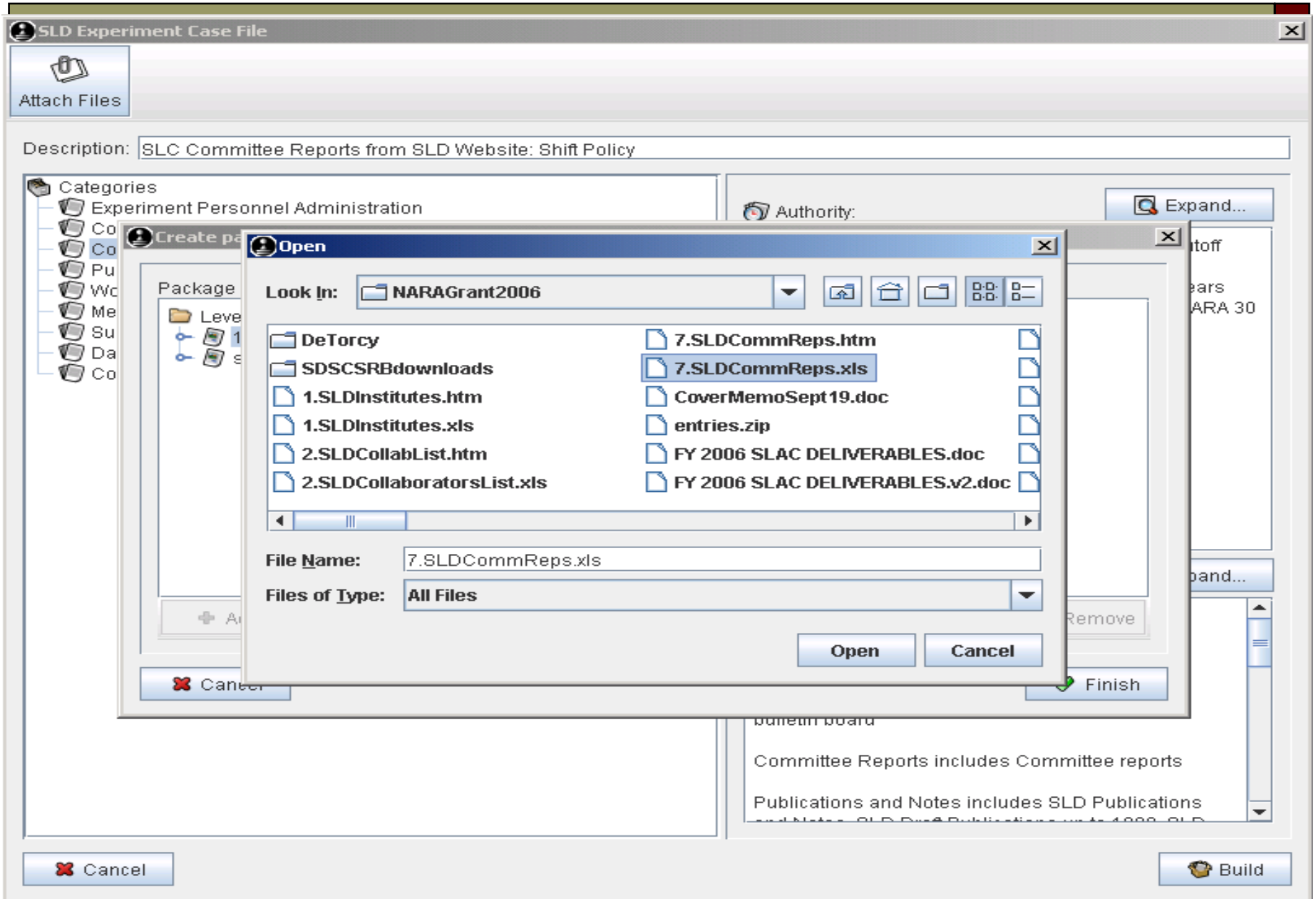

PAWN
Syste
SLD Experiment Case File
Attach Files
Description: SLC Committee Reports from SLD Website: Shift Policy
Categories
Experiment Personnel Administration
Communications
Committee Reports
Publications and Notes
Working Group Records
Meeting Papers and Talks
Supporting Technical Information
Data
Computer Code Documentation
Authority:
Expand...
1. Level I (see criteria at 1. above). Permanent. Cutoff after project/program completion, cancellation, or termination, or in 5 year blocks. Retire to FRC 2 years after termination of project/program. Transfer to NARA 30 years after termination of project/program. (N1-434-96-9.1A1)
Plan Notes:
Expand...
Experiment personnel administration includes list of SLD Institutes, Collaborators list, Mug-shot gallery
Communications includes general news and the bulletin board
Committee Reports includes Committee reports
Publications and Notes includes SLD Publications
Cancel
Build

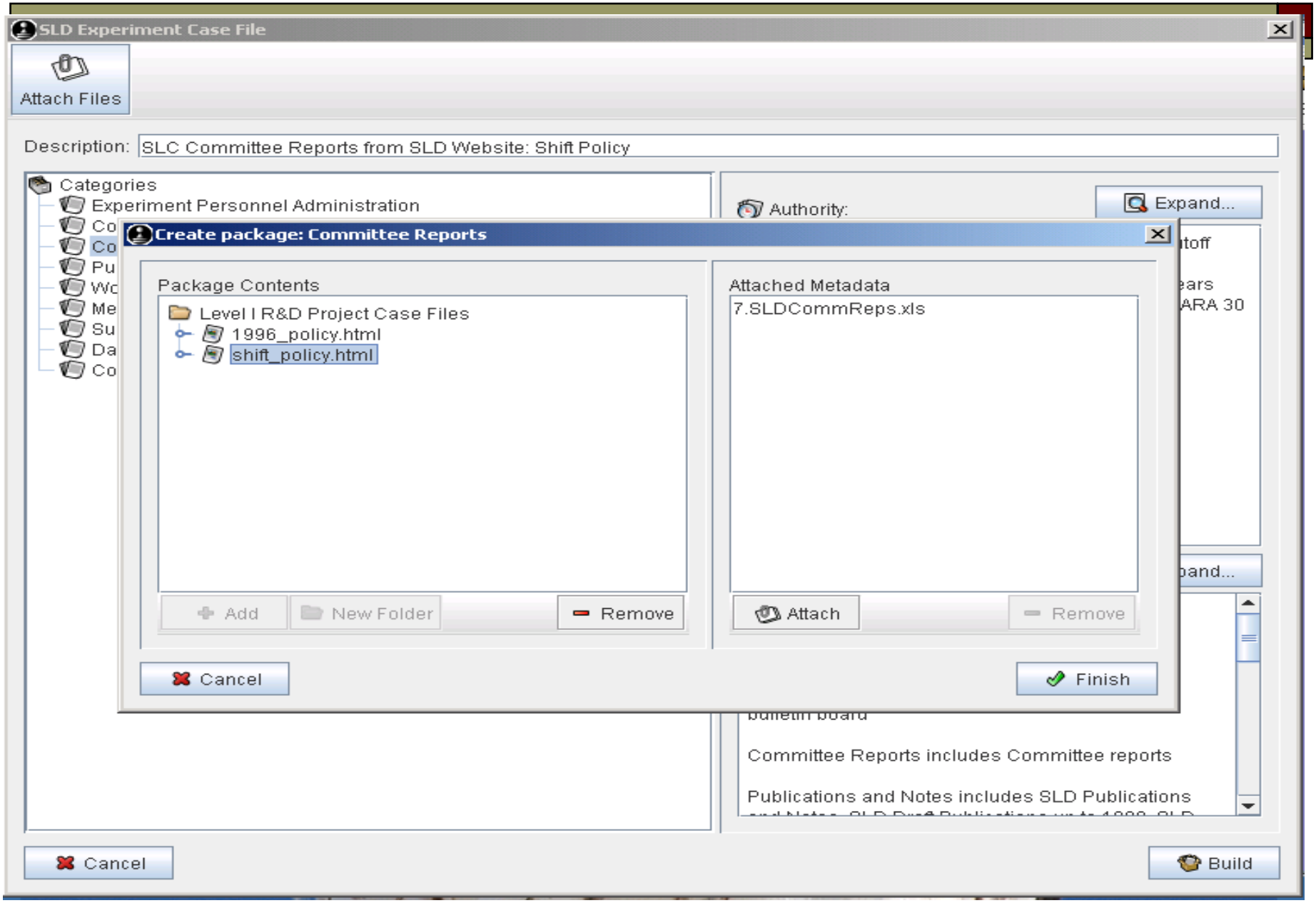

SLD Experiment Case File
Attach Files
Description:
SLC Committee Reports from SLD Website: Shift Policy
Categories
Experiment Personnel Administration
Authority:
Expand...
Create package: Committee Reports
Package Contents
Level I R&D Project Case Files
Attached Metadata
Add
New Folder
Remove
Attach
Remove
Cancel
Finish
Committee Reports includes Committee reports
Publications and Notes includes SLD Publications
Cancel
Build

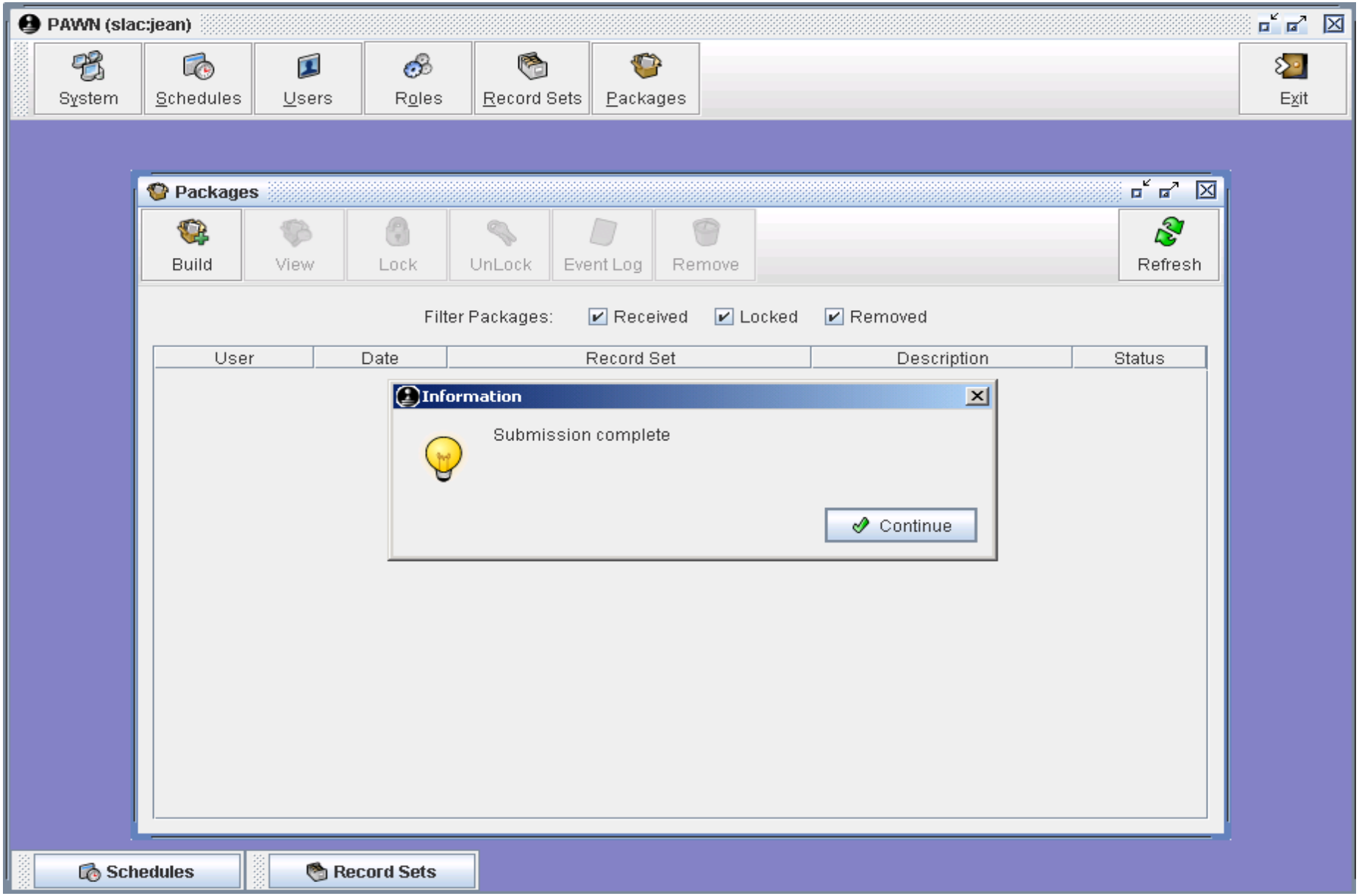

SLD Experiment Case File
Attach Files
Description:
SLC Committee Reports from SLD Website: Shift Policy
Categories
Experiment Personnel Administration
Authority:
Expand...
Open
Look In:
My Documents
My Music
My Pictures
client.htm
desktop.ini
launch.jnlp
SoiT-wc.png
File Name:
Files of Type:
All Files
Open
Cancel
Package
Remove
Finish
bulletin board
Committee Reports includes Committee reports
Publications and Notes includes SLD Publications
Cancel
Build

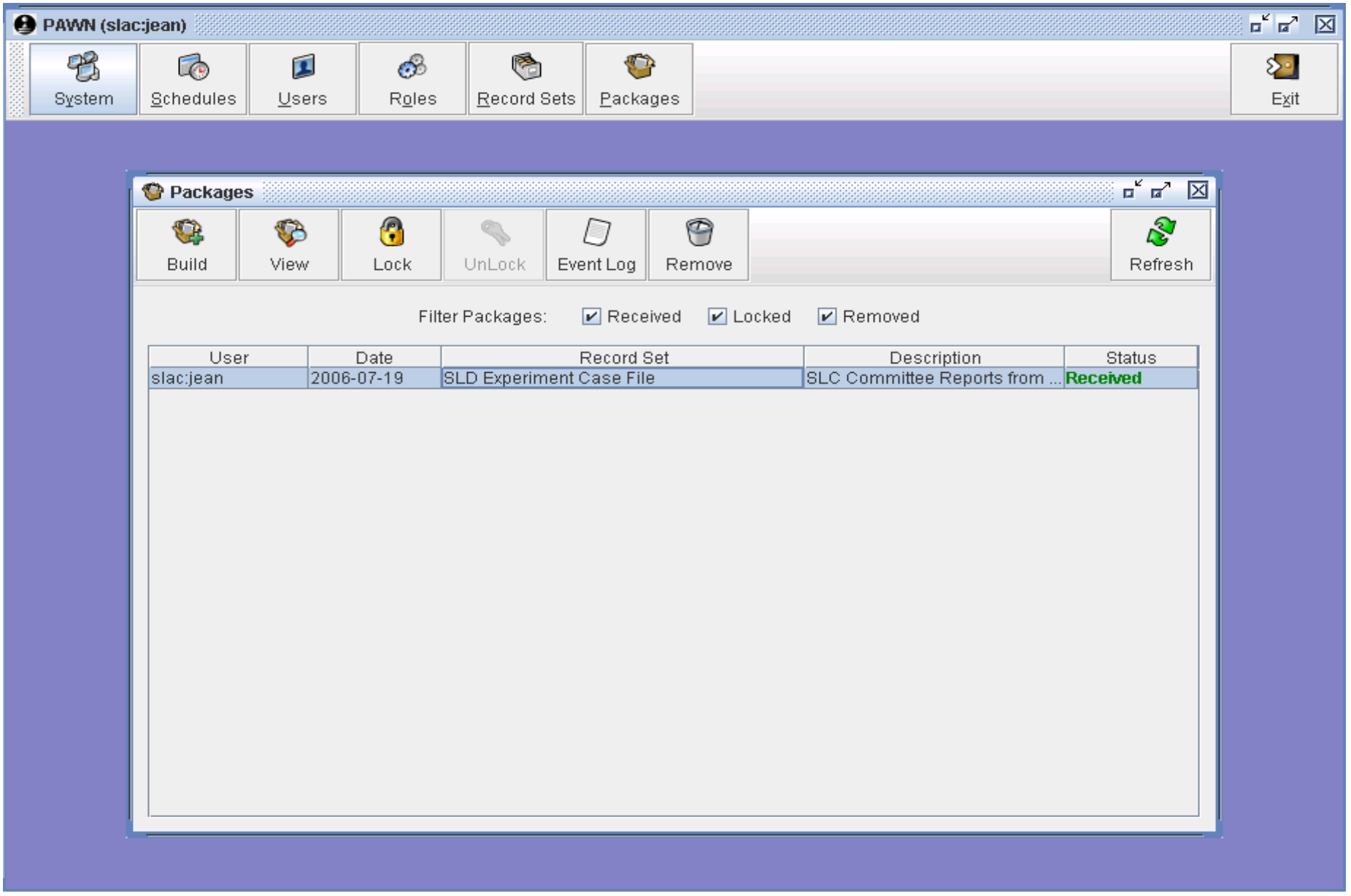

SLD Experiment Case File
Attach Files
Description:
SLC Committee Reports from SLD Website: Shift Policy
Categories
Experiment Personnel Administration
Authority:
Expand...
Open
Look In:
CommitteeReports
images
1996_policy.html
8mm.html
_coller.html
ctf.html
datacat.html
gdb.html
gnats.html
mainlist.html
maintainers.html
radioshack.html
reports.html
shift_policy.html
sld-gnats-mail.html
sld-working.html
sldext.html
stacker.html
File Name:
shift_policy.html
Files of Type:
All Files
Open
Cancel
Package
Remove
Finish
Committee Reports includes Committee reports
Publications and Notes includes SLD Publications
Cancel
Build

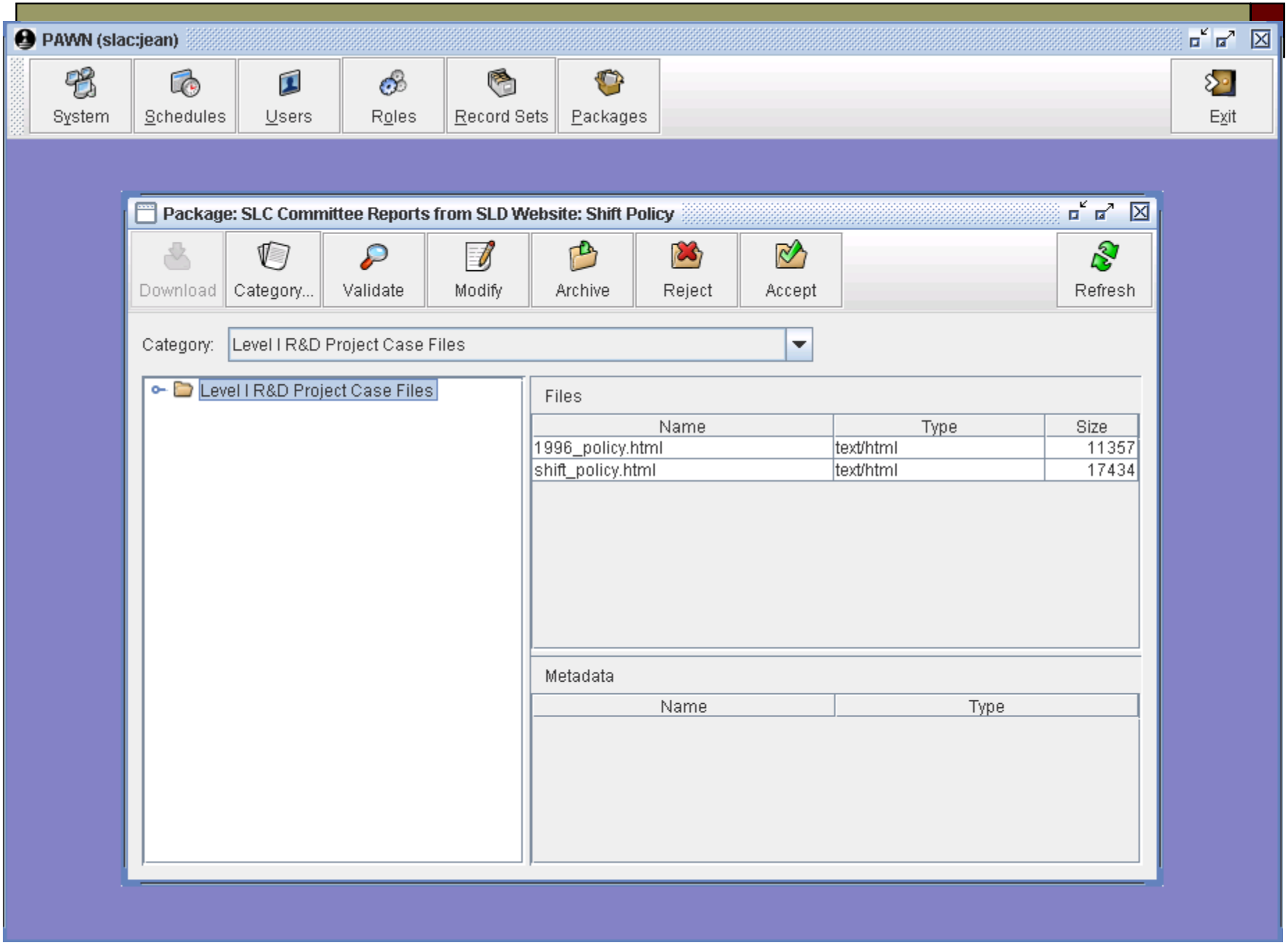

SLD Experiment Case File
Attach Files
Description:
SLC Committee Reports from SLD Website: Shift Policy
Categories
Experiment Personnel Administration
Authority:
Expand...
Create package: Committee Reports
Package Contents
Level I R&D Project Case Files
1996_policy.html
shift_policy.html
Attached Metadata
Add
New Folder
Remove
Attach
Remove
Cancel
Finish
Committee Reports includes Committee reports
Publications and Notes includes SLD Publications
Cancel
Build

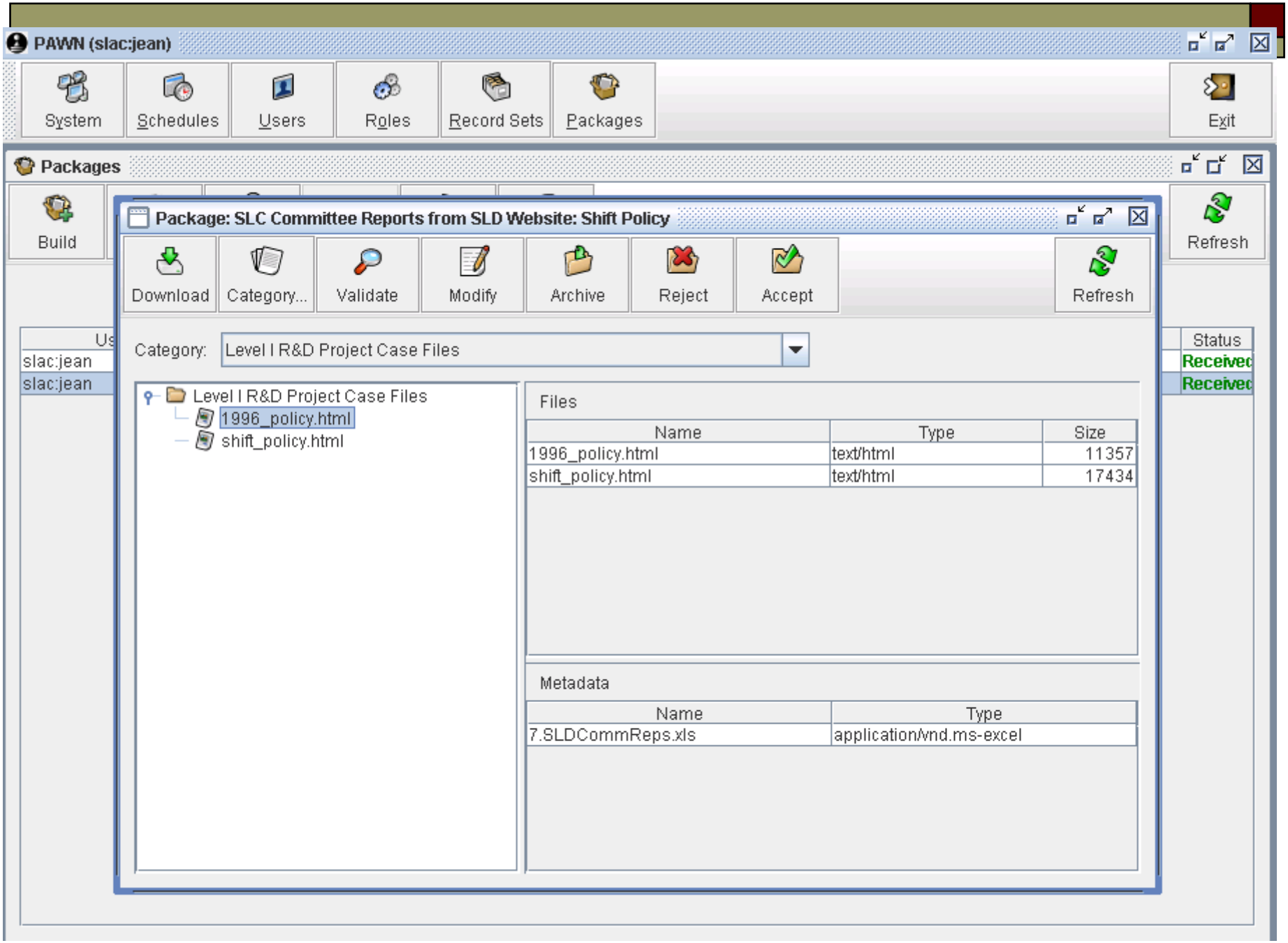

SLD Experiment Case File
Attach Files
Description:
SLC Committee Reports from SLD Website: Shift Policy
Categories
Experiment Personnel Administration
Authority:
Expand...
Open
Look In:
NARAGrant2006
DeTorcy
SDSCSRBdownloads
1.SLDInstitutes.htm
1.SLDInstitutes.xls
2.SLDCollabList.htm
2.SLDCollaboratorsList.xls
7.SLDCommReps.htm
7.SLDCommReps.xls
CoverMemoSept19.doc
entries.zip
FY 2006 SLAC DELIVERABLES.doc
FY 2006 SLAC DELIVERABLES.v2.doc
File Name:
7.SLDCommReps.xls
Files of Type:
All Files
Open
Cancel
Package
Remove
Finish
Committee Reports includes Committee reports
Publications and Notes includes SLD Publications
Cancel
Build


SLD Experiment Case File
Attach Files
Description:
SLC Committee Reports from SLD Website: Shift Policy
Categories
Experiment Personnel Administration
Authority:
Expand...
Create package: Committee Reports
Package Contents
Level I R&D Project Case Files
1996_policy.html
shift_policy.html
Attached Metadata
7.SLDCommReps.xls
Add
New Folder
Remove
Attach
Remove
Cancel
Finish
Committee Reports includes Committee reports
Publications and Notes includes SLD Publications
Cancel
Build


PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Packages
Build
View
Lock
UnLock
Event Log
Remove
Refresh
Filter Packages:
Received
Locked
Removed
User
Date
Record Set
Description
Status
Information
Submission complete
Continue
Schedules
Record Sets


PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Packages
Build
View
Lock
UnLock
Event Log
Remove
Refresh
Filter Packages:
Received
Locked
Removed
User
Date
Record Set
Description
Status
slac:jean
2006-07-19
SLD Experiment Case File
SLC Committee Reports from ...
Received


PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Package: SLC Committee Reports from SLD Website: Shift Policy
Download
Category...
Validate
Modify
Archive
Reject
Accept
Refresh
Category:
Level I R&D Project Case Files
Level I R&D Project Case Files
Files
Name
Type
Size
1996_policy.html
text/html
11357
shift_policy.html
text/html
17434
Metadata
Name
Type


PAWN (slac:jean)
System
Schedules
Users
Roles
Record Sets
Packages
Exit
Packages
Build
Refresh
Package: SLC Committee Reports from SLD Website: Shift Policy
Download
Category...
Validate
Modify
Archive
Reject
Accept
Refresh
Category:
Level I R&D Project Case Files
Level I R&D Project Case Files
1996_policy.html
shift_policy.html
Files
Name
Type
Size
1996_policy.html
text/html
11357
shift_policy.html
text/html
17434
Metadata
Name
Type
7.SLDCommReps.xls
application/vnd.ms-excel
Status
slac:jean
slac:jean
Received
Received

Conclusions / Future issues …

- Metadata
  - What
  - Where
  - How
- Generalize lessons to other collections
  - GLAST project (NASA records schedule)
  - Non-science archival collections
    - First US Web site
    - Public Affairs photographs

For further information….

- TPAP at SLAC: http://www.slac.stanford.edu/history/projects.shtml
- PAWN:
  - http://www.umiacs.umd.edu/research/adapt
  - Or "I'm feeling lucky" Google keywords: ADAPT UMIACS *

*A Digital Approach to Preservation Technology,
University of Maryland Institute for Advanced Computer Studies

Contact information….

Jean Marie Deken

Archivist and Head

Archives and History Office

Stanford Linear Accelerator Center

jmdeken@slac.stanford.edu