9th ETD Conference - 2006

A Prototype for Preservation and Harvesting of International ETDs using LOCKSS and OAI-PMH

Kamini Santhanagopalan
Department of Computer Science, Virginia Tech
ksanthan@vt.edu

Dr. Edward A. Fox
Department of Computer Science, Virginia Tech
fox@vt.edu

Prof. Gail McMillan
Digital Library and Archives, University Libraries, Virginia Tech
gailmac@vt.edu

Acknowledgements

I would like to thank Dr. Edward A. Fox for being my advisor and guide. I am grateful to him for his continuous support and invaluable inputs he has been providing me through the development of the project. This work would not have been possible without his support and encouragement.

I would like to express my sincere thanks to Prof. Gail McMillan for being my client and project leader in the place where I'm working as a graduate assistant. I am grateful to her for her continuous support, invaluable inputs and encouragement. This work would not have been possible without her smile, support and encouragement.

I would like to thank all the collaborators for this project (the six universities) who had agreed to participate in this project of testing the prototype of preservation and harvesting of International ETDs. Without their co-operation, this project would not have been successful.

I would also like to thank Mr. Thomas Robertson and Mr. Seth Morabito of Stanford University Libraries for their extended help in testing the plug-in implementation and hosting the collections in their servers.

I would like to thank my parents and all of my dear friends who have been rendering continuous moral support, encouragement, and helping me complete the task successfully.

Abstract

Students, faculty, researchers, and administrators desire, and in some cases require, that ETDs be preserved and made accessible for long periods of time, often 50 years or more, with the goal of support for the very long term. Part of the challenge of digital data preservation is for the "bits to be preserved", in spite of problems with infrastructure or disaster. Toward that end, the LOCKSS software (Lots of Copies Keep Stuff Safe, http://lockss.stanford.edu/) supports replication, distribution, and reliable preservation of content. Since 2002, discussions have been underway regarding the use of LOCKSS in connection with NDLTD. The following six universities

- Pontifícia Universidade Católica do Rio de Janeiro, Brazil
- Humboldt-Universität, Germany
- University of Cape Town, South Africa
- Florida State University, USA
- Georgia Tech, USA
- Virginia Tech, USA

are collaborating to demonstrate the functionality of the LOCKSS and OAI-PMH for international preservation of ETDs. This report describes in detail the implementation for preservation and harvesting of International ETDs using LOCKSS tool. An analysis of the test results is also presented.

TABLE OF CONTENTS

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

1. Introduction

1.1 Background

Digital Data Preservation is an important term used to describe the safeguarding of a digital resource into the distant future. The main goal of digital preservation is to ensure that the digital information remains readable and useable in the future. For any digitization project, one should ensure that only globally recognized standards [1] are implemented in the digitization methodology.

Hence, a establishing and implementing a digital preservation strategy (which is reliable for decades to come) must be considered one of the vital aspects of a digitization project. As we know, the earlier the issues of digital data preservation are tackled, the easier it is to make the preservation strategy that matches the project specifications [1]. Data retrieval should continue to be supported across platforms into the future.

1.2 Motivation

The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organization which aims to promote the creation, use, and preservation of electronic theses and dissertations in lieu of the traditional paper-based theses and dissertations. Traditional paper based theses and dissertations are difficult to manage, use and preserve. Hence, many institutions in the world are advocating ETD submissions to individual students. We need a standard documentation especially for the preservation of ETDs.

1.2.1 Existing Preservation Techniques

Preservation of digital data involves the retention of the information object and its meaning. Hence, it is necessary for preservation techniques to be able to understand and re-create the original form [2] of the information to make sure that it is authentic and accessible. The art of preservation of digital information is complex due to the fact that the information has a dependency factor on its technical environment [3]. A consequential issue in digital persistence is that the one should ensure the integrity of information over decades and through the software development life cycle [6].

How are digital data stored? Digital resources can be stored on any medium such as a CD-ROM or a DVD (these media can represent the data's binary digits or bits). Hence, digital preservation involves copying the information into a newer media before the old media becomes stale or unusable. This is called copying or refreshing [4]. But, simply copying the digital data into such medias is not enough for preserving them. One should be able to retrieve this information in the future and process it further, if needed.

Different countries have different preservation strategies. Most of the existing preservation strategies involve replication of contents that are to be preserved (which are usually done in a central server). This is explained in the table below.

Table 1 – Preservation Strategies used in different countries

Country	**Preservation Strategy/Technique used**
India	1. The digital collection is stored in floppy drives, CDs and Hard disk drives. 2. The goal is to achieve a collaborative digital library [7].
Brazil	1. Lacks standards-related preservation activities 2. Scientific data is being stored in information centers' databases that are proprietary in nature [8]. 3. Current practices are done by using traditional preservation technique (paper based)
South Africa	1. Digital Imaging Project of South Africa - DISA [10] implement digital technologies to so that scholars and researchers from around the world can access South African material of high socio-political interest 2. No concrete preservation training till now
Germany	1. Preservation strategies are being planned but, not implemented till now 2. The technical infrastructure is under implementation

1.2.2 Digital Data Preservation

OAIS is a reference model for an *Open Archival Information System.* It describes all the functions of a digital repository such as how digital objects can be prepared, submitted to an archive, stored for long periods, maintained, and retrieved as needed [11]. But, it does not address specific technologies, archiving techniques, or types of content.

A checksum is a form of redundancy check which is a simple measure of protecting data integrity. Preserving the integrity of data is very important in any preservation technique. We should make sure that the digital data bits are not corrupted. Checksum techniques add up the basic components of a message and store the resulting value. The same operation is performed on the data bits; the results are compared to the authentic checksum, and if the sums match, it means that the message was not corrupted. MD5 (Message Digest algorithm 5) is a commonly used cryptographic hash function used for checking the integrity of files (digital data). MD5 hashes are used to provide some assurance to the users that the downloaded file has not been altered.

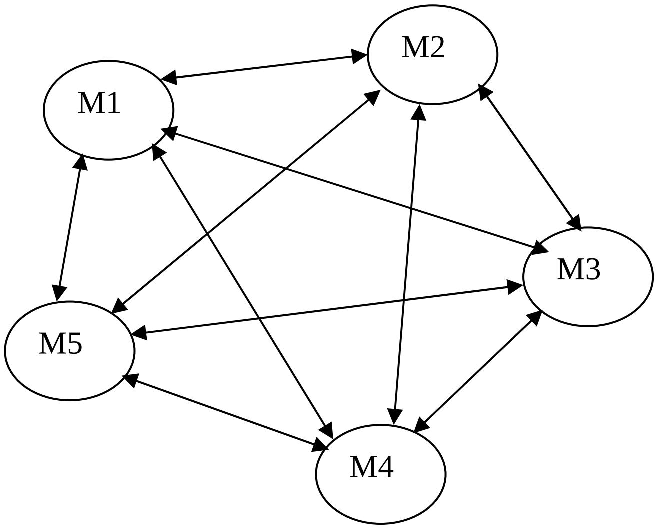
1.2.3 What is LOCKSS?

LOCKSS (Lots of Copies Keep Stuff Safe) is a peer to peer digital preservation system. It is an open source software [5] which provides an inexpensive way to collect, store and preserve the local copy of authorized content. LOCKSS does not require so much of a technical administration or hardware requirements. Thus, LOCKSS converts a PC into a digital preservation machine (not just a place for backup).

There are four main functions [5] to be performed by a library that uses the LOCKSS software. They are given below.

- It should collect the published content from the e-journals by using a web-crawler (similar to internet search engines)
- It should constantly compare the collected content with the contents collected by other machines, detect and repair differences (damaged or missing pages) if found
- It should act as a web proxy and provide access to the preserved contents
- It should provide a web-based administrative interface that lets the librarians to include new journals for preservation, monitor the existing journals, and provide controlled access to the journals which are being preserved.

In the figure below, M1, M2, M3, M4, and M5 represent different universities (or institutions) who wish to preserve digital data. They are all connected to each other. In this preservation technique, the contents of one university are being preserved in all the 5 locations. For instance, M1's collections are being harvested and preserved in M1, M2, M3, M4, and M5. By doing this, we create multiple copies of the same digital data in different locations. Hence, this becomes a distributed preservation technique. The main advantage of a distributed preservation technique is that - the data is not stored in a centralized server. Hence, loss of the collection in one or 2 of the servers will not affect the preservation, since there are copies in other servers. This is illustrated in the figure below.



Points to be noted:

1) Nodes M1 through M5 represents participating universities.
2) Preservation – we have multiple copies here; NOT just a backup.

Figure 1.01

LOCKSS tool is built using Java, and it helps write plug-ins for harvesting and preserving the digital data collection. The plug-in is generated in XML format. The following table explains the meaning of some of the commonly used terms in digital data preservation using LOCKSS.

Table 2 – Technical terms in using the LOCKSS tool

Some important terms	Meaning of the term with respect to LOCKSS and preservation
AU – Archival Unit	It is a unit of digital data which is to be preserved. This could be old or new e-journals, ETDs, etc.
Plug-in Name	It is a user understandable name given to the plug-in (for easy identification).
Publisher Manifest page (or, Manifest page)	1. For LOCKSS to identify that the digital data (e-journal/ETDs) is to be preserved, it needs a permission page which gives rights to LOCKSS to preserve the collection. 2. This is just a static HTML page with the permission statement written on it, with the link to the BASE_URL
BASE_URL	1. The URL which is some kind of a home page for the digital collection (ETDs or e-journals) 2. For example, the e journal collection “Virginia Libraries” has http://scholar.lib.vt.edu/ejournals/VALib/ as its BASE_URL.
Crawl rules	These are rules are written using Regular Expression, which helps the LOCKSS tool to identify the files to be crawled and harvested. This is similar to the search engines.

2. A Tutorial for creating Plug-ins using LOCKSS tool – For Developers

(http://scholar.lib.vt.edu/lockss/introduction.htm)

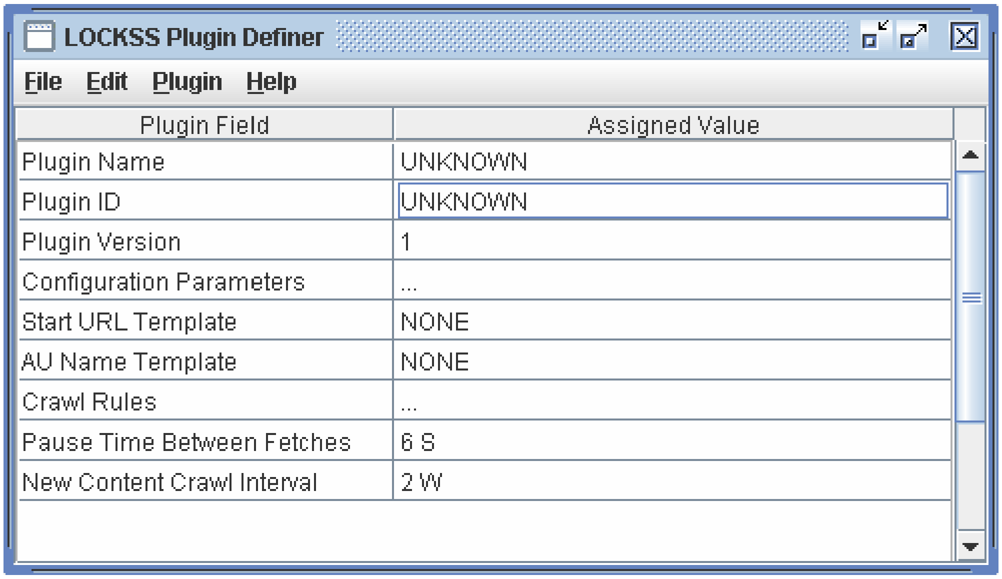
2.1 Introduction to LOCKSS Plug-in Generation Tool

The LOCKSS plug-in generation tool provides a user interface that allows the user to create and test a plug-in for simple and complex journals. The user input includes the details such as name of the plug-in, name of the plug-in ID (which is taken as the Java class name), name of the collection/journal, the URL from which it should be fetched, extra parameters (if any), pause time between fetches, etc. The tool outputs a definition of the plug-in for the chosen journal in XML. The next section starts with the tutorial explaining how to create plug-ins with an assumption that LOCKSS is being installed in your machine.

2.2 The Main window of LOCKSS Plug-in Definer:

The following screen shot shows how the LOCKSS tool would look like when you execute the batch file '**runtool.bat**' from the corresponding location in the command prompt. There are totally 9 fields in the LOCKSS plug-in definer tool (given below). The following sub-sections discuss these parameters in depth.

1. Plug-in Name
2. Plug-in ID
3. Plug-in Version
4. Configuration Parameters
5. Start URL Template
6. AU Name Template
7. Crawl Rules
8. Pause Time Between Fetches
9. New Content Crawl Interval



Plugin Field	Assigned Value
Plugin Name	UNKNOWN
Plugin ID	UNKNOWN
Plugin Version	1
Configuration Parameters	...
Start URL Template	NONE
AU Name Template	NONE
Crawl Rules	...
Pause Time Between Fetches	6 S
New Content Crawl Interval	2 W

Figure 2.01

2.2.1 Plug-in Name

The name that the plug-in will appear under in the configuration menu of the LOCKSS user interface. Spaces are allowed.

Example: VirginiaLibraries

2.2.2 Plug-in ID

The Java class name for the plug-in. The package structure should be maintained consistently.

Example: edu.vt.lib.plugin.VirginiaLibraries

2.2.3 Plug-in Version:

A version number that differentiates different versions of the same plug-in. It should start at 1.

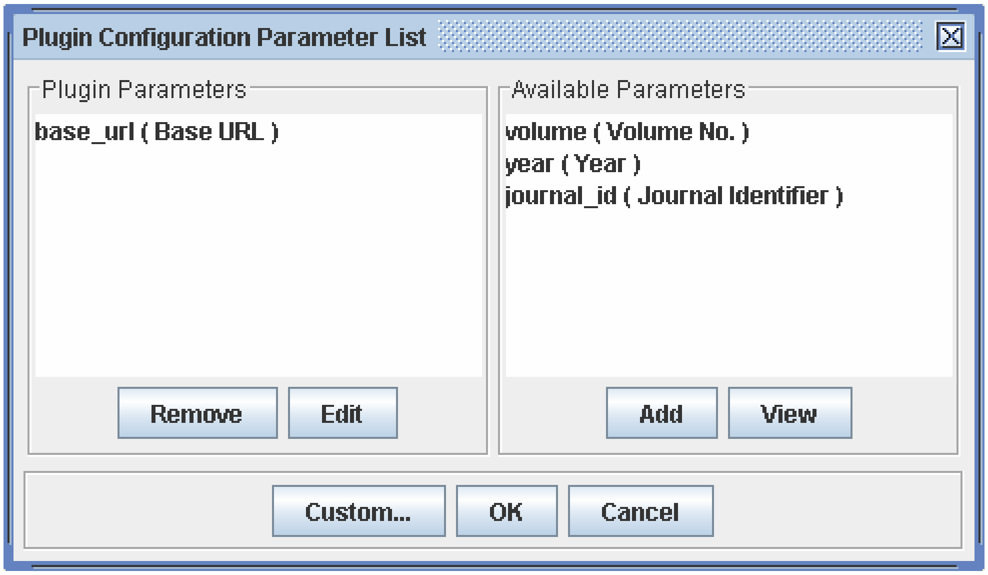
Example: 1

2.2.4 Configuration Parameters:

This defines the set of parameters that the plug-in will use to identify and differentiate between each AU that uses the plug-in.

1. Example: The following screenshot represents the Configuration Parameters window. Consider the e-journal, Virginia Libraries: The Base URL being http://scholar.lib.vt.edu/ejournals/VALib/

2. In this website, we find that the journal issues are classified or grouped in terms of 'Volume Numbers'. Since we need to crawl the contents related to Volume numbers and issues, we can add the '**Volume (Volume No.)**' parameter from the available parameters section (in addition to the Base URL which is already added) to the Plug-in Parameters section. (see Figure 3.02)



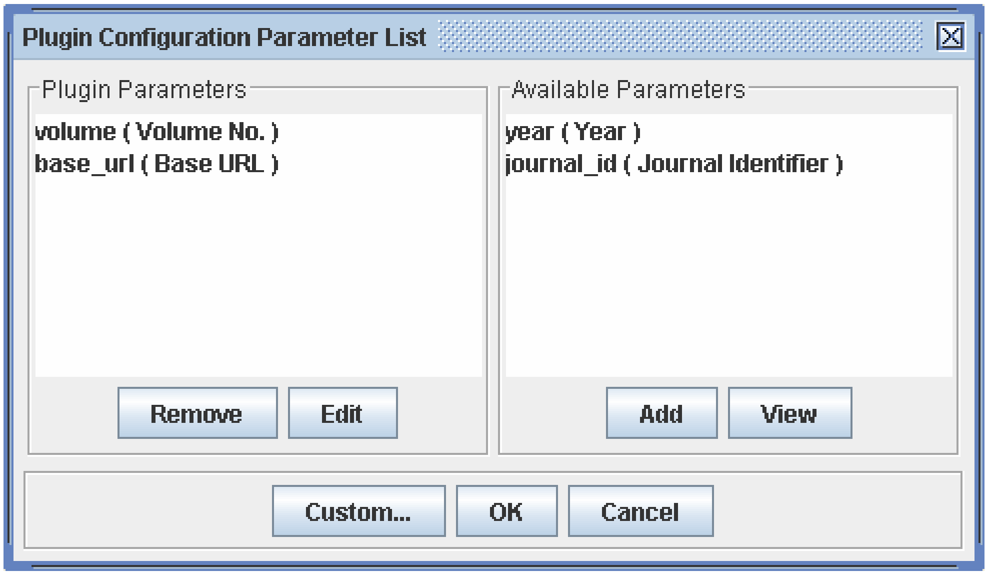
Plugin Parameters:

base_url (Base URL)

Available Parameters:

volume (Volume No.)
year (Year)
journal_id (Journal Identifier)

Figure 2.02



Plugin Parameters:

volume (Volume No.)
base_url (Base URL)

Available Parameters:

year (Year)
journal_id (Journal Identifier)

Figure 2.03

2.2.5 AU Name Template:

This field is used to create the name of each Archival Unit (AU) using the plug-in used by the LOCKSS administrative UI.

1. In the LOCKSS plug-in definer window, click on the NONE beside AU Name template. A new window as shown below would pop up.

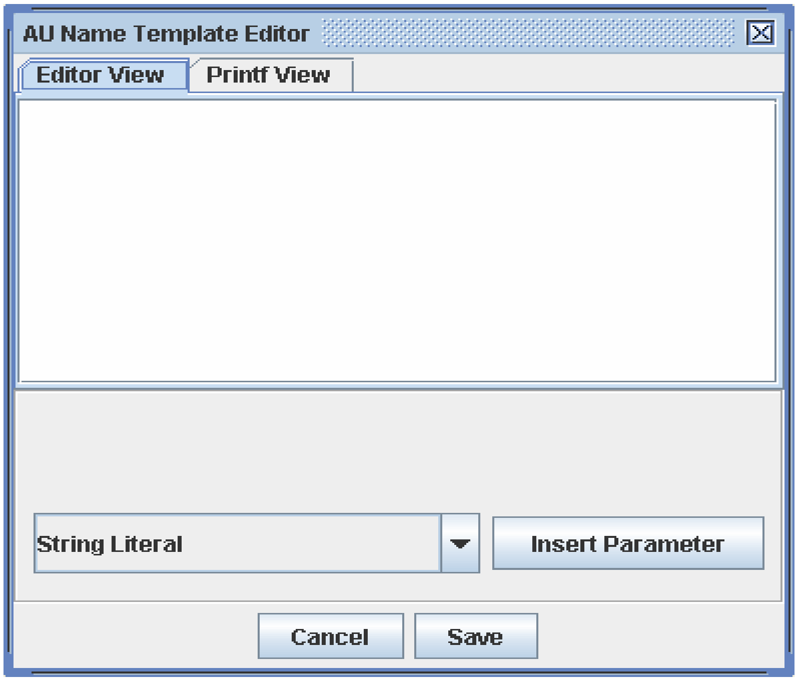
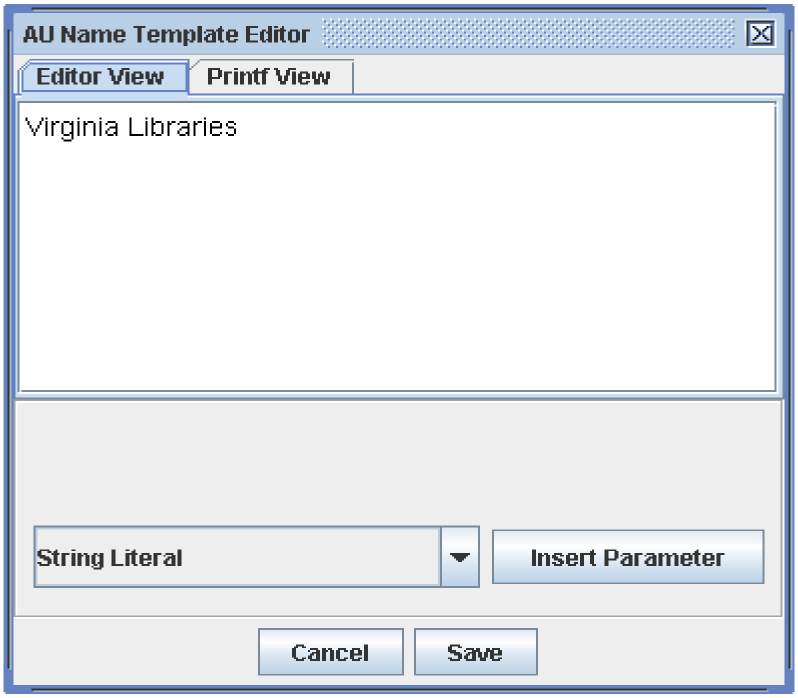


Figure 2.04

2. Select 'String Literal' from the combo box and type in the corresponding name (here, Virginia Libraries) in the dialog box. Now, click OK in the dialog box and click on the 'Save' button.
3. *Alternate condition:* If you have a unique BASE_URL and want to display that as the name instead of any string literal, select 'BASE_URL' from the combo box and click on the 'Save' button. Similarly, if you want to display the BASE_URL followed by the parameter 'Volume Number' or 'Year', select the corresponding parameter from the combo box and click on the 'Save' button.



Virginia Libraries

Figure 2.05

2.2.6 Start URL Template:

This template tells the plug-in where to find the *publisher manifest page* for each AU of the journal.

1. In the LOCKSS plug-in definer window, click on the NONE beside Starting URL Name template. A new window as shown below would pop up.

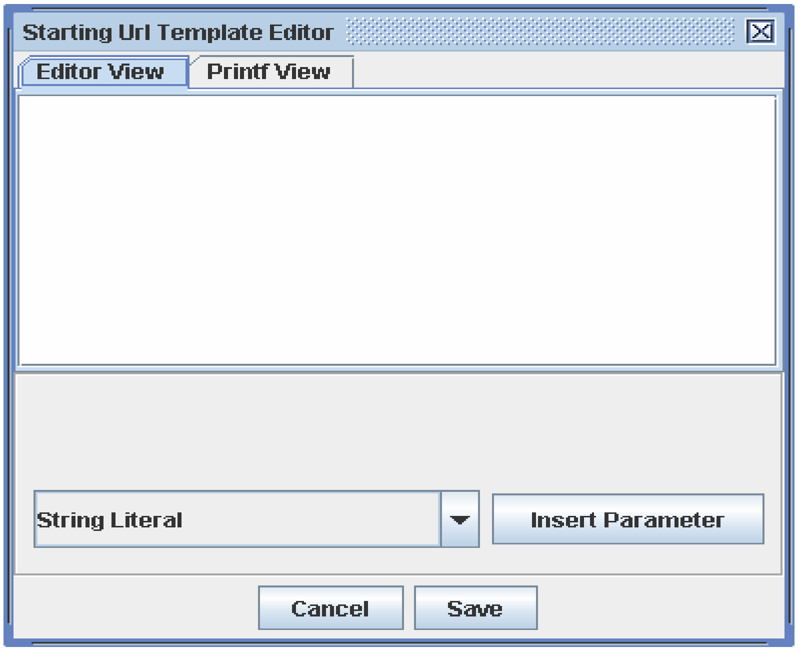
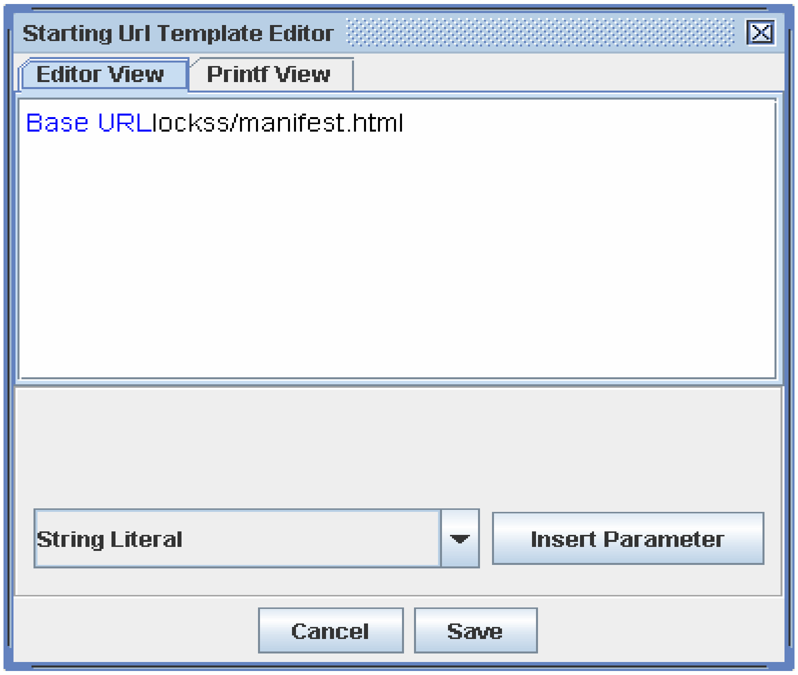


Figure 2.06

2. Select BASE_URL from the combo box and click on 'Insert Parameter' button.
3. After the BASE_URL gets inserted, select 'String Literal' from the combo box, and click on 'Insert Parameter'. A new dialog box would pop up and the exact location in which the publisher manifest page lies should be entered there. Save the changes.
    a. Example: In this case, the publisher manifest page is 'manifest.html' and it its location is http://scholar.lib.vt.edu/ejournals/VALib/lockss/manifest.html



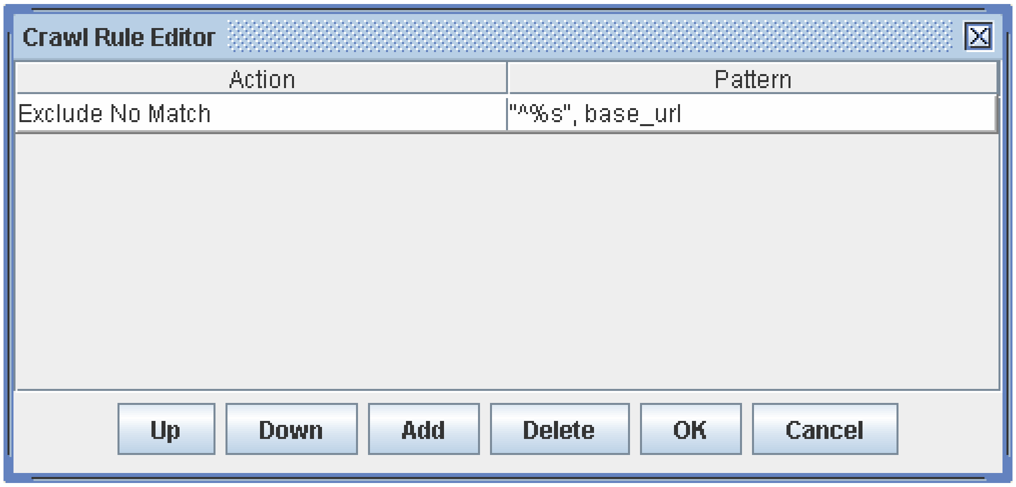
Base URLlockss/manifest.html

Figure 2.07

2.2.7 Crawl Rules:

The crawl rules define the boundaries of an AU in the journal's web site. An AU is normally a year's run or a volume of the journal.

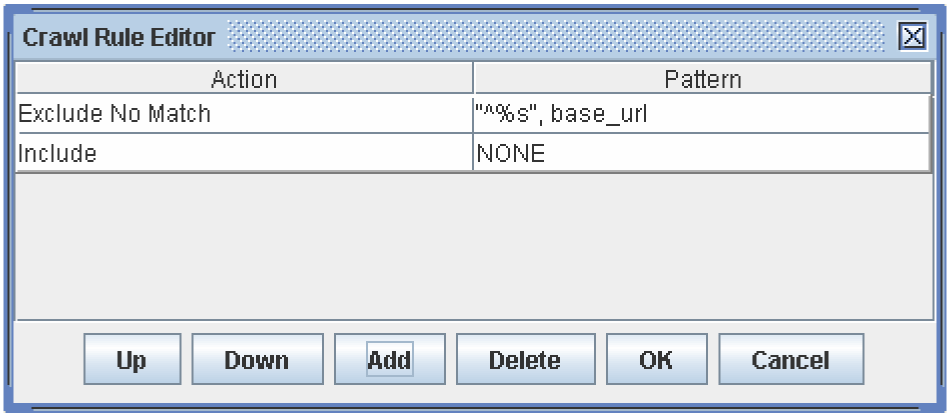
1. In the LOCKSS plug-in definer window, click on the '…' beside 'Crawl Rules' cell. A new window as shown below would pop up.



Action	Pattern
Exclude No Match	"^%s", base_url

Figure 2.08

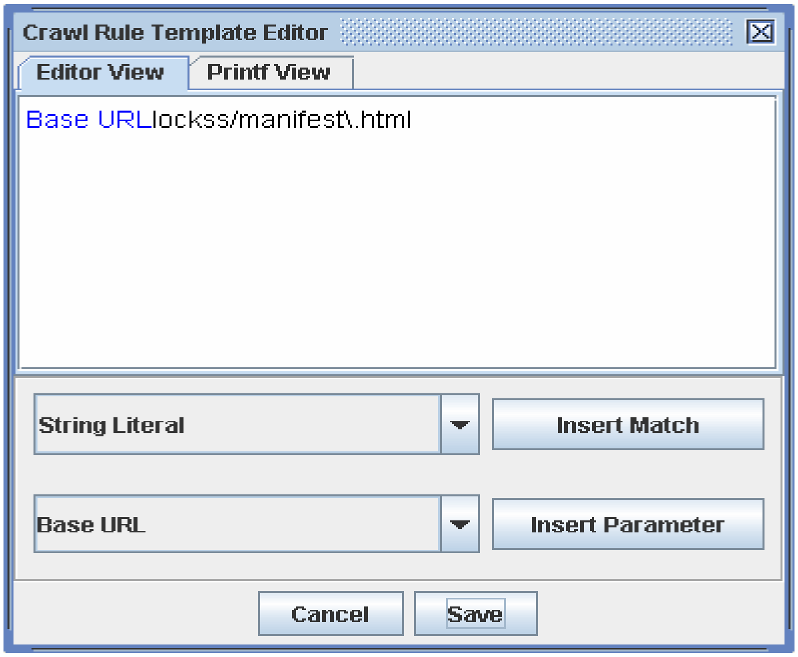
2. In this Crawl Rule editor, you must start entering your rules for crawling. You might need to understand a few things about Regular Expressions. Add a new rule by clicking on the 'Add' button. Now, click on the pattern 'NONE' beside the action 'Include'.



Action	Pattern
Exclude No Match	"^%s", base_url
Include	NONE

Figure 2.09

3. A new window as shown below would pop up, allowing you to enter or define the crawl rule. Select BASE_URL from the combo box and do an Insert Parameter. Then, as before type in the String Literal (exact location in which the publisher's manifest page is located) in the dialog box and save the changes.



Base URLlockss/manifest\.html

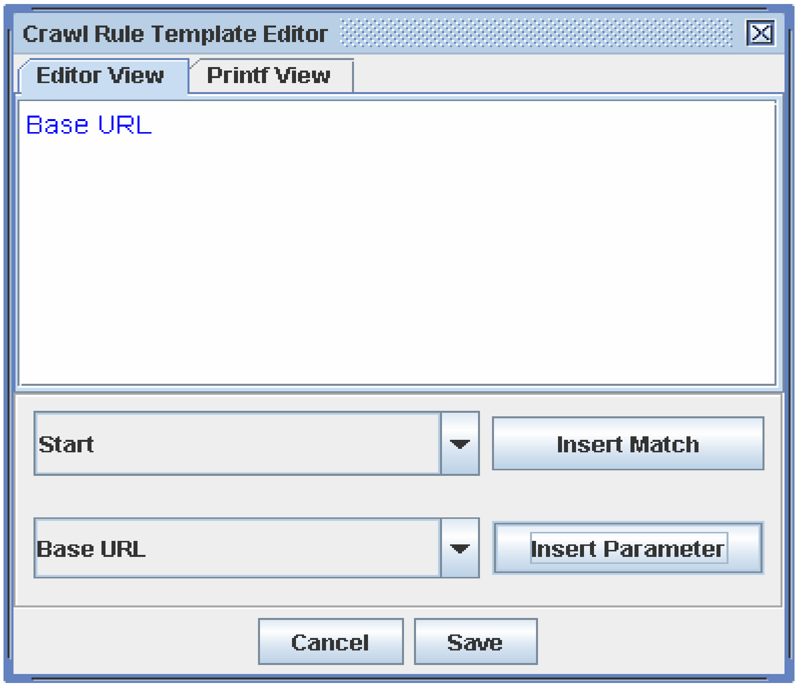
String Literal

Base URL

Figure 2.10

a. Now, let us see how to include rules for a complex structure containing volume numbers. In the URL for VirginiaLibraries (http://scholar.lib.vt.edu/ejournals/VALib/), consider the PDF file which is available in the location given by http://scholar.lib.vt.edu/ejournals/VALib/v50_n2/v50n2.pdf.

b. In the above PDF location, the BASE_URL is followed by a series of characters, which can be represented using regular expressions. First select the BASE_URL from the combo box and insert the parameter.



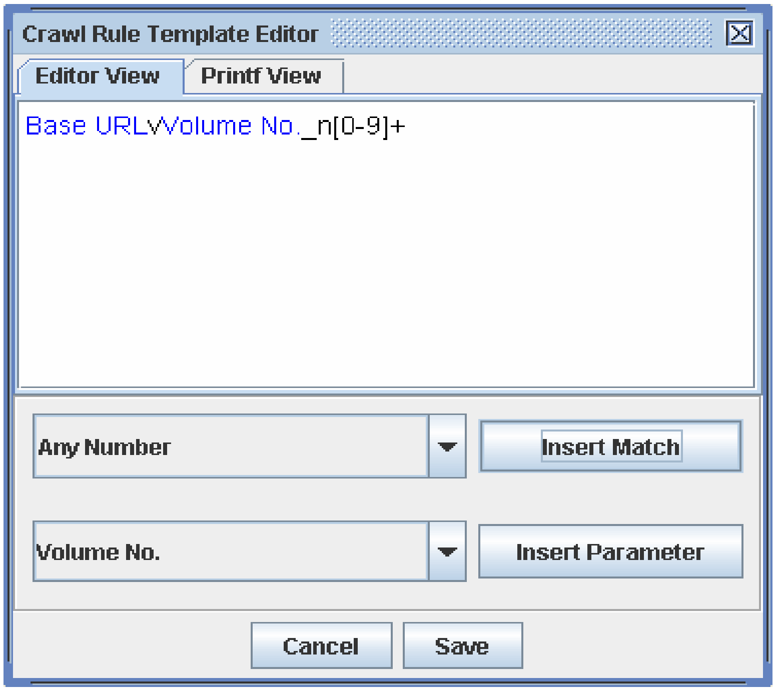
Base URL

Base URL

Figure 2.11

c. The second step is to enter a String Literal ‘v’, followed by the insertion of the Volume Number parameter. (When you insert the Volume number, a new window would pop up, asking you to specify the padding value. Give OK for the default padding value of ‘0’.)

d. The third step is to continue the rule – Having an underscore (‘_’) followed by ‘n’ as the string literal. After inserting the String Literal, select ‘Any Number’ from the combo box and give an insert match. Now, [0-9]+ will appear on the window, which is the regular expression format for Any Number. This means that the ‘_n’ can be followed by any integer number from 0 to 9.



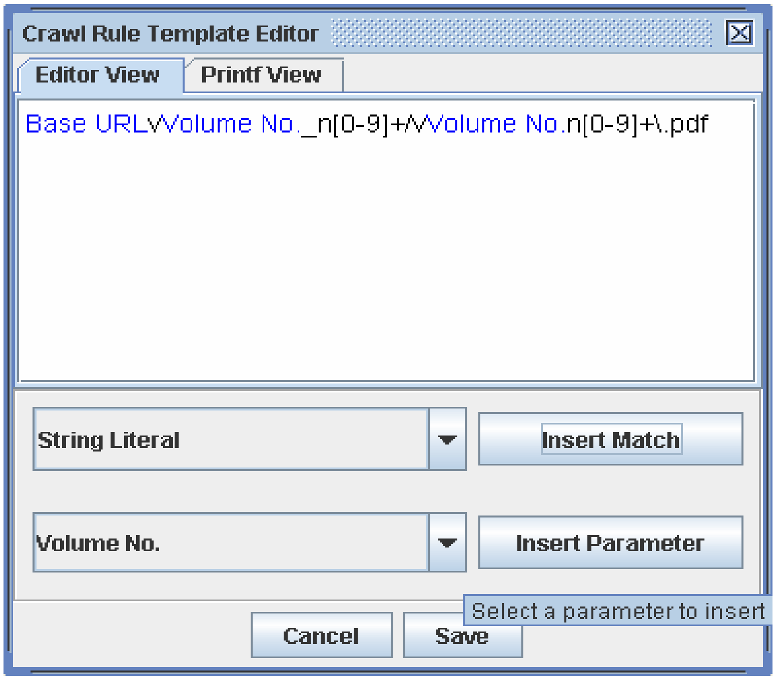
```
Base URLvVolume No._n[0-9]+
```

Any Number

Volume No.

Figure 2.12

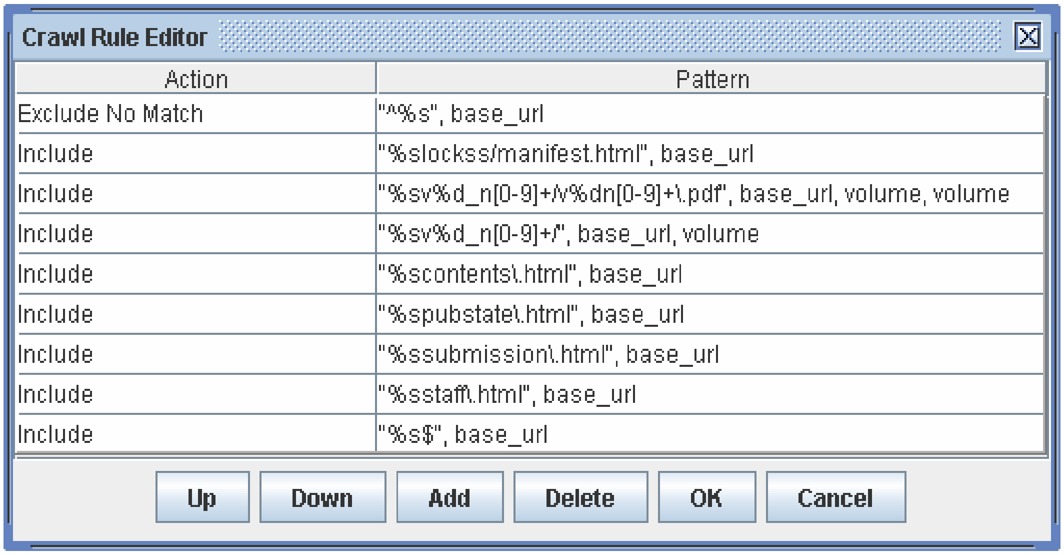
e. Finally, the string literals ‘/v’ followed by insertion of Volume Number, the string literal ‘n’, and ‘any number’ match has to be specified again, because the complete URL is http://scholar.lib.vt.edu/ejournals/VALib/v50_n2/v50n2.pdf. Now, enter the String literal ‘.pdf’ to specify the file type that has to be crawled. You will notice that a backslash is introduced before the dot. This is the representation of a dot in Regular Expressions format.



Base URL\/Volume No._n[0-9]+/\/Volume No.n[0-9]+\.pdf

Figure 2.13

f. Now, continue creating other rules to complete the Crawl rules section of the plug-in. The final Crawl rule template editor would look like the figure shown below.



Action	Pattern
Exclude No Match	"^%s", base_url
Include	"%slockss/manifest.html", base_url
Include	"%s\/%d_n[0-9]+/\/%dn[0-9]+\.pdf", base_url, volume, volume
Include	"%s\/%d_n[0-9]+/", base_url, volume
Include	"%scontents\.html", base_url
Include	"%spubstate\.html", base_url
Include	"%ssubmission\.html", base_url
Include	"%sstaff\.html", base_url
Include	"%s$", base_url

Figure 2.14

g. In the last rule shown above, we have added a $ to the end of that pattern (which will match the end of a string). If this is $ is not given, the locations of all the files would be fetched irrespective of the volume number given during the test run.

2.2.8 Pause Time between Fetches:

The time for which the LOCKSS daemon waits after fetching each page from the publisher's web site.

Example: 6S (This is usually set at 6 seconds)

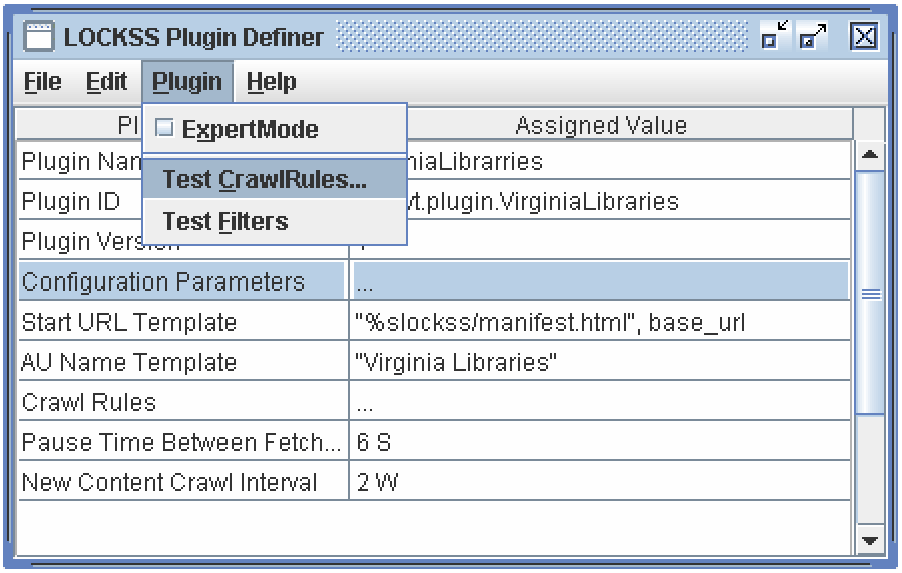
2.2.9 New Content Crawl Interval:

The time between attempts by the LOCKSS daemon to find new content on the publisher's web site.

Example: 2W (This is usually set as 2 weeks)

2.3 Results:

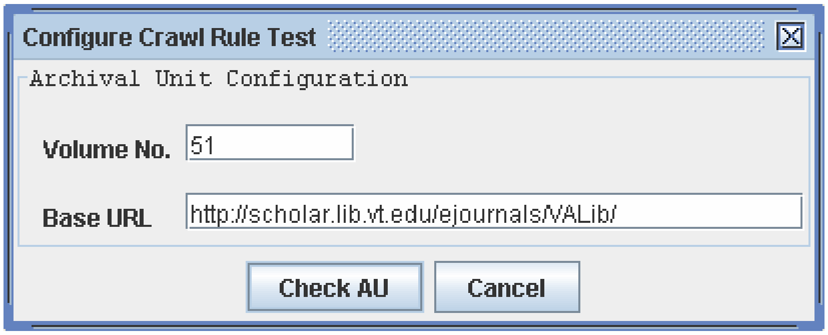
To view the crawl results, click on the ‘Plug-in’ pull down menu and click on “Test Crawl Rules” menu button, as shown in the screen shot.



Pl	Assigned Value
Plugin Nan	iaLibrarries
Plugin ID	t.plugin.VirginiaLibraries
Plugin Vers	
Configuration Parameters	...
Start URL Template	"%slockss/manifest.html", base_url
AU Name Template	"Virginia Libraries"
Crawl Rules	...
Pause Time Between Fetch...	6 S
New Content Crawl Interval	2 W

Figure 2.15

1. Now, a new window, 'Configure Crawl Rule Test' would pop up. Enter the details such the volume number for which you would like to view the test results and the BASE_URL. Click on Check AU button.



Configure Crawl Rule Test

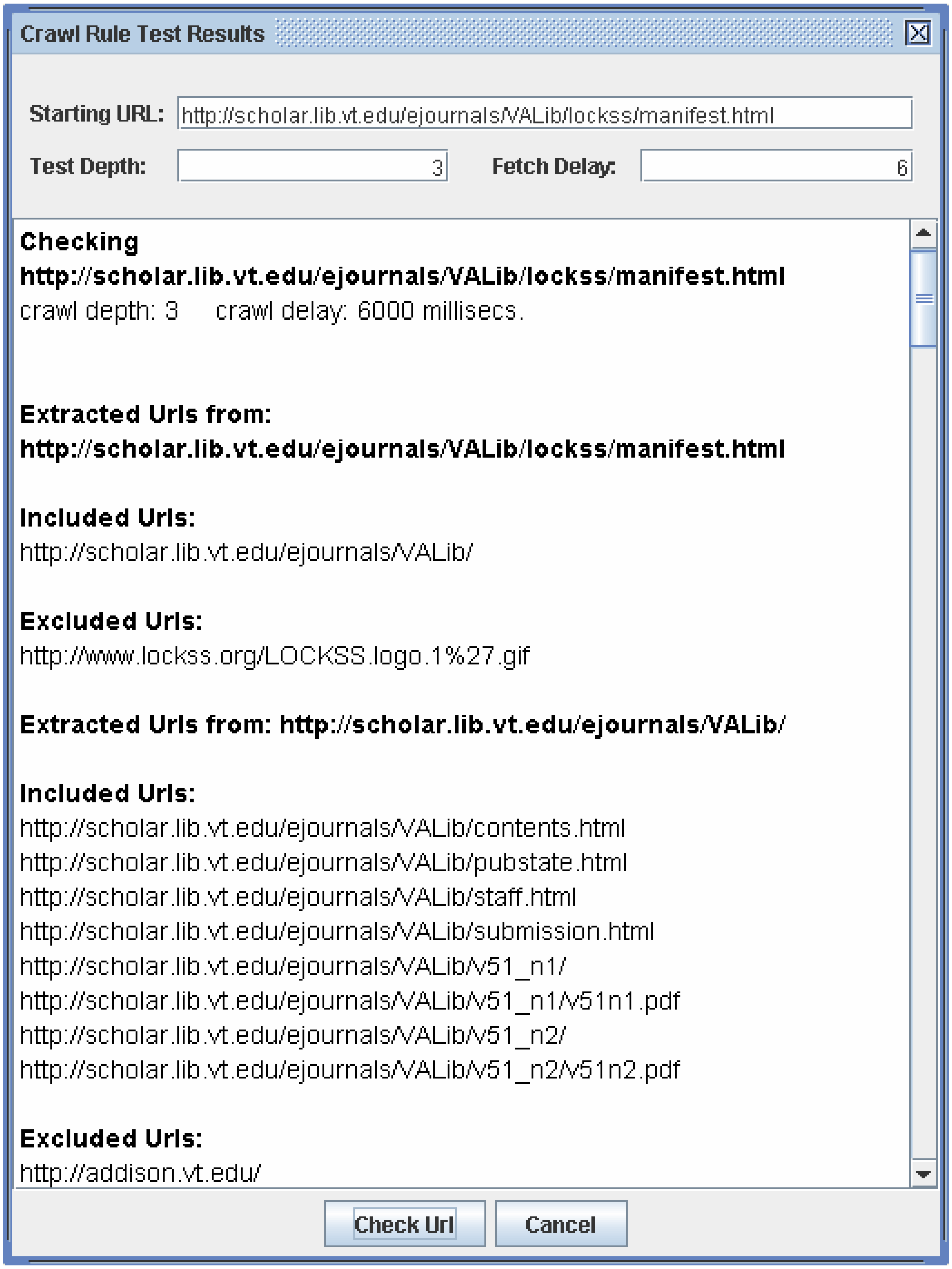
Archival Unit Configuration

Volume No. 51

Base URL http://scholar.lib.vt.edu/ejournals/VALib/

Figure 2.16

2. The results pane would be displayed in a new window, showing the test results. Scroll down to see detailed crawl results depending on the 'Test Depth' entered. (Note: While giving the Test Depth. Smaller integers are allowed, but, if you give large numbers like 15 or 20, a 'FileNotFound' Exception will be thrown at the backend.)



Starting URL: http://scholar.lib.vt.edu/ejournals/VALib/lockss/manifest.html

Test Depth: 3

Fetch Delay: 6

Checking http://scholar.lib.vt.edu/ejournals/VALib/lockss/manifest.html

crawl depth: 3   crawl delay: 6000 millisecs.

Extracted Urls from: http://scholar.lib.vt.edu/ejournals/VALib/lockss/manifest.html

Included Urls:

http://scholar.lib.vt.edu/ejournals/VALib/

Excluded Urls:

http://www.lockss.org/LOCKSS.logo.1%27.gif

Extracted Urls from: http://scholar.lib.vt.edu/ejournals/VALib/

Included Urls:

http://scholar.lib.vt.edu/ejournals/VALib/contents.html
http://scholar.lib.vt.edu/ejournals/VALib/pubstate.html
http://scholar.lib.vt.edu/ejournals/VALib/staff.html
http://scholar.lib.vt.edu/ejournals/VALib/submission.html
http://scholar.lib.vt.edu/ejournals/VALib/v51_n1/
http://scholar.lib.vt.edu/ejournals/VALib/v51_n1/v51n1.pdf
http://scholar.lib.vt.edu/ejournals/VALib/v51_n2/
http://scholar.lib.vt.edu/ejournals/VALib/v51_n2/v51n2.pdf

Excluded Urls:

http://addison.vt.edu/

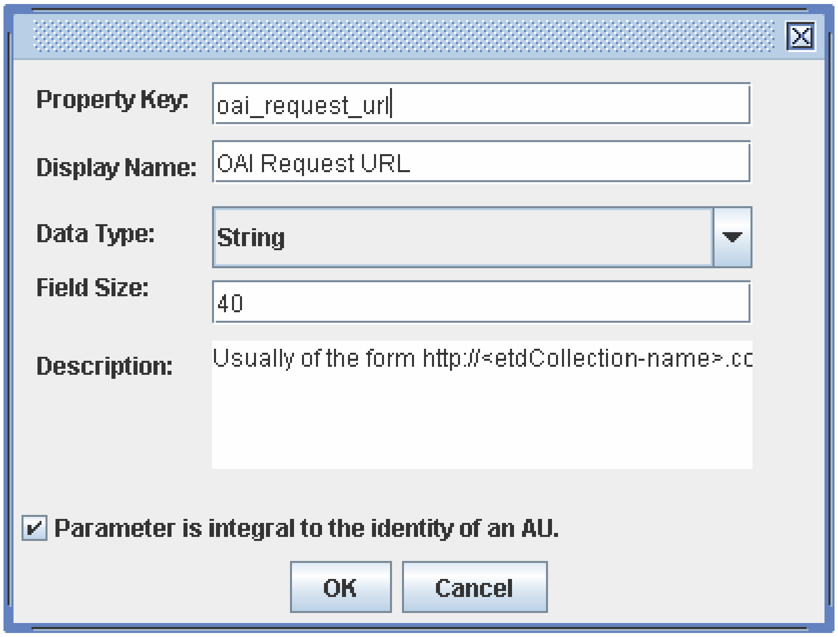
Figure 2.17

2.4 Writing OAI Plug-ins:

The current version of LOCKSS does not have a support for writing OAI plug-ins. So, we need to edit the XML file after we create the basic plug-in. We also need the OAI

Request URL for the plug-in. For that, we need to do some changes in section 2.2.4. The following steps explain the procedure for including OAI Request URL.

- In figure 3.1, click on "Custom" button. A new 'Custom Property Add' screen would be popped up. The values across each of the blanks are filled out as shown in the figure below.



Property Key: oai_request_url

Display Name: OAI Request URL

Data Type: String

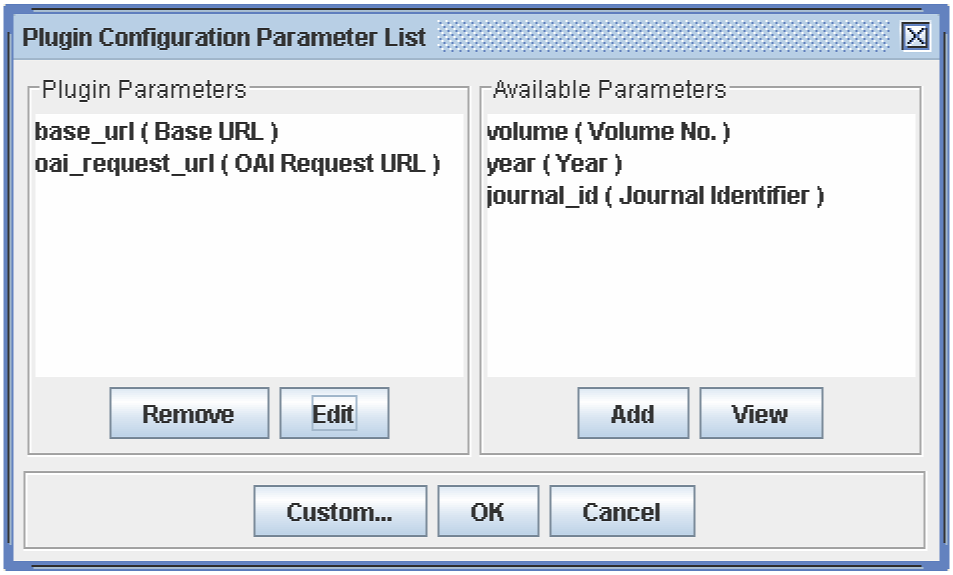
Field Size: 40

Description: Usually of the form http://<etdCollection-name>.cc

☑ Parameter is integral to the identity of an AU.

Figure 2.18

- On clicking the "OK" button, the updated Configuration Parameter list will look like as shown below.



Plugin Configuration Parameter List

Plugin Parameters

base_url (Base URL)
oai_request_url (OAI Request URL)

Available Parameters

volume (Volume No.)
year (Year)
journal_id (Journal Identifier)

Figure 2.19

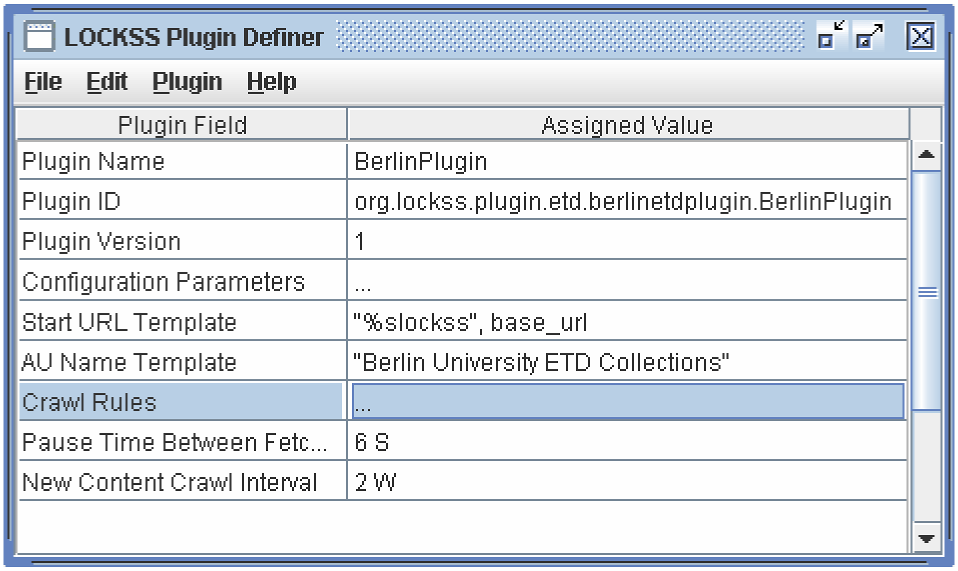
- Subsequently, the other steps as explained in section 3 are followed and the plug-in is completed.
- Once the plug-in is completed, the XML file should be opened in any editor, and the Element value should be added as given by the XML statement: <ElementValue xsi:type="java:java.lang.String" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">OAI</ElementValue>. This tells the plug-in to use an OAI crawl instead of a regular crawl.
- Since this is an OAI plug-in, we need to tell that where to find the OAI plug-in. So, the Element value should be given just after the previous step. <ElementValue xsi:type="java:java.util.ArrayList" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"> <string xsi:type="java:java.lang.String">"%slockss_permission.html", base_url</string>

3. Implementation

For LOCKSS to harvest the ETD collections at any university, it needs a permissions page from the university, which grants the LOCKSS permission to use their ETDs for preservation and harvesting.

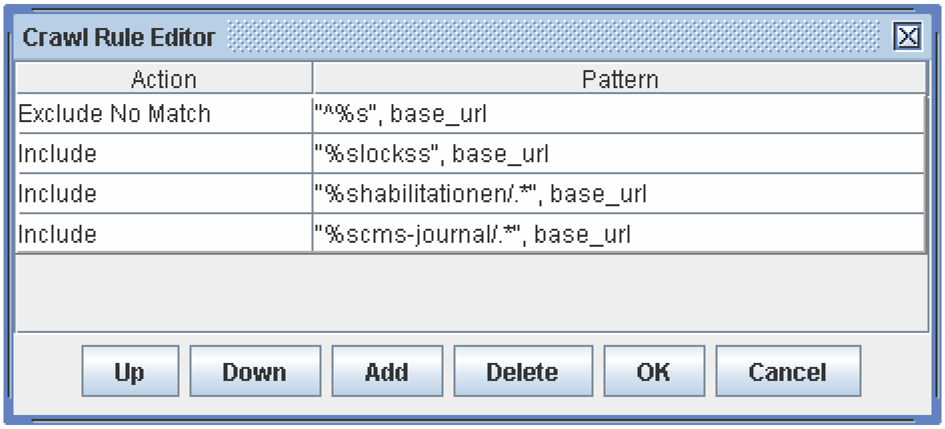
3.1 Humboldt-Universität, Berlin, Germany

The permissions page (popularly called as manifest page) for the ETD collections of Humboldt-Universität, Berlin, Germany is available at the location http://edoc.hu-berlin.de/lockss//. The plug-in for this ETD collection is written using the LOCKSS tool. The screen shots of crawl rules and the test crawl results are given below. Since this is an OAI plug-in, it needs a manual change in the XML file which is generated. In this case, the BASE_URL is http://edoc.hu-berlin.de/lockss/



Plugin Field	Assigned Value
Plugin Name	BerlinPlugin
Plugin ID	org.lockss.plugin.etd.berlinetdplugin.BerlinPlugin
Plugin Version	1
Configuration Parameters	...
Start URL Template	"%slockss", base_url
AU Name Template	"Berlin University ETD Collections"
Crawl Rules	...
Pause Time Between Fetc...	6 S
New Content Crawl Interval	2 W

Figure 3.01

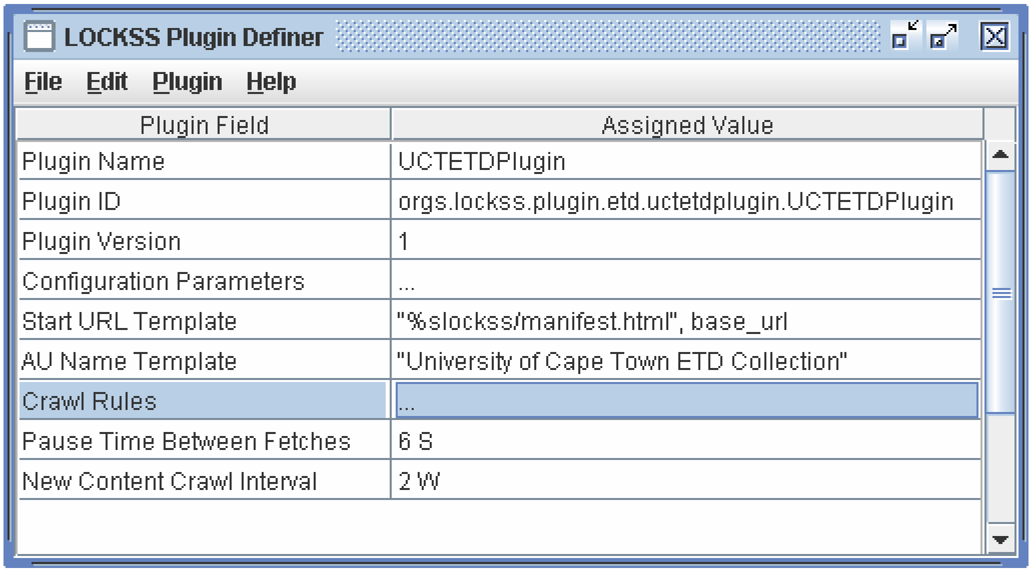


Action	Pattern
Exclude No Match	"^%s", base_url
Include	"%slockss", base_url
Include	"%shabilitationen/.*", base_url
Include	"%scms-journal/.*", base_url

Figure 3.02

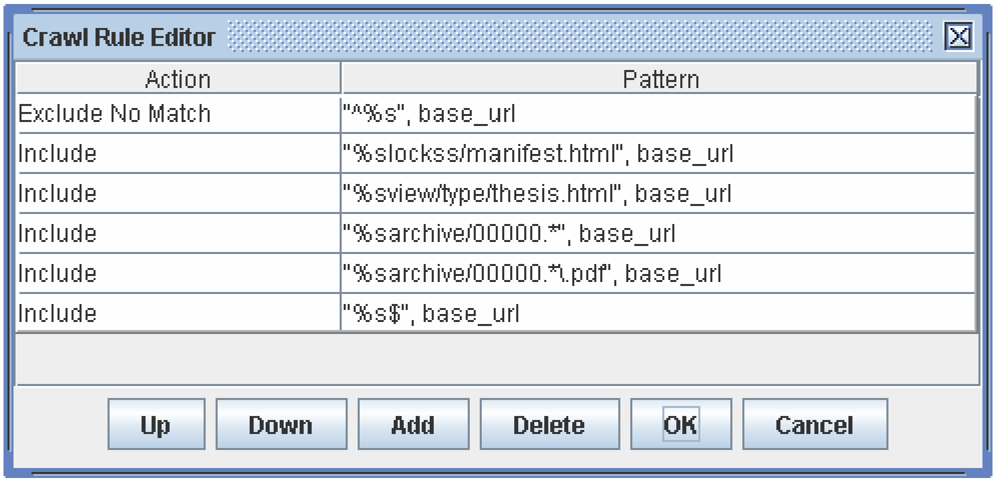
3.2 University of Cape Town, South Africa

The manifest page for the ETD collections of University of Cape Town, South Africa is available at the location http://pubs.cs.uct.ac.za/lockss/manifest.html. The plug-in for this ETD collection is written using the LOCKSS tool. The screen shots of crawl rules and the test crawl results are given below. Since this is an OAI plug-in, it needs a manual change in the XML file which is generated. In this case, the BASE_URL is http://pubs.cs.uct.ac.za/



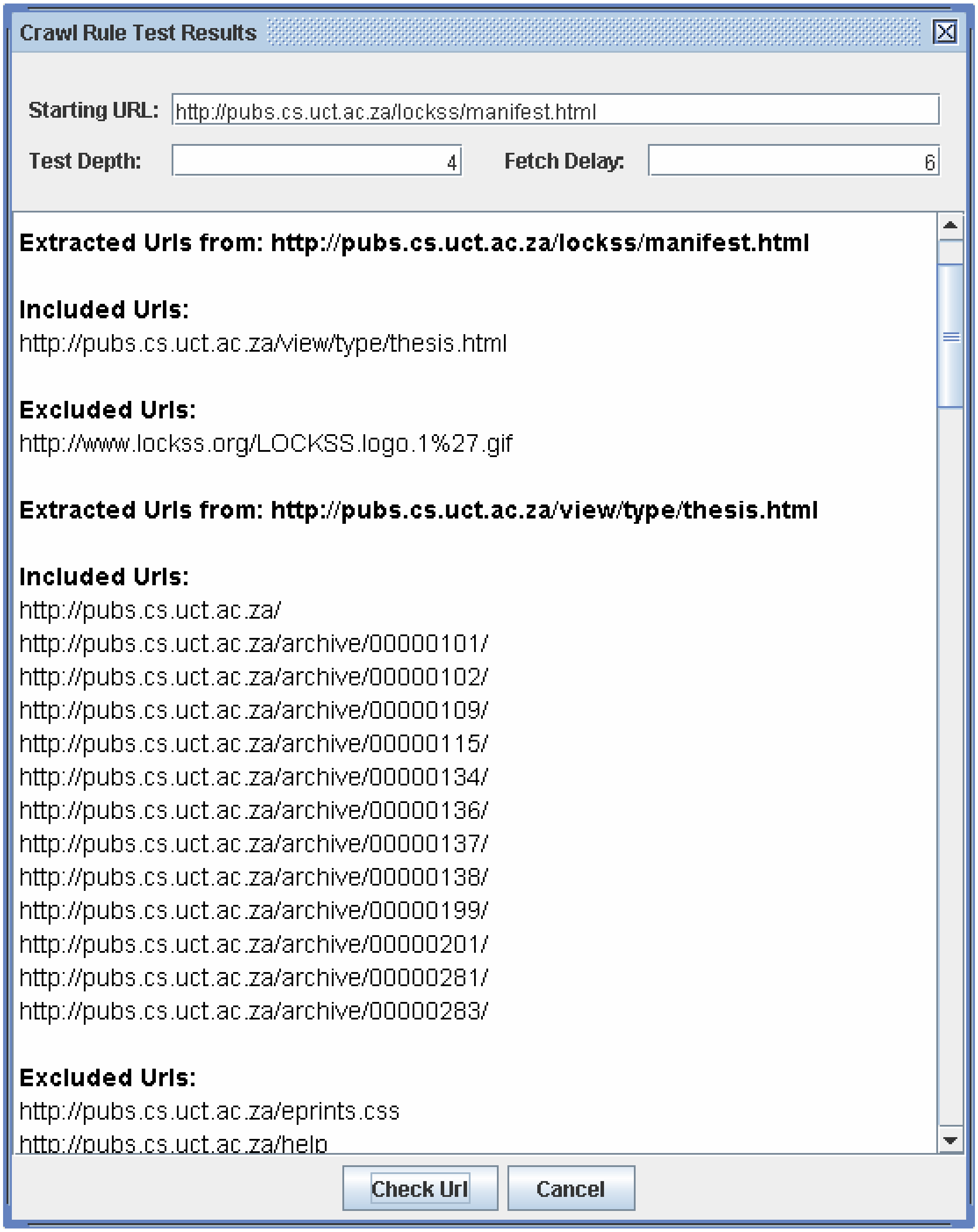
Plugin Field	Assigned Value
Plugin Name	UCTETDPlugin
Plugin ID	orgs.lockss.plugin.etd.uctetdplugin.UCTETDPlugin
Plugin Version	1
Configuration Parameters	...
Start URL Template	"%slockss/manifest.html", base_url
AU Name Template	"University of Cape Town ETD Collection"
Crawl Rules	...
Pause Time Between Fetches	6 S
New Content Crawl Interval	2 W

Figure 3.03



Action	Pattern
Exclude No Match	"^%s", base_url
Include	"%slockss/manifest.html", base_url
Include	"%sview/type/thesis.html", base_url
Include	"%sarchive/00000.*", base_url
Include	"%sarchive/00000.*\.pdf", base_url
Include	"%s$", base_url

Figure 3.04



Starting URL: http://pubs.cs.uct.ac.za/lockss/manifest.html

Test Depth: 4

Fetch Delay: 6

Extracted Urls from: http://pubs.cs.uct.ac.za/lockss/manifest.html

Included Urls:
http://pubs.cs.uct.ac.za/view/type/thesis.html

Excluded Urls:
http://www.lockss.org/LOCKSS.logo.1%27.gif

Extracted Urls from: http://pubs.cs.uct.ac.za/view/type/thesis.html

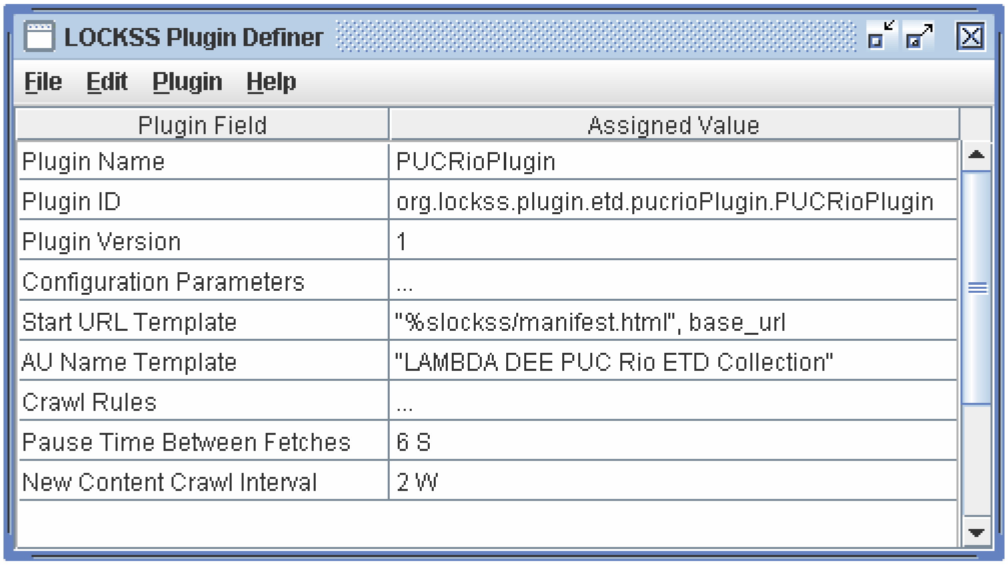
Included Urls:
http://pubs.cs.uct.ac.za/
http://pubs.cs.uct.ac.za/archive/00000101/
http://pubs.cs.uct.ac.za/archive/00000102/
http://pubs.cs.uct.ac.za/archive/00000109/
http://pubs.cs.uct.ac.za/archive/00000115/
http://pubs.cs.uct.ac.za/archive/00000134/
http://pubs.cs.uct.ac.za/archive/00000136/
http://pubs.cs.uct.ac.za/archive/00000137/
http://pubs.cs.uct.ac.za/archive/00000138/
http://pubs.cs.uct.ac.za/archive/00000199/
http://pubs.cs.uct.ac.za/archive/00000201/
http://pubs.cs.uct.ac.za/archive/00000281/
http://pubs.cs.uct.ac.za/archive/00000283/

Excluded Urls:
http://pubs.cs.uct.ac.za/eprints.css
http://pubs.cs.uct.ac.za/help

Figure 3.05

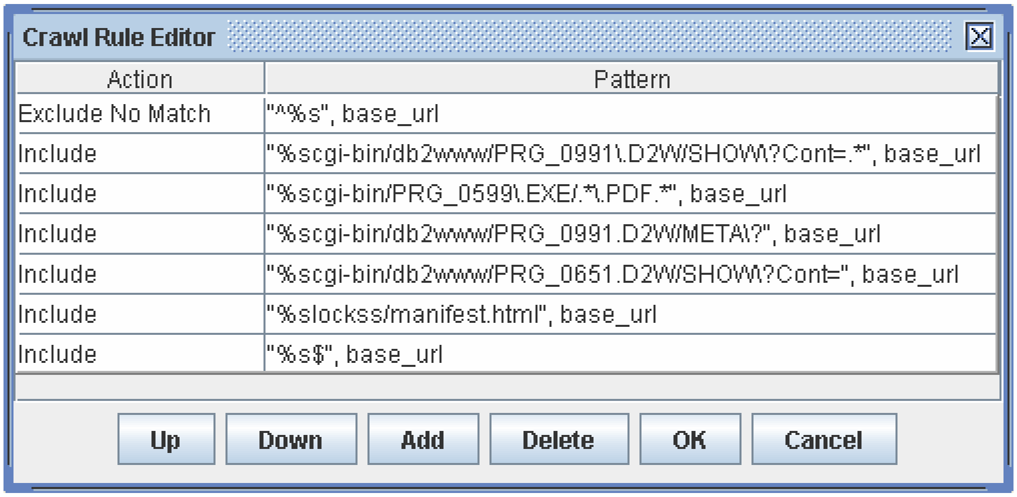
3.3 Pontifícia Universidade Católica do Rio de Janeiro, Brazil

The manifest page for the ETD collections of Pontifícia Universidade Católica do Rio de Janeiro, Brazil is available at the location http://www.maxwell.lambda.ele.puc-rio.br/lockss/manifest.html. The plug-in for this ETD collection is written using the LOCKSS tool. The screen shots of crawl rules and the test crawl results are given below. Since this is an OAI plug-in, it needs a manual change in the XML file which is generated. In this case, the BASE_URL is http://www.maxwell.lambda.ele.puc-rio.br/.



Plugin Field	Assigned Value
Plugin Name	PUCRioPlugin
Plugin ID	org.lockss.plugin.etd.pucrioPlugin.PUCRioPlugin
Plugin Version	1
Configuration Parameters	...
Start URL Template	"%slockss/manifest.html", base_url
AU Name Template	"LAMBDA DEE PUC Rio ETD Collection"
Crawl Rules	...
Pause Time Between Fetches	6 S
New Content Crawl Interval	2 W

Figure 3.06

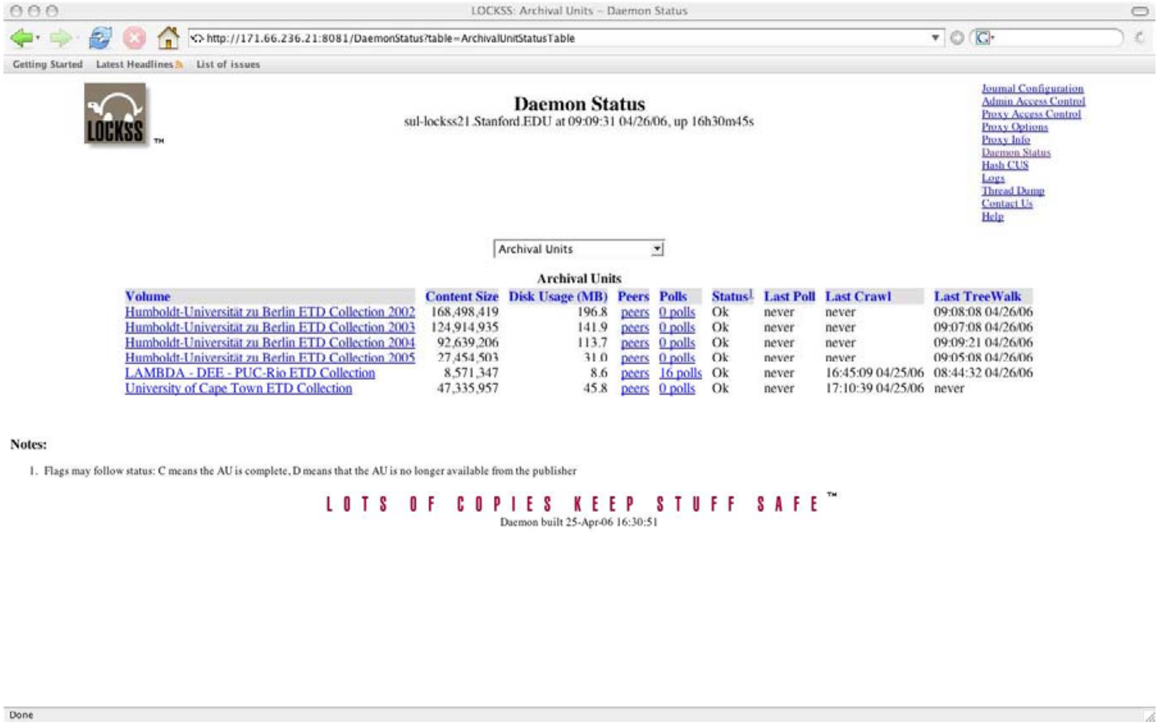


Action	Pattern
Exclude No Match	"^%s", base_url
Include	"%scgi-bin/db2www/PRG_0991\.D2W/SHOW\?Cont=.*", base_url
Include	"%scgi-bin/PRG_0599\.EXE/.*\.PDF.*", base_url
Include	"%scgi-bin/db2www/PRG_0991.D2W/META\?", base_url
Include	"%scgi-bin/db2www/PRG_0651.D2W/SHOW\?Cont=", base_url
Include	"%slockss/manifest.html", base_url
Include	"%s$", base_url

Figure 3.07

4. Analysis and Results

The following screen shots shows the statistics related to the harvest results of International ETDs. These were provided by Stanford University, as a proof that International ETDs are indeed being harvested properly.



Daemon Status

sul-lockss21.Stanford.EDU at 09:09:31 04/26/06, up 16h30m45s

Archival Units

Volume	Content Size	Disk Usage (MB)	Peers	Polls	Status[1]	Last Poll	Last Crawl	Last TreeWalk
Humboldt-Universität zu Berlin ETD Collection 2002	168,498,419	196.8	peers	0 polls	Ok	never	never	09:08:08 04/26/06
Humboldt-Universität zu Berlin ETD Collection 2003	124,914,935	141.9	peers	0 polls	Ok	never	never	09:07:08 04/26/06
Humboldt-Universität zu Berlin ETD Collection 2004	92,639,206	113.7	peers	0 polls	Ok	never	never	09:09:21 04/26/06
Humboldt-Universität zu Berlin ETD Collection 2005	27,454,503	31.0	peers	0 polls	Ok	never	never	09:05:08 04/26/06
LAMBDA - DEE - PUC-Rio ETD Collection	8,571,347	8.6	peers	16 polls	Ok	never	16:45:09 04/25/06	08:44:32 04/26/06
University of Cape Town ETD Collection	47,335,957	45.8	peers	0 polls	Ok	never	17:10:39 04/25/06	never

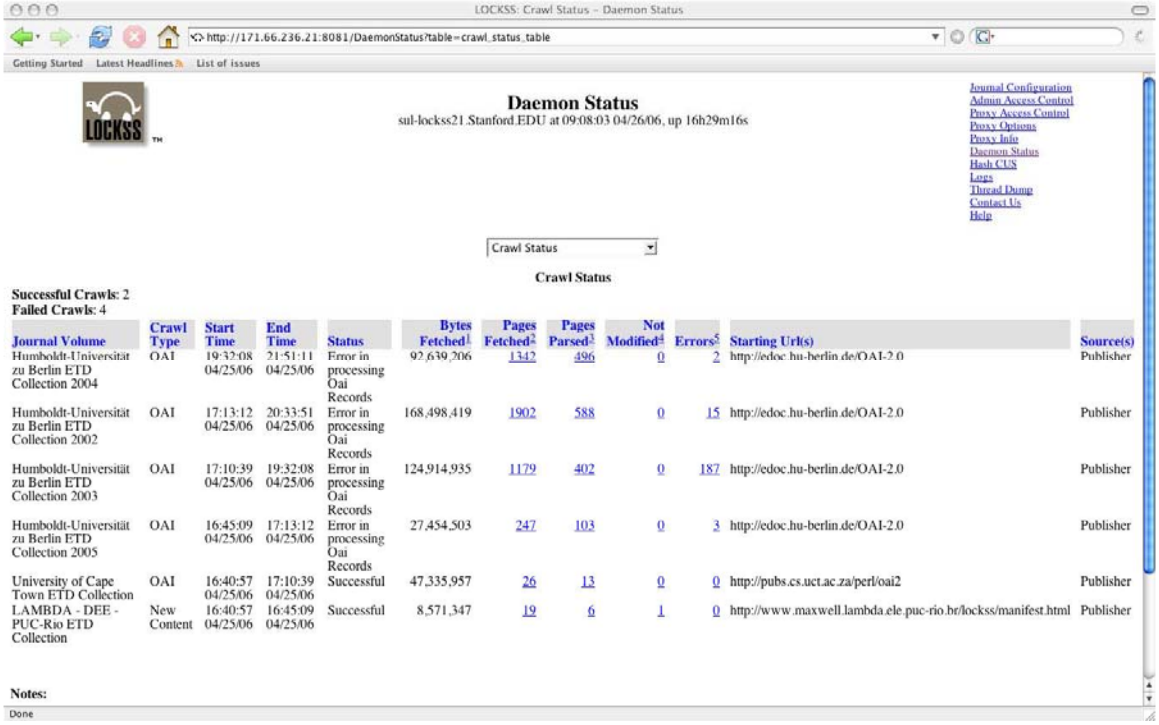
Notes:

1. Flags may follow status: C means the AU is complete, D means that the AU is no longer available from the publisher

LOTS OF COPIES KEEP STUFF SAFE™

Daemon built 25-Apr-06 16:30:51

Figure 4.01



Daemon Status

sul-lockss21.Stanford.EDU at 09:08:03 04/26/06, up 16h29m16s

Crawl Status

Successful Crawls: 2
Failed Crawls: 4

Journal Volume	Crawl Type	Start Time	End Time	Status	Bytes Fetched[1]	Pages Fetched[2]	Pages Parsed[3]	Not Modified[4]	Errors[5]	Starting Url(s)	Source(s)
Humboldt-Universität zu Berlin ETD Collection 2004	OAI	19:32:08 04/25/06	21:51:11 04/25/06	Error in processing Oai Records	92,639,206	1342	496	0	2	http://edoc.hu-berlin.de/OAI-2.0	Publisher
Humboldt-Universität zu Berlin ETD Collection 2002	OAI	17:13:12 04/25/06	20:33:51 04/25/06	Error in processing Oai Records	168,498,419	1902	588	0	15	http://edoc.hu-berlin.de/OAI-2.0	Publisher
Humboldt-Universität zu Berlin ETD Collection 2003	OAI	17:10:39 04/25/06	19:32:08 04/25/06	Error in processing Oai Records	124,914,935	1179	402	0	187	http://edoc.hu-berlin.de/OAI-2.0	Publisher
Humboldt-Universität zu Berlin ETD Collection 2005	OAI	16:45:09 04/25/06	17:13:12 04/25/06	Error in processing Oai Records	27,454,503	247	103	0	3	http://edoc.hu-berlin.de/OAI-2.0	Publisher
University of Cape Town ETD Collection	OAI	16:40:57 04/25/06	17:10:39 04/25/06	Successful	47,335,957	26	13	0	0	http://pubs.cs.uct.ac.za/perl/oai2	Publisher
LAMBDA - DEE - PUC-Rio ETD Collection	New Content	16:40:57 04/25/06	16:45:09 04/25/06	Successful	8,571,347	19	6	1	0	http://www.maxwell.lambda.ele.puc-rio.br/lockss/manifest.html	Publisher

Notes:

Figure 4.02

5. Future Work

5.1 Providing Public and Private Access

Once the ETD collection is being harvested, the next step is to present it to the users. This should be done in a way which would be easily read and interpreted by the users – both the librarians and the people from the universities who wish to see the harvest status of their ETD collections. This could be done in the following way.

5. 1. 1 For the Users (Open Archive or Open access)

We can design a 3 tier Client/Server architecture for the online portal system (which should be individually deployed in the participating universities) wherein the users can login and check the preservation status of their ETD collection. This statistical result need not contain intricate details like the 'last crawl depth', 'content size', etc. It should only contain the details required by the public users. Since they are most interested in viewing the status (harvest successful or unsuccessful) of their ETD collection, only such details need to be provided. The three tier client server architecture that could be designed for this purpose is shown in the following figure.

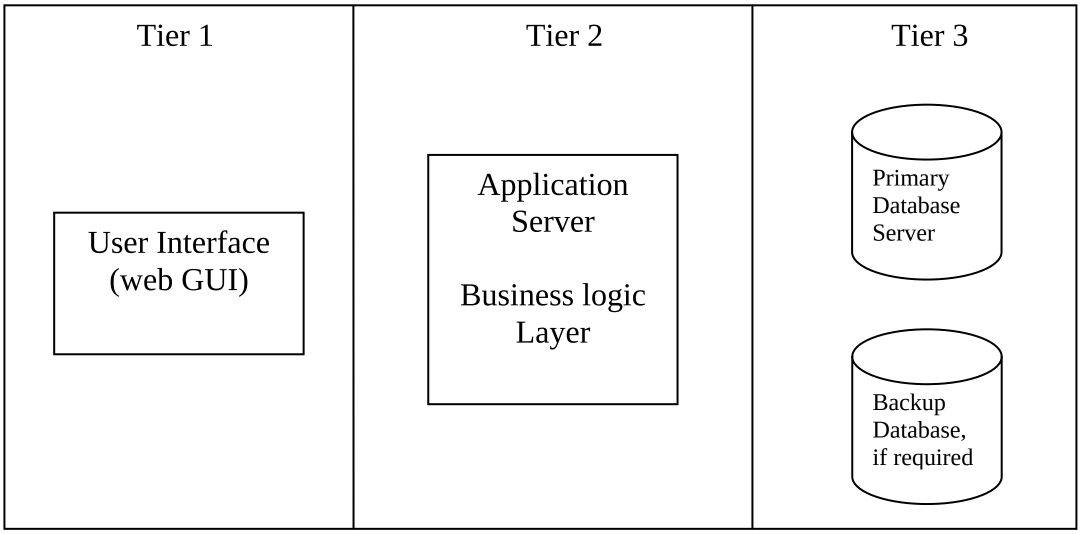
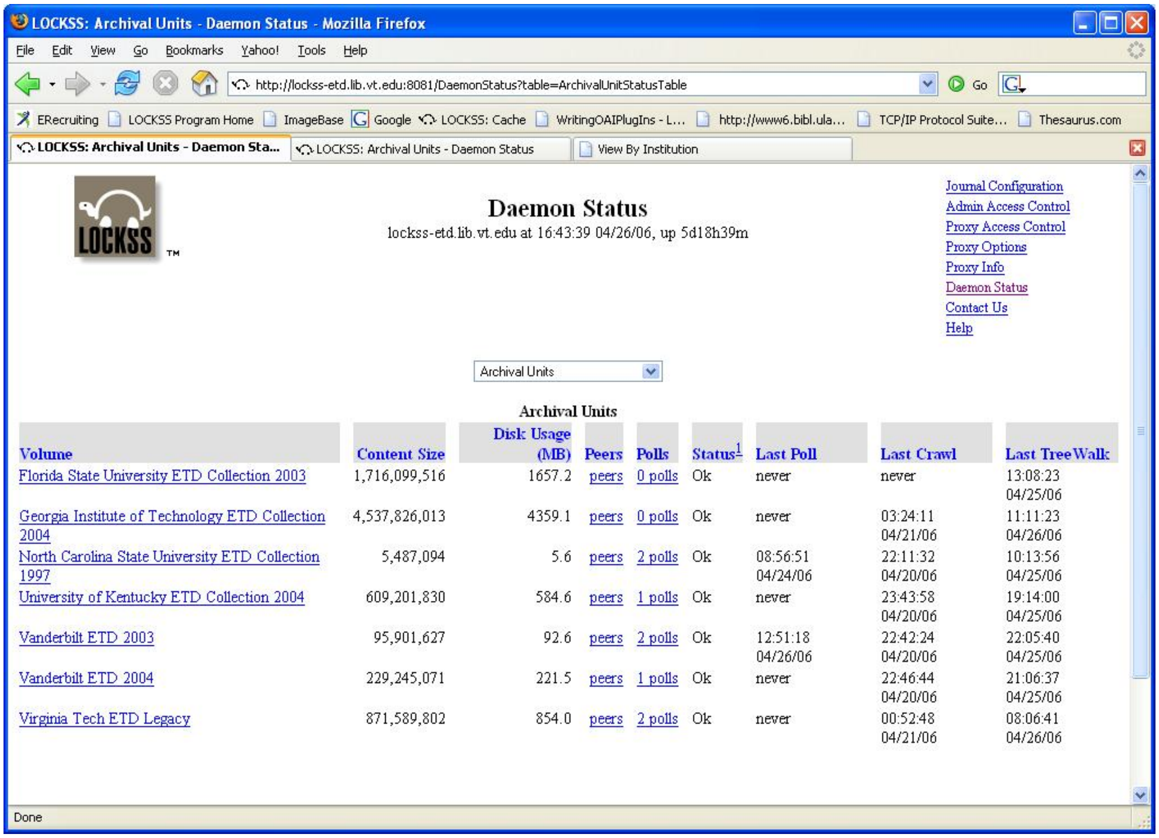


Figure 5.01

5.1.2 For the developers (Dark Archive or Restricted/Private Access)

Each university stores the contents of every other university it its local cache – To enable the developers and participating universities to view the statistical results, a detailed report of failures in harvesting could also be presented in a secure (restricted access) online portal. This would make the maintenance of the harvest results easier, and would help the universities involved to track the errors and correct them as and when needed. This is called a **“Dark Archive”** (restricted access given to the developers and participating universities).

A similar application (private access) is already being designed, and is in use for the Virginia Tech collection, as shown in the figure below.


LOCKSS: Archival Units - Daemon Status - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://lockss-etd.lib.vt.edu:8081/DaemonStatus?table=ArchivalUnitStatusTable
Go
ERecruiting LOCKSS Program Home ImageBase Google LOCKSS: Cache WritingOAIPlugIns - L... http://www6.bibl.ula... TCP/IP Protocol Suite... Thesaurus.com
LOCKSS: Archival Units - Daemon Sta...
LOCKSS: Archival Units - Daemon Status
View By Institution
LOCKSS
Journal Configuration
Admin Access Control
Proxy Access Control
Proxy Options
Proxy Info
Daemon Status
Contact Us
Help
Archival Units
Done

Daemon Status

lockss-etd.lib.vt.edu at 16:43:39 04/26/06, up 5d18h39m

Archival Units

Volume	Content Size	Disk Usage (MB)	Peers	Polls	Status1	Last Poll	Last Crawl	Last TreeWalk
Florida State University ETD Collection 2003	1,716,099,516	1657.2	peers	0 polls	Ok	never	never	13:08:23 04/25/06
Georgia Institute of Technology ETD Collection 2004	4,537,826,013	4359.1	peers	0 polls	Ok	never	03:24:11 04/21/06	11:11:23 04/26/06
North Carolina State University ETD Collection 1997	5,487,094	5.6	peers	2 polls	Ok	08:56:51 04/24/06	22:11:32 04/20/06	10:13:56 04/25/06
University of Kentucky ETD Collection 2004	609,201,830	584.6	peers	1 polls	Ok	never	23:43:58 04/20/06	19:14:00 04/25/06
Vanderbilt ETD 2003	95,901,627	92.6	peers	2 polls	Ok	12:51:18 04/26/06	22:42:24 04/20/06	22:05:40 04/25/06
Vanderbilt ETD 2004	229,245,071	221.5	peers	1 polls	Ok	never	22:46:44 04/20/06	21:06:37 04/25/06
Virginia Tech ETD Legacy	871,589,802	854.0	peers	2 polls	Ok	never	00:52:48 04/21/06	08:06:41 04/26/06

Figure 5.02

6. References

[1] "An Introduction to Digital Preservation", Technical Advisory Service for images Advice (TASI), March 2002

[2] K. Lee, O. Slattery, R. Lu, X. Tang, and V. McCrary, "The State of the Art and Practice in Digital Preservation", Journal of Research of the National Institute of Standards and Technology , Volume 107, Number 1, 2002

[3] D. S. H. Rosenthal, M. Roussopoulos, T.J. Giuli, P. Maniatis, Mary Baker, "Using Hard Disks For Digital Preservation"

[4] T. Hendley, "Comparison of Methods & Costs of Digital Preservation", British Library Research and Innovation Report 106, British Library Research and Innovation Center, West Yorkshire (1998), 121 pp.

[5] LOCKSS – Lots of Copies Keep Stuff Safe, A Preservation tool developed by Stanford University, http://www.lockss.com/

[6] Council on Library and Information Resources, Annual Report 1997-1998, http://www.clir.org/pubs/annual/annrpt97/preservation.html

[7] Manjula Iyer, "Final project report (CS6604)"- http://pubs.dlib.vt.edu:9090/34/01/ManjulaIyerFinalReport.pdf

[8] Miguel Ángel Márdero Arellano, "Digital Preservation of Scientific Information in Brazil: an initial approach of existing models", ICCC 8th International Conference on Electronic Publishing, 2004

[9] Dale Peters, "Developing Digital Libraries in South Africa: Breaking out of the Bookish Mould", Durban, South Africa, http://www.ukzn.ac.za/citte/papers/id48.pdf

[10] "DISA Insights of an African Model for Digital Library Development", D-Lib Magazine November 2001Volume 7 Number 11 ISSN 1082-9873

[11] OAIS Activities - Reference Model for an Open Archival Information System, http://www.rlg.org/en/page.php?Page_ID=3201