ManifoldCF- End-user Documentation

Table of contents

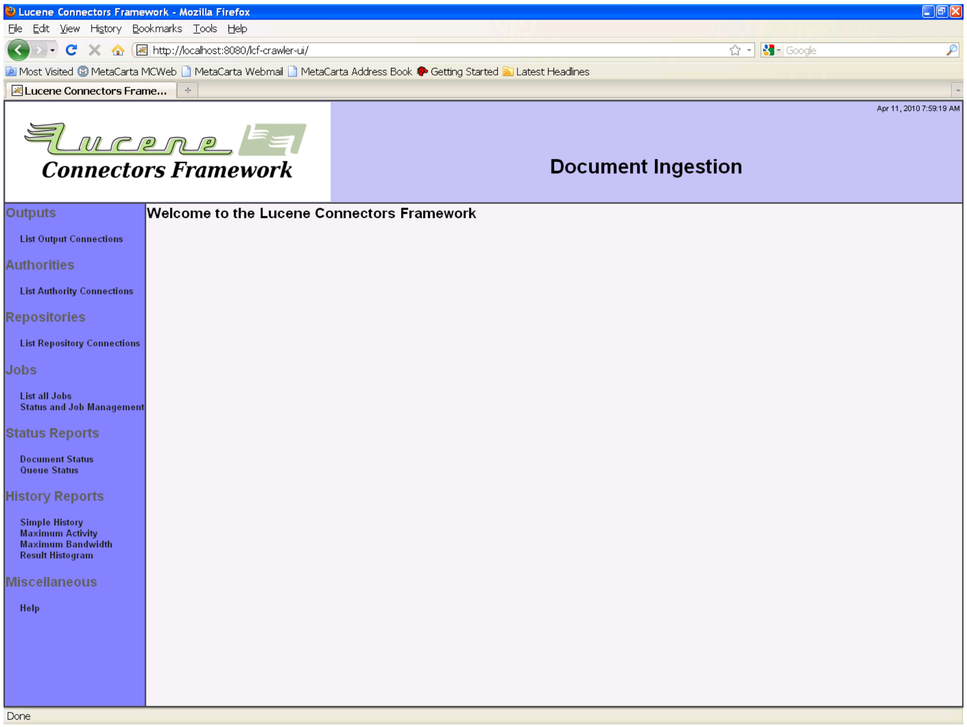
1 Overview

This manual is intended for an end-user of ManifoldCF. It is assumed that the Framework has been properly installed, either by you or by a system integrator, with all required services running and desired connection types properly registered. If you think you need to know how to do that yourself, please visit the "Developer Resources" page.

Most of this manual describes how to use the ManifoldCF user interface. On a standard ManifoldCF deployment, you would reach that interface by giving your browser a URL something like this: `http://my-server-name:8080/acf-crawler-ui`. This will, of course, differ from system to system. Please contact your system administrator to find out what URL is appropriate for your environment.

The ManifoldCF UI has been tested with Firefox and various incarnations of Internet Explorer. If you use another browser, there is a small chance that the UI will not work properly. Please let your system integrator know if you find any browser incompatibility problems.

When you do manage to enter the Framework user interface the first time, you should see a screen that looks something like this:


Lucene Connectors Framework - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Apr 11, 2010 7:59:19 AM
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

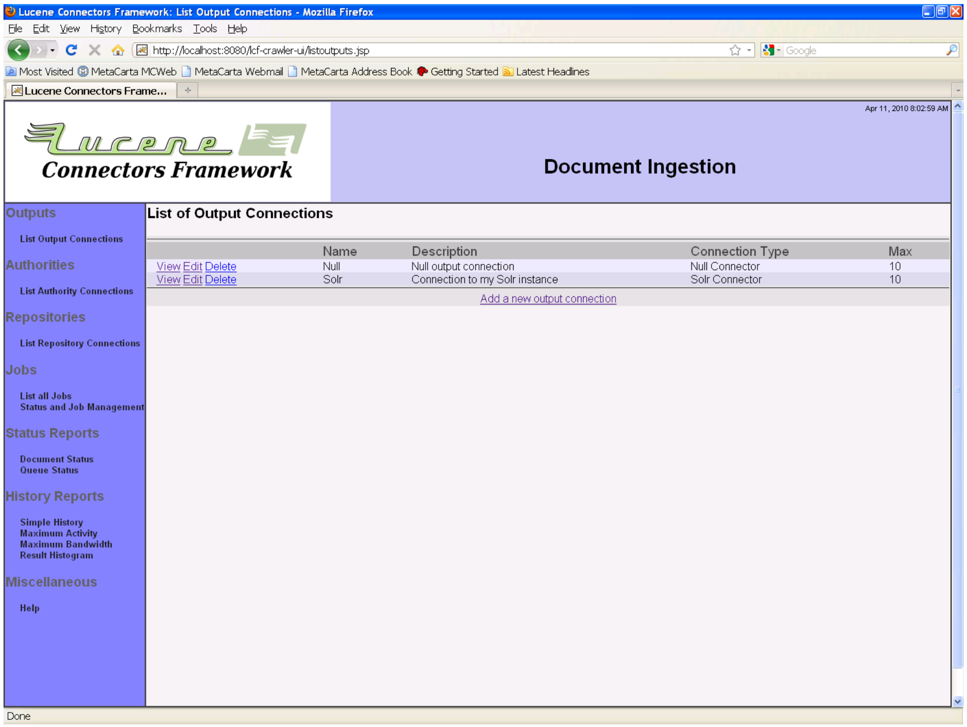
Welcome to the Lucene Connectors Framework

On the left, there are menu options you can select. The main pane on the right shows a welcome message, but depending on what you select on the left, the contents of the main pane will change. Before you try to accomplish anything, please take a moment to read the descriptions below of the menu selections, and thus get an idea of how the Framework works as a whole.

1.1 Defining Output Connections

The Framework UI's left-side menu contains a link for listing output connections. An output connection is a connection to a system or place where documents fetched from various repositories can be written to. This is often a search engine.

All jobs must specify an output connection. You can create an output connection by clicking the "List Output Connections" link in the left-side navigation menu. When you do this, the following screen will appear:

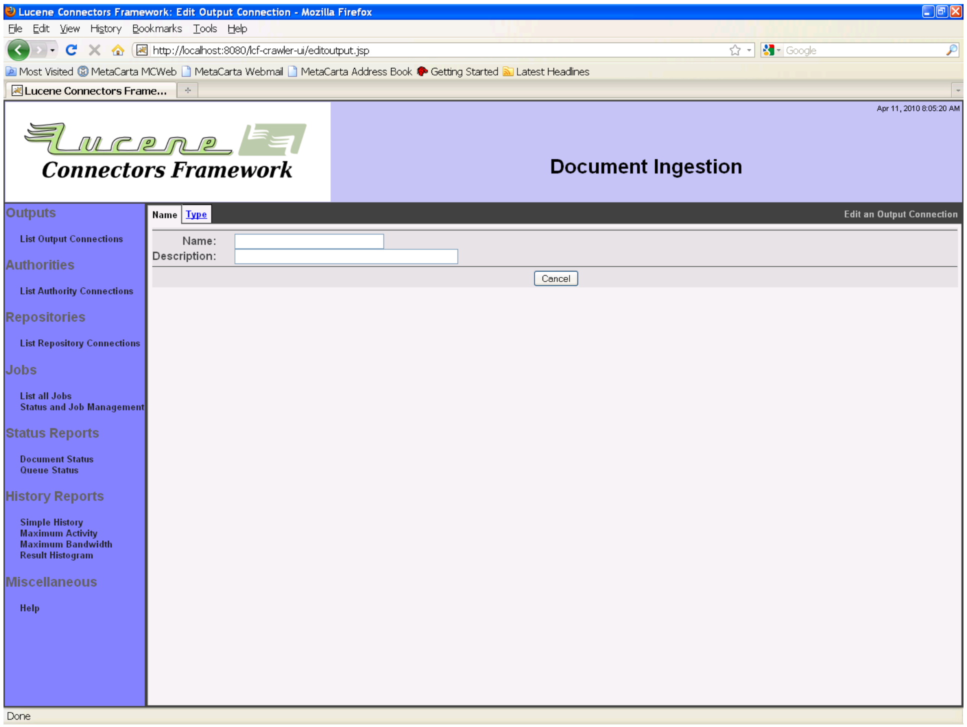

Lucene Connectors Framework: List Output Connections - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/listoutputs.jsp
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 11, 2010 8:02:59 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

List of Output Connections

	Name	Description	Connection Type	Max
View Edit Delete	Null	Null output connection	Null Connector	10
View Edit Delete	Solr	Connection to my Solr instance	Solr Connector	10

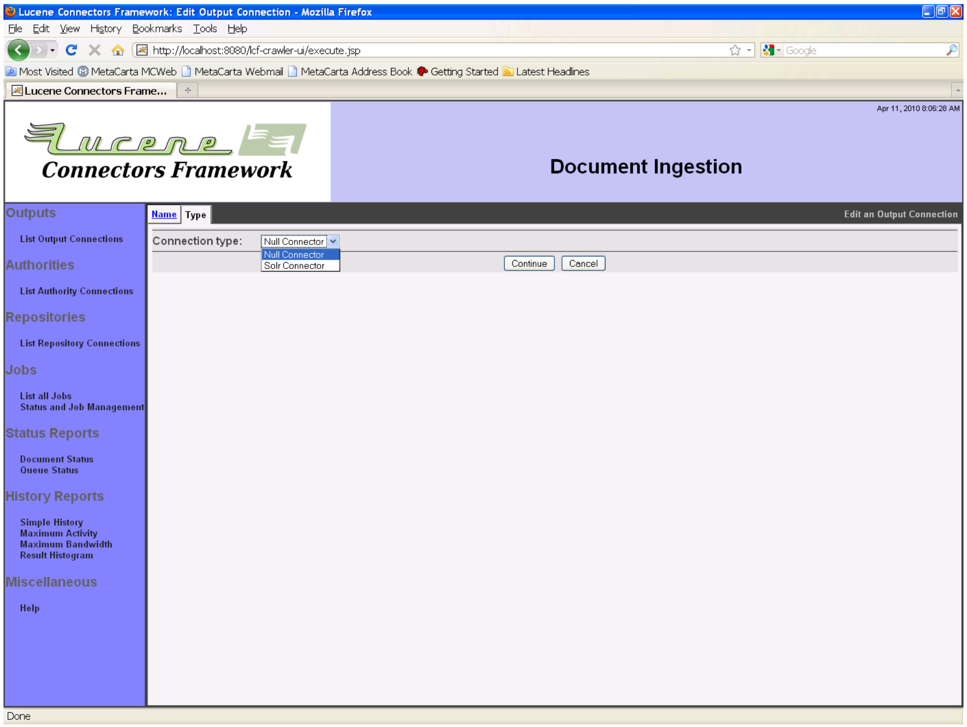
Add a new output connection

On a freshly created system, there may well be no existing output connections listed. If there are already output connections, they will be listed on this screen, along with links that allow you to view, edit, or delete them. To create a new output connection, click the "Add new output connection" link at the bottom. The following screen will then appear:


Lucene Connectors Framework: Edit Output Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/editoutput.jsp
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 11, 2010 8:05:20 AM
Lucene
Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Edit an Output Connection
Name:
Description:
Cancel
Done

The tabs across the top each present a different view of your output connection. Each tab allows you to edit a different characteristic of that connection. The exact set of tabs you see depends on the connection type you choose for the connection.

Start by giving your connection a name and a description. Remember that all output connection names must be unique, and cannot be changed after the connection is defined. The name must be no more than 32 characters long. The description can be up to 255 characters long. When you are done, click on the "Type" tab. The Type tab for the connection will then appear:

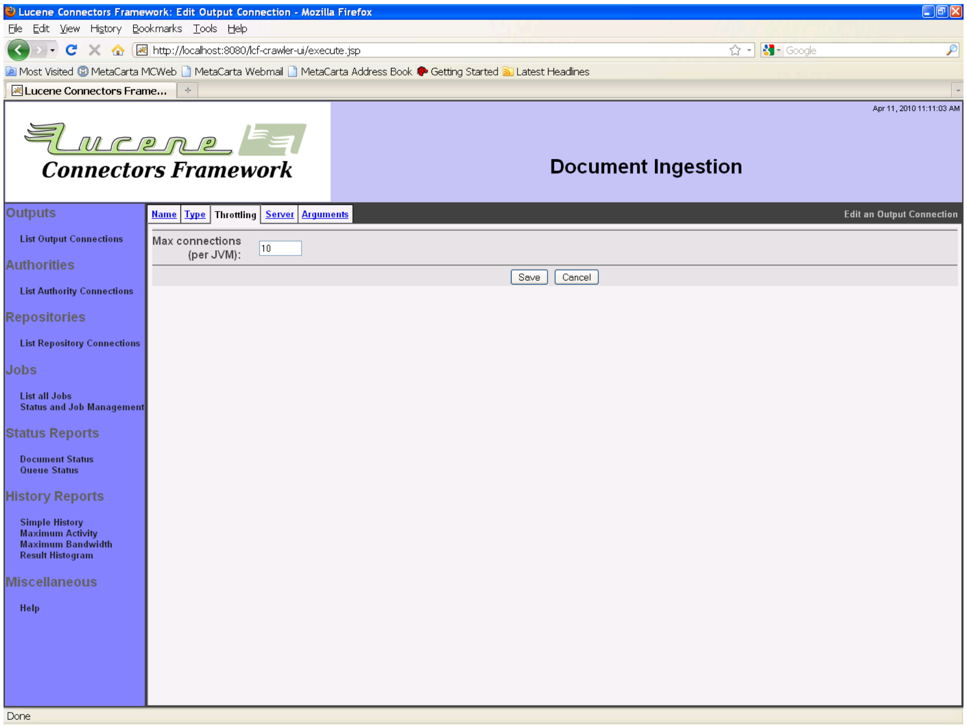

Lucene Connectors Framework: Edit Output Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Document Ingestion
Apr 11, 2010 8:06:28 AM
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Edit an Output Connection
Null Connector
Solr Connector
Continue
Cancel
Done

Connection type: Null Connector

The list of output connection types in the pulldown box, and what they are each called, is determined by your system integrator. The configuration tabs for each different kind of output connection type are described in separate sections below.

After you choose an output connection type, click the "Continue" button at the bottom of the pane. You will then see all the tabs appropriate for that kind of connection appear, and a "Save" button will also appear at the bottom of the pane. You must click the "Save" button when you are done in order to create your connection. If you click "Cancel" instead, the new connection will not be created. (The same thing will happen if you click on any of the navigation links in the left-hand pane.)

Every output connection has a "Throttling" tab. The tab looks like this:


Lucene Connectors Framework: Edit Output Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 11, 2010 11:11:03 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Throttling
Server
Arguments
Edit an Output Connection
Save
Cancel
Done

Max connections (per JVM): 10

On this tab, you can specify only one thing: how many open connections are allowed at any given time to the system the output connection talks with. This restriction helps prevent that system from being overloaded, or in some cases exceeding its license limitations. Conversely, making this number larger allows for greater overall throughput. The default value is 10, which may not be optimal for all types of output connections. Please refer to the section of the manual describing your output connection type for more precise recommendations.

Please refer to the section of the manual describing your chosen output connection type for a description of the tabs appropriate for that connection type.

After you save your connection, a summary screen will be displayed that describes your connection's configuration. This summary screen contains a line where the connection's status is displayed. If you did everything correctly, the message "Connection working" will be displayed as a status. If there was a problem, you will see a connection-type-specific diagnostic message instead. If this happens, you will need to correct the problem, by either fixing your infrastructure, or by editing the connection configuration appropriately, before the output connection will work correctly.

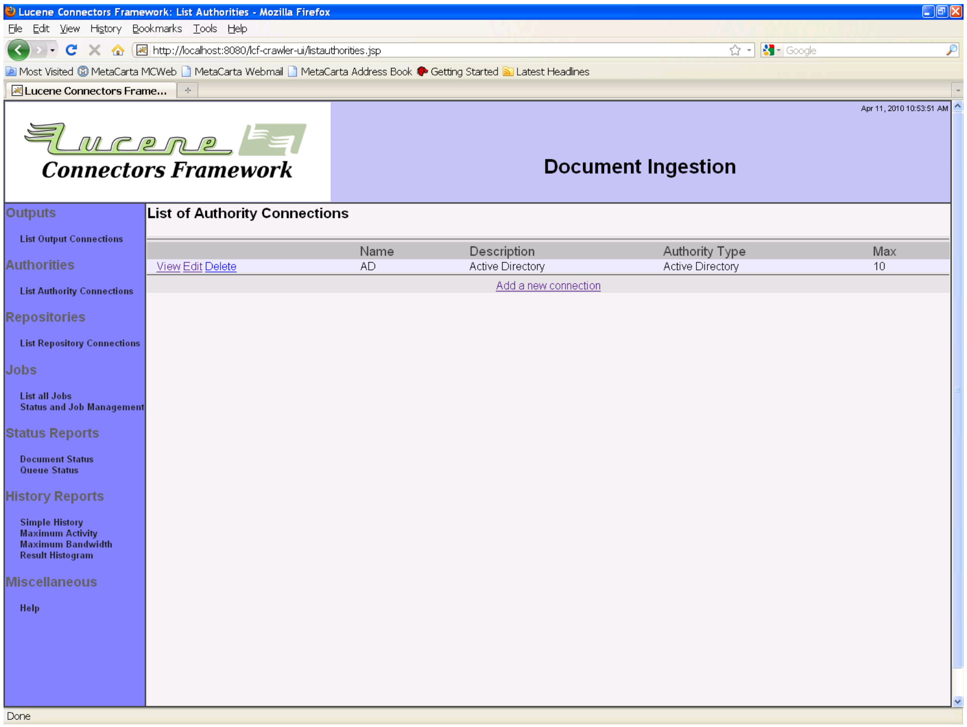
1.2 Defining Authority Connections

The Framework UI's left-side menu contains a link for listing authority connections. An authority connection is a connection to a system that defines a particular security environment. For example, if you want to index some documents that are protected by Active Directory, you would need to configure an Active Directory authority connection.

You may not need an authority if you do not mind that portions of all the documents you want to index are visible to everyone. For web, RSS, and Wiki crawling, this might be the situation. Most other repositories have some security mechanism, however.

You should define your authority connections before setting up your repository connections. While it is possible to change the relationship between a repository connection and its authority after-the-fact, in practice such changes may cause many documents to require reindexing.

You can create an authority connection by clicking the "List Authority Connections" link in the left-side navigation menu. When you do this, the following screen will appear:


Lucene Connectors Framework: List Authorities - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene
Connectors Framework
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

http://localhost:8080/lcf-crawler-ui/listauthorities.jsp

Apr 11, 2010 10:53:51 AM

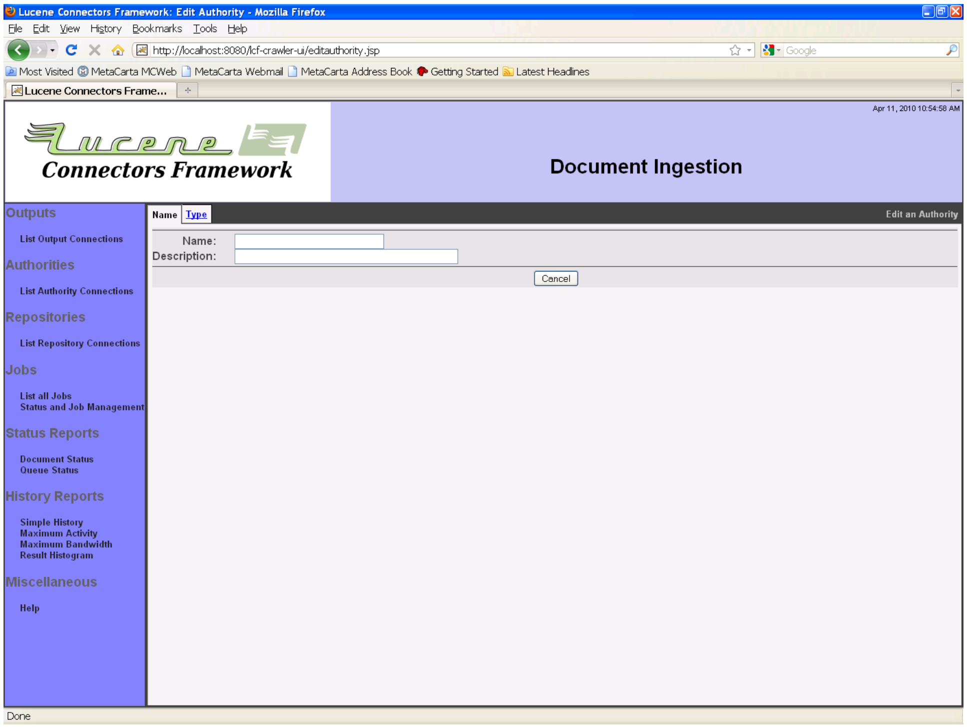
Document Ingestion

List of Authority Connections

	Name	Description	Authority Type	Max
View Edit Delete	AD	Active Directory	Active Directory	10

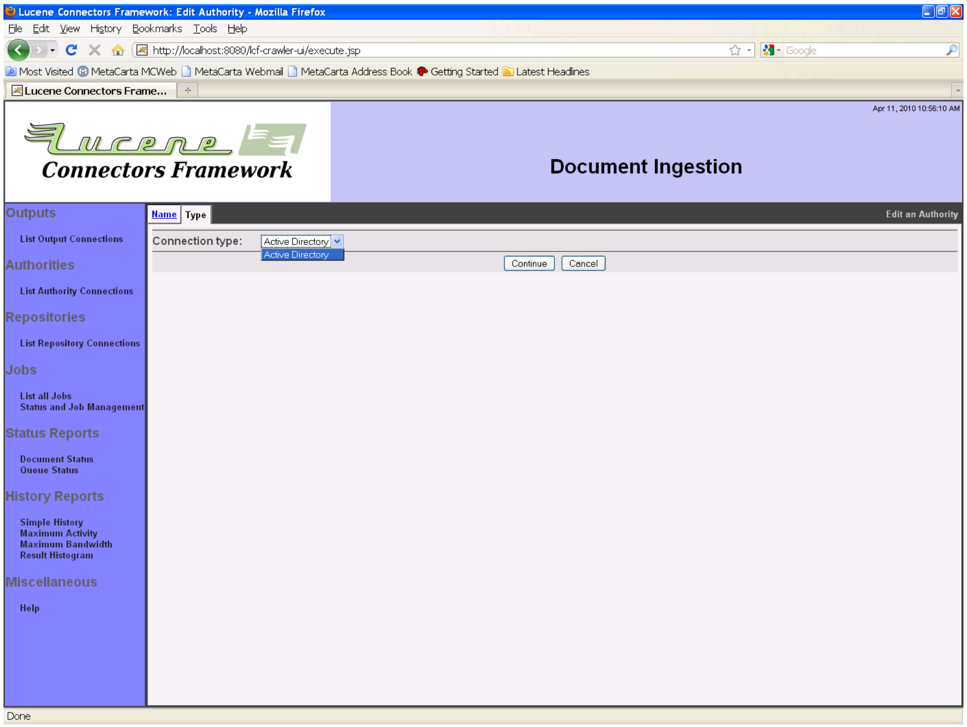
Add a new connection

On a freshly created system, there may well be no existing authority connections listed. If there are already authority connections, they will be listed on this screen, along with links that allow you to view, edit, or delete them. To create a new authority connection, click the "Add a new connection" link at the bottom. The following screen will then appear:


Lucene Connectors Framework: Edit Authority - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/editauthority.jsp
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 11, 2010 10:54:58 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Edit an Authority
Name:
Description:
Cancel
Done

The tabs across the top each present a different view of your authority connection. Each tab allows you to edit a different characteristic of that connection. The exact set of tabs you see depends on the connection type you choose for the connection.

Start by giving your connection a name and a description. Remember that all authority connection names must be unique, and cannot be changed after the connection is defined. The name must be no more than 32 characters long. The description can be up to 255 characters long. When you are done, click on the "Type" tab. The Type tab for the connection will then appear:



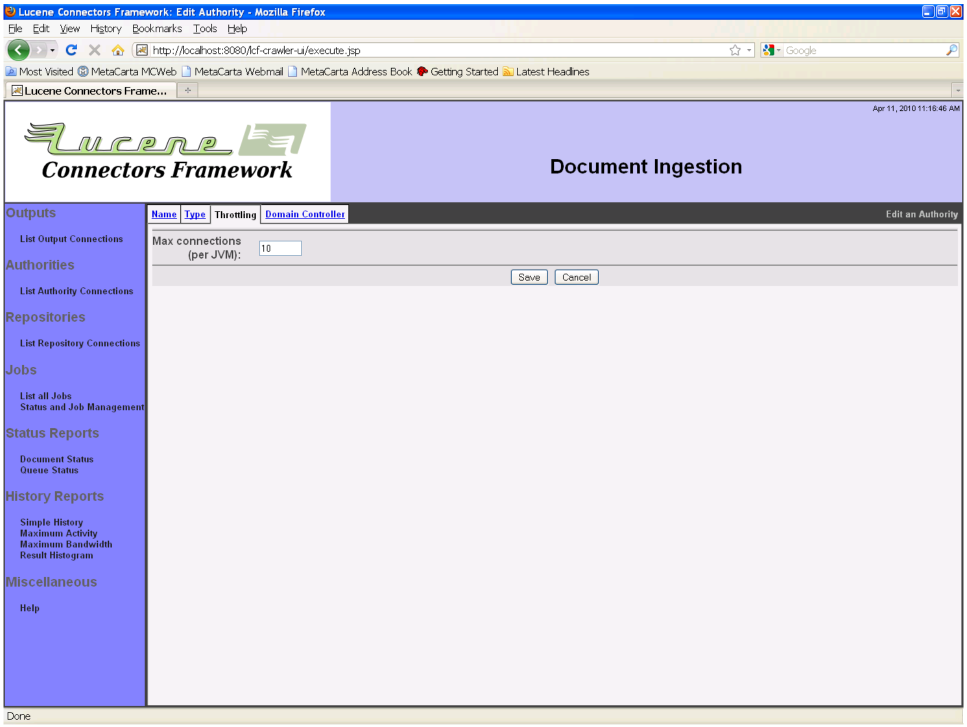
Lucene Connectors Framework: Edit Authority - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 11, 2010 10:56:10 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Edit an Authority
Active Directory
Continue
Cancel
Done

Connection type: Active Directory

The list of authority connection types in the pulldown box, and what they are each called, is determined by your system integrator. The configuration tabs for each different kind of authority connection type are described in separate sections below.

After you choose an authority connection type, click the "Continue" button at the bottom of the pane. You will then see all the tabs appropriate for that kind of connection appear, and a "Save" button will also appear at the bottom of the pane. You must click the "Save" button when you are done in order to create your connection. If you click "Cancel" instead, the new connection will not be created. (The same thing will happen if you click on any of the navigation links in the left-hand pane.)

Every authority connection has a "Throttling" tab. The tab looks like this:


Lucene Connectors Framework: Edit Authority - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 11, 2010 11:16:46 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Throttling
Domain Controller
Edit an Authority
Save
Cancel
Done

Max connections (per JVM): 10

On this tab, you can specify only one thing: how many open connections are allowed at any given time to the system the authority connection talks with. This restriction helps prevent that system from being overloaded, or in some cases exceeding its license limitations. Conversely, making this number larger allows for smaller average search latency. The default value is 10, which may not be optimal for all types of authority connections. Please refer to the section of the manual describing your authority connection type for more precise recommendations.

Please refer to the section of the manual describing your chosen authority connection type for a description of the tabs appropriate for that connection type.

After you save your connection, a summary screen will be displayed that describes your connection's configuration. This summary screen contains a line where the connection's status is displayed. If you did everything correctly, the message "Connection working" will be displayed as a status. If there was a problem, you will see a connection-type-specific diagnostic message instead. If this happens, you will need to correct the problem, by either fixing your infrastructure, or by editing the connection configuration appropriately, before the authority connection will work correctly.

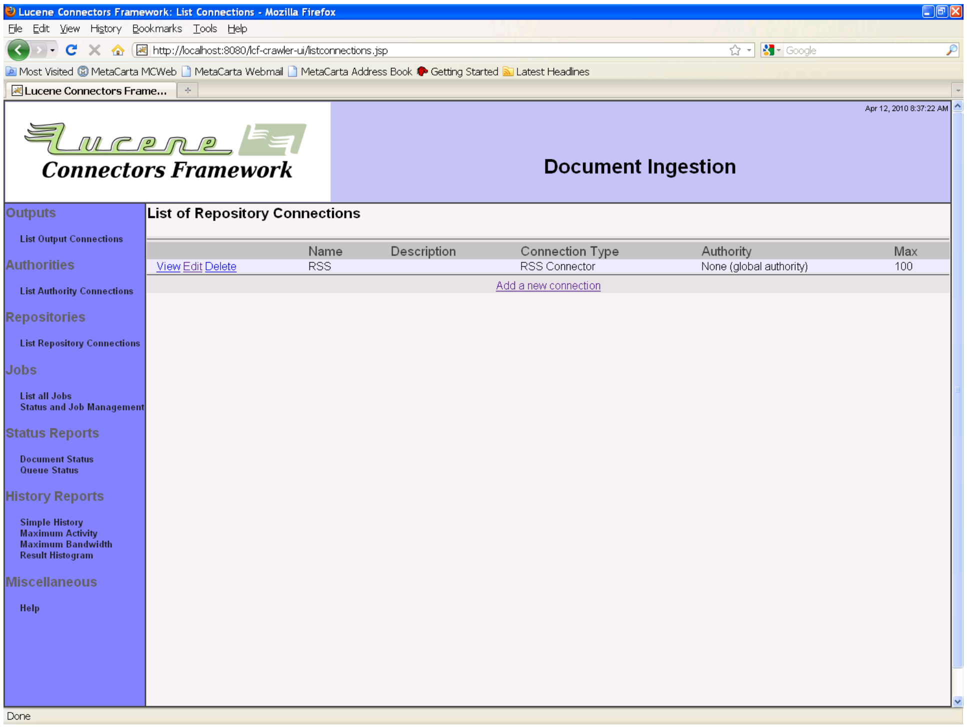
1.3 Defining Repository Connections

The Framework UI's left-hand menu contains a link for listing repository connections. A repository connection is a connection to the repository system that contains the documents that you are interested in indexing.

All jobs require you to specify a repository connection, because that is where they get their documents from. It is therefore necessary to create a repository connection before indexing any documents.

A repository connection also may have an associated authority connection. This specified authority determines the security environment in which documents from the repository connection are placed. While it is possible to change the specified authority for a repository connection after a crawl has been done, in practice this will require that all documents associated with that repository connection be reindexed. Therefore, we recommend that you set up your desired authority connection before defining your repository connection.

You can create a repository connection by clicking the "List Repository Connections" link in the left-side navigation menu. When you do this, the following screen will appear:

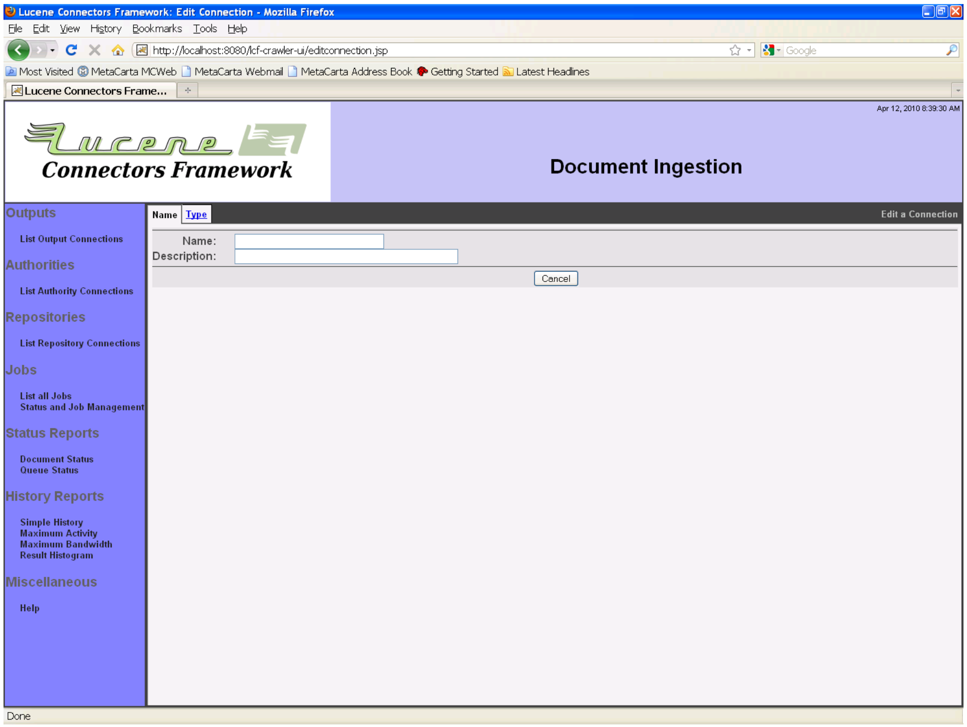

Lucene Connectors Framework: List Connections - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/listconnections.jsp
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 12, 2010 8:37:22 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

List of Repository Connections

	Name	Description	Connection Type	Authority	Max
View Edit Delete	RSS		RSS Connector	None (global authority)	100

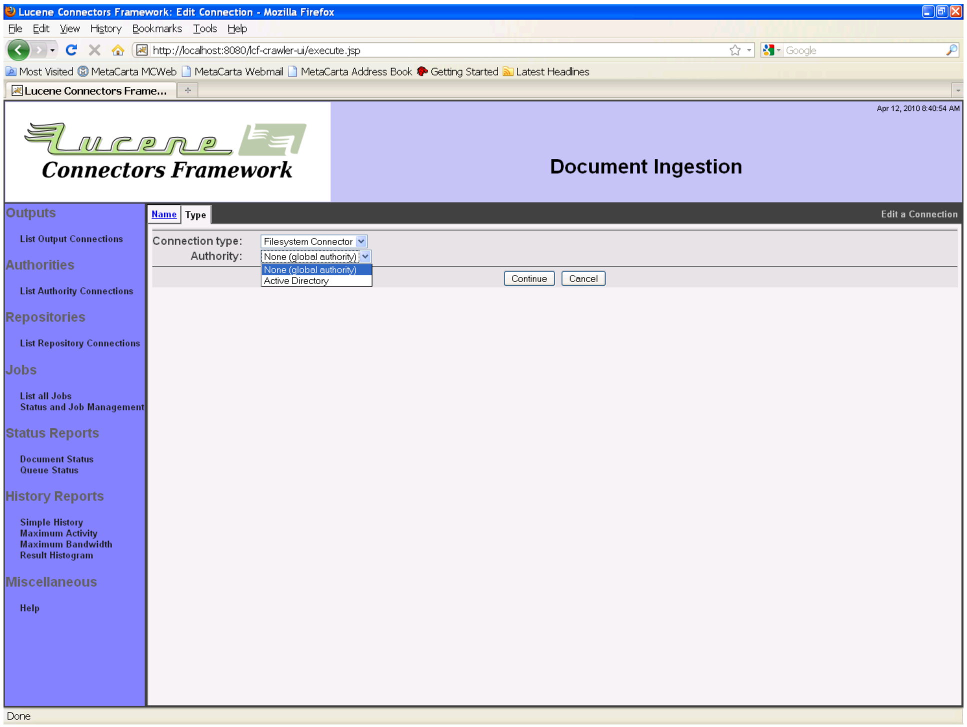
Add a new connection

On a freshly created system, there may well be no existing repository connections listed. If there are already repository connections, they will be listed on this screen, along with links that allow you to view, edit, or delete them. To create a new repository connection, click the "Add a new connection" link at the bottom. The following screen will then appear:


Lucene Connectors Framework: Edit Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/editconnection.jsp
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 12, 2010 8:39:30 AM
Lucene
Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Edit a Connection
Name:
Description:
Cancel
Done

The tabs across the top each present a different view of your repository connection. Each tab allows you to edit a different characteristic of that connection. The exact set of tabs you see depends on the connection type you choose for the connection.

Start by giving your connection a name and a description. Remember that all repository connection names must be unique, and cannot be changed after the connection is defined. The name must be no more than 32 characters long. The description can be up to 255 characters long. When you are done, click on the "Type" tab. The Type tab for the connection will then appear:


Lucene Connectors Framework: Edit Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Apr 12, 2010 8:40:54 AM
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Edit a Connection
None (global authority)
Active Directory
Continue
Cancel
Done

Connection type: Filesystem Connector

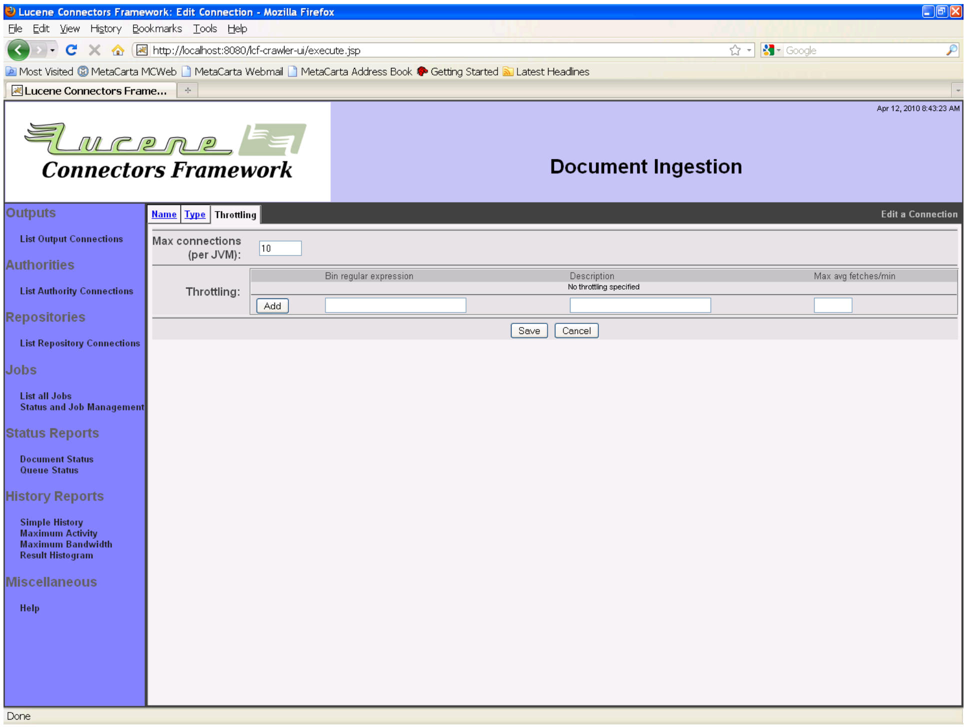
Authority: None (global authority)

The list of repository connection types in the pulldown box, and what they are each called, is determined by your system integrator. The configuration tabs for each different kind of repository connection type are described in separate sections below.

You may also at this point select the authority connection to secure all documents fetched from this repository with. Bear in mind that only some authority connection types are compatible with any given repository connection types. Read the details of your desired repository or authority connection type to understand its intentions, and how it is expected to be used.

After you choose the desired repository connection type and an authority connection, click the "Continue" button at the bottom of the pane. You will then see all the tabs appropriate for that kind of connection appear, and a "Save" button will also appear at the bottom of the pane. You must click the "Save" button when you are done in order to create or update your connection. If you click "Cancel" instead, the new connection will not be created. (The same thing will happen if you click on any of the navigation links in the left-hand pane.)

Every repository connection has a "Throttling" tab. The tab looks like this:



Max connections (per JVM): 10

Throttling:

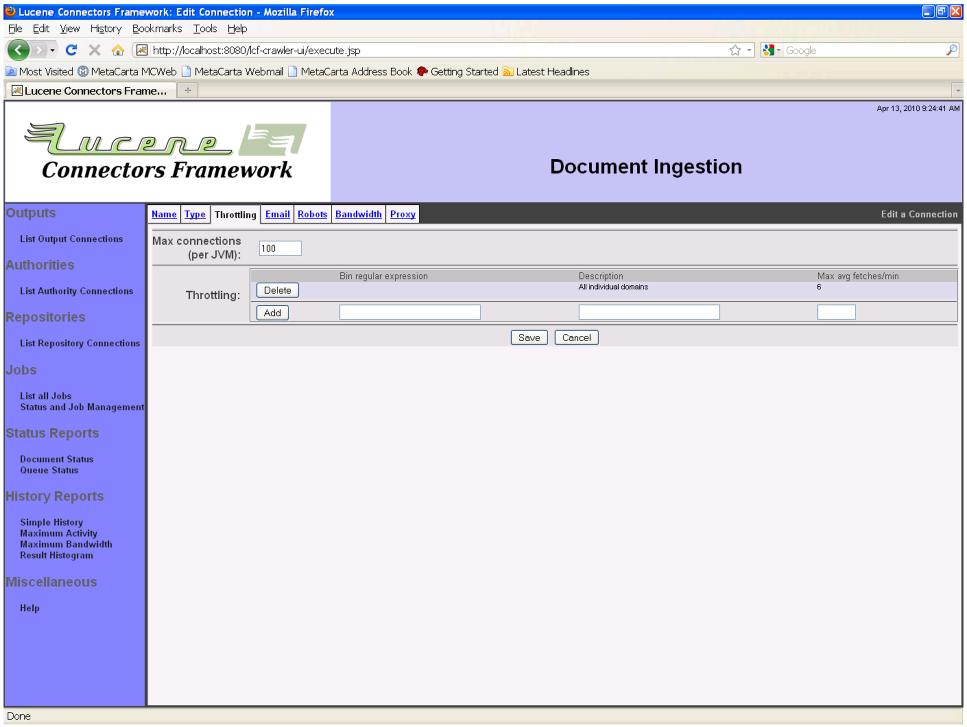
Bin regular expression	Description	Max avg fetches/min
	No throttling specified	

On this tab, you can specify two things. The first is how many open connections are allowed at any given time to the system the authority connection talks with. This restriction helps prevent that system from being overloaded, or in some cases exceeding its license limitations. Conversely, making this number larger allows for smaller average search latency. The default value is 10, which may not be optimal for all types of repository connections. Please refer to the section of the manual describing your authority connection type for more precise recommendations. The second specifies how rapidly, on average, the crawler will fetch documents via this connection.

Each connection type has its own notion of "throttling bin". A throttling bin is the name of a resource whose access needs to be throttled. For example, the Web connection type uses a document's server name as the throttling bin associated with the document, since (presumably) it will be access to each individual server that will need to be throttled independently.

On the repository connection "Throttling" tab, you can specify an unrestricted number of throttling descriptions. Each throttling description consists of a regular expression that describes a family of throttling bins, plus a helpful description, plus an average number of fetches per minute for each of the throttling bins that matches the regular expression. If a given throttling bin matches more than one throttling description, the most conservative fetch rate is chosen.

The simplest regular expression you can use is the empty regular expression. This will match all of the connection's throttle bins, and thus will allow you to specify a default throttling policy for the connection. Set the desired average fetch rate, and click the "Add" button. The throttling tab will then appear something like this:


Lucene Connectors Framework: Edit Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Apr 13, 2010 9:24:41 AM
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Throttling
Email
Robots
Bandwidth
Proxy
Edit a Connection
Delete
Add
Save
Cancel
Done

Max connections (per JVM): 100

Throttling:

	Bin regular expression	Description	Max avg fetches/min
Delete		All individual domains	6
Add			

If no throttle descriptions are added, no fetch-rate throttling will be performed.

Please refer to the section of the manual describing your chosen repository connection type for a description of the tabs appropriate for that connection type.

After you save your connection, a summary screen will be displayed that describes your connection's configuration. This summary screen contains a line where the connection's status is displayed. If you did everything correctly, the message "Connection working" will be displayed as a status. If there was a problem, you will see a connection-type-specific diagnostic message instead. If this happens, you will need to correct the problem, by either fixing your infrastructure, or by editing the connection configuration appropriately, before the authority connection will work correctly.

1.4 Creating Jobs

A "job" in ManifoldCF is a description of a set of documents. The Framework's job is to fetch this set of documents come from a specific repository connection, and send them to a specific output connection. The repository connection that is associated with the job will determine exactly how this set of documents is described, and to some degree how they are indexed. The output connection associated with the job can also affect how each document is indexed.

Every job is expected to be run more than once. Each time a job is run, it is responsible not only for sending new or changed documents to the output connection, but also for notifying the output connection of any documents that are no longer part of the set. Note that there are two ways for a document to no longer be part of the included set of documents: Either the document may have been deleted from the repository, or the document may no longer be included in the allowed set of documents. The Framework handles each case properly.

Deleting a job causes the output connection to be notified of deletion for all documents belonging to that job. This makes sense because the job represents the set of documents, which would otherwise be orphaned when the job was removed. (Some users make the assumption that a ManifoldCF job represents nothing more than a task, which is an incorrect assumption.)

Note that the Framework allows jobs that describe overlapping sets of documents to be defined. Documents that exist in more than one job are treated in the following special ways:

- When a job is deleted, the output connection is notified of deletion of documents belonging to that job only if they don't belong to another job
- The version of the document sent to the output connection depends on which job was run last

The subtle logic of overlapping documents means that you probably want to avoid this situation entirely, if it is at all feasible.

A typical non-continuous run of a job has the following stages of execution:

1. Adding the job's new, changed, or deleted starting points to the queue ("seeding")
2. Fetching documents, discovering new documents, and detecting deletions
3. Removing no-longer-included documents from the queue

Jobs can also be run "continuously", which means that the job never completes, unless it is aborted. A continuous run has different stages of execution:

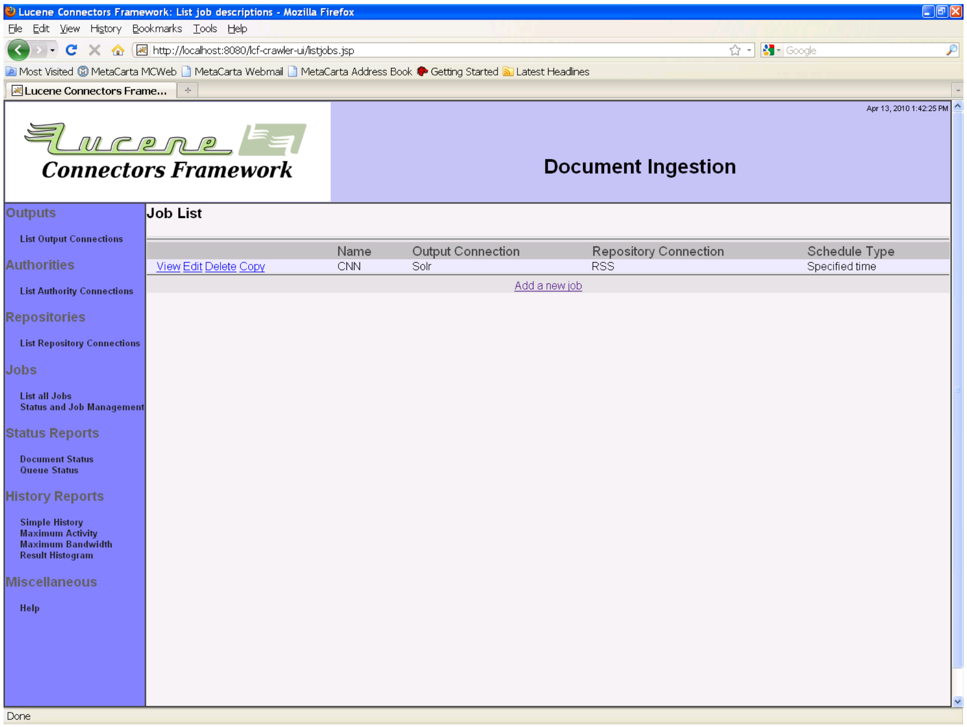
1. Adding the job's new, changed, or deleted starting points to the queue ("seeding")
2. Fetching documents, discovering new documents, and detecting deletions, while reseeding periodically

Note that continuous jobs cannot remove no-longer-included documents from the queue. They can only remove documents that have been deleted from the repository.

A job can independently be configured to start when explicitly started by a user, or to run on a user-specified schedule. If a job is set up to run on a schedule, it can be made to start only at the beginning of a schedule window, or to start again within any remaining schedule window when the previous job run completes.

There is no restriction in ManifoldCF as to how many jobs many running at any given time.

You create a job by first clicking on the "List All Jobs" link on the left-side menu. The following screen will appear:

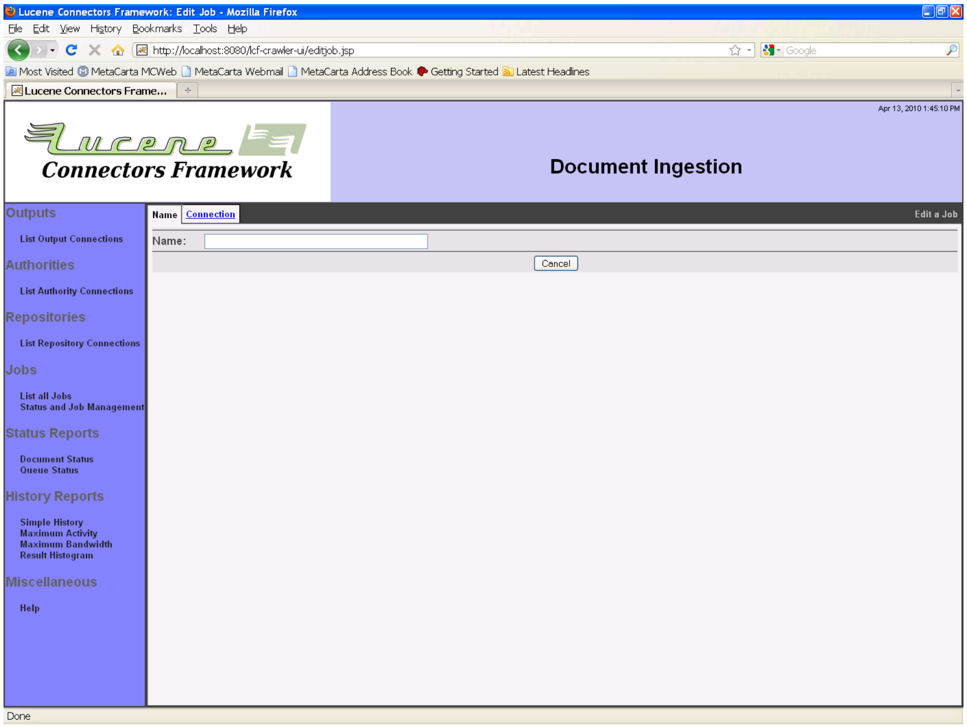


Job List

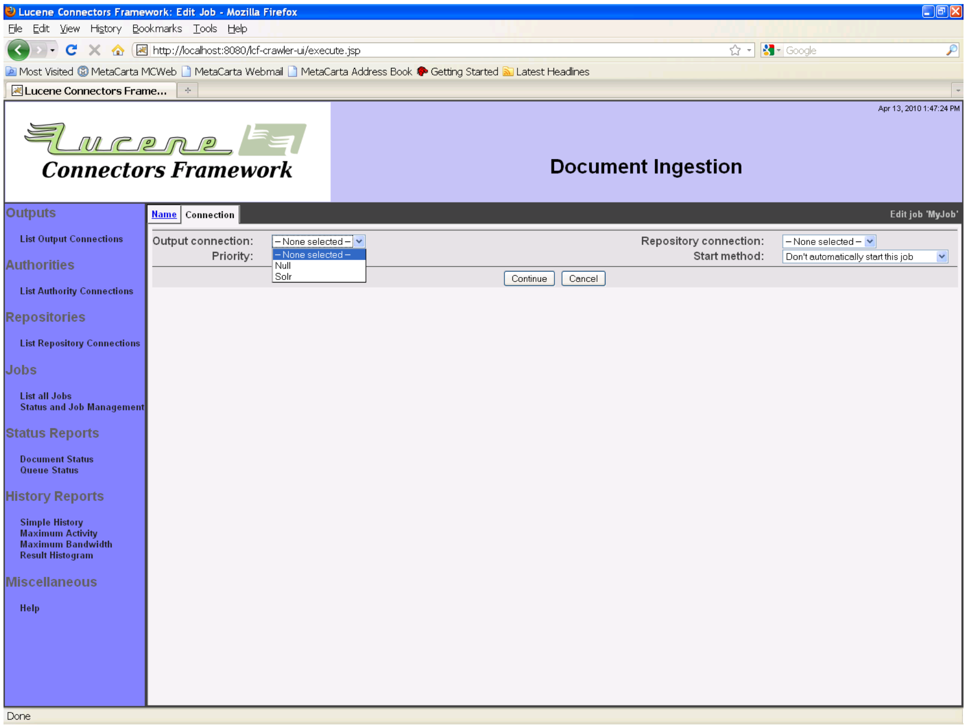
	Name	Output Connection	Repository Connection	Schedule Type
View Edit Delete Copy	CNN	Solr	RSS	Specified time

Add a new job

You may view, edit, or delete any existing jobs by clicking on the appropriate link. You may also create a new job that is a copy of an existing job. But to create a brand-new job, click the "Add a new job" link at the bottom. You will then see the following page:



Give your job a name. Note that job names do not have to be unique, although it is probably less confusing to have a different name for each one. Then, click the "Connection" tab:



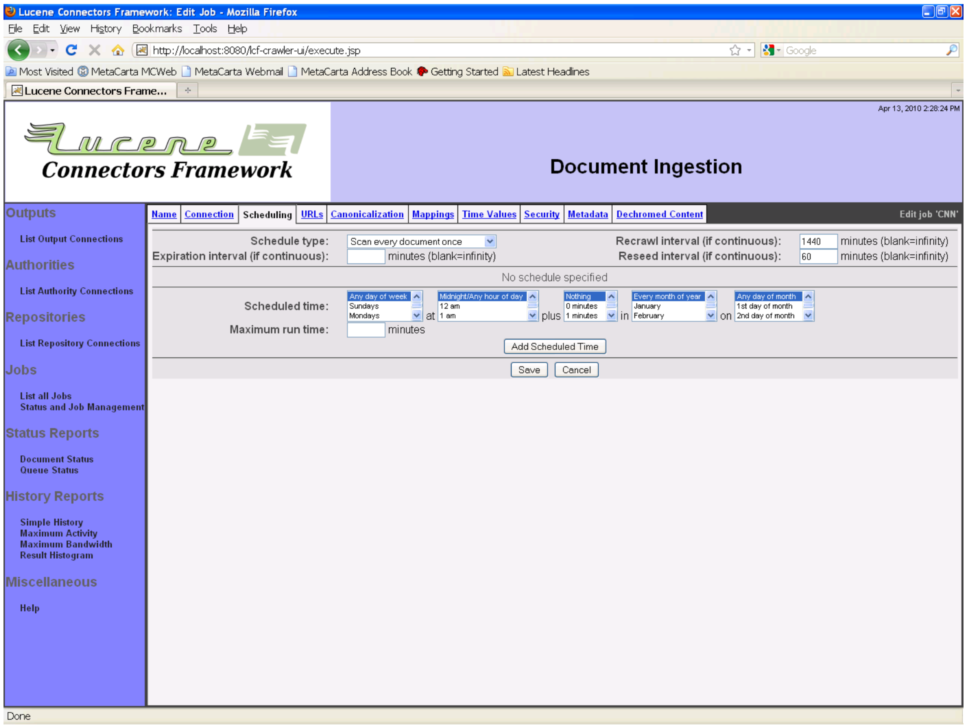
Output connection: – None selected – (– None selected – / Null / Solr)
Priority:
Repository connection: – None selected –
Start method: Don't automatically start this job

Now, you should select both the output connection name, and the repository connection name. Bear in mind that whatever you select cannot be changed after the job is saved the first time.

You also have the opportunity to modify the job's priority and start method at this time. The priority controls how important this job's documents are, relative to documents from any other job. The higher the number, the more important it is considered for that job's documents to be fetched first. The start method is as previously described; you get a choice of manual start, starting on the beginning of a scheduling window, or starting whenever possible within a scheduling window.

Make your selections, and click "Continue". The rest of the job's tabs will now appear, and a "Save" button will also appear at the bottom of the pane. You must click the "Save" button when you are done in order to create or update your job. If you click "Cancel" instead, the new job will not be created. (The same thing will happen if you click on any of the navigation links in the left-hand pane.)

All jobs have a "Scheduling" tab. The scheduling tab allows you to set up schedule-related configuration information:


Lucene Connectors Framework: Edit Job - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Apr 13, 2010 2:28:24 PM
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Connection
Scheduling
URLs
Canonicalization
Mappings
Time Values
Security
Metadata
Dechromed Content
Edit job 'CNN'
Add Scheduled Time
Save
Cancel
Done

Document Ingestion

Schedule type: Scan every document once
Expiration interval (if continuous): minutes (blank=infinity)
Recrawl interval (if continuous): 1440 minutes (blank=infinity)
Reseed interval (if continuous): 60 minutes (blank=infinity)

No schedule specified

Scheduled time: Any day of week / Sundays / Mondays at Midnight/Any hour of day / 12 am / 1 am plus Nothing / 0 minutes / 1 minutes in Every month of year / January / February on Any day of month / 1st day of month / 2nd day of month
Maximum run time: minutes

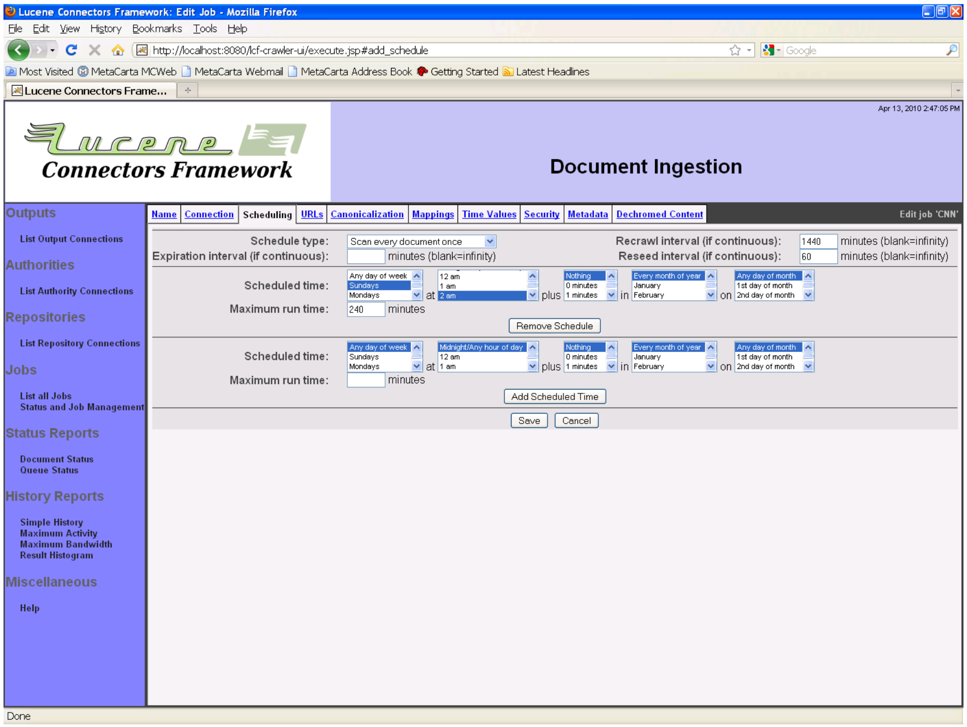
On this tab, you can specify the following parameters:

- Whether the job runs continuously, or scans every document once
- How long a document should remain alive before it is 'expired', and removed from the index
- How long an interval before a document is re-checked, to see if it has changed
- How long to wait before reseeding initial documents

The last three parameters only make sense if a job is a continuously running one, as the UI indicates.

The other thing you can do on this time is to define an appropriate set of scheduling records. Each scheduling record defines some related set of intervals during which the job can run. The intervals are determined by the starting time (which is defined by the day of week, month, day, hour, and minute pulldowns), and the maximum run time in minutes, which determines when the interval ends. It is, of course, possible to select multiple values for each of the pulldowns, in which case you be describing a starting time that had to match at least one of the selected values for each of the specified fields.

Once you have selected the schedule values you want, click the "Add Scheduled Time" button:



Lucene Connectors Framework: Edit Job - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp#add_schedule
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 13, 2010 2:47:05 PM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name Connection Scheduling URLs Canonicalization Mappings Time Values Security Metadata Dechromed Content
Edit job 'CNN'
Remove Schedule
Add Scheduled Time
Save
Cancel
Done

Schedule type: Scan every document once
Expiration interval (if continuous): minutes (blank=infinity)
Recrawl interval (if continuous): 1440 minutes (blank=infinity)
Reseed interval (if continuous): 60 minutes (blank=infinity)

Scheduled time: Sundays at 2 am plus Nothing in Every month of year on Any day of month
Maximum run time: 240 minutes

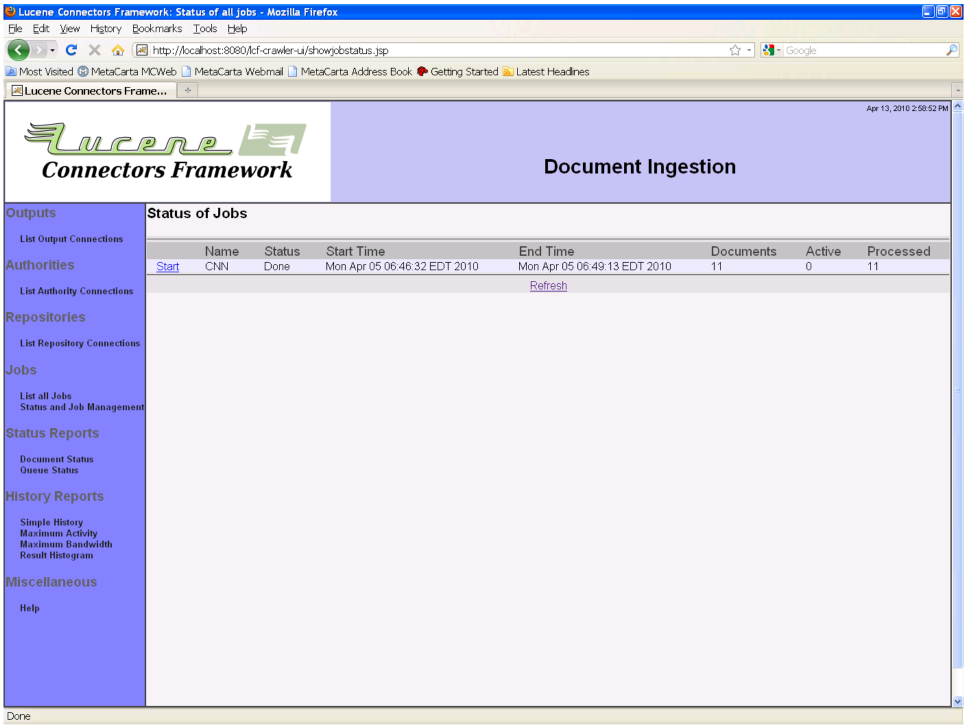
Scheduled time: Any day of week at Midnight/Any hour of day plus Nothing in Every month of year on Any day of month
Maximum run time: minutes

The example shows a schedule where crawls are run on Saturday and Sunday nights at 2 AM, and run for no more than 4 hours.

The rest of the job tabs depend on the types of the connections you selected. Please refer to the section of the manual describing the appropriate connection types corresponding to your chosen repository and output connections for a description of the job tabs that will appear for those connections.

1.5 Executing Jobs

You can follow what is going on, and control the execution of your jobs, by clicking on the "Status and Job Management" link on the left-side navigation menu. When you do, you might see something like this:


Lucene Connectors Framework: Status of all jobs - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/showjobstatus.jsp
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 13, 2010 2:58:52 PM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

Status of Jobs

	Name	Status	Start Time	End Time	Documents	Active	Processed
Start	CNN	Done	Mon Apr 05 06:46:32 EDT 2010	Mon Apr 05 06:49:13 EDT 2010	11	0	11

Refresh

From here, you can click the "Refresh" link at the bottom of the main pane to see an updated status display, or you can directly control the job using the links in the leftmost status column. Allowed actions you may see at one point or another include:

- Start (start the job)
- Abort(abort the job)
- Pause (pause the job)
- Resume (resume the job)
- Restart (equivalent to aborting the job, and starting it all over again)

The columns "Documents", "Active", and "Processed" have very specific means as far as documents in the job's queue are concerned. The "Documents" column counts all the documents that belong to the job. The "Active" column counts all of the documents for that job that are queued up for processing. The "Processed" column counts all documents that are on the queue for the job that have been processed at least once in the past.

1.6 Status Reports

Every job in ManifoldCF describes a set of documents. A reference to each document in the set is kept in a job-specific queue. It is sometimes valuable for diagnostic reasons to examine this queue for information. The Framework UI has several canned reports which do just that.

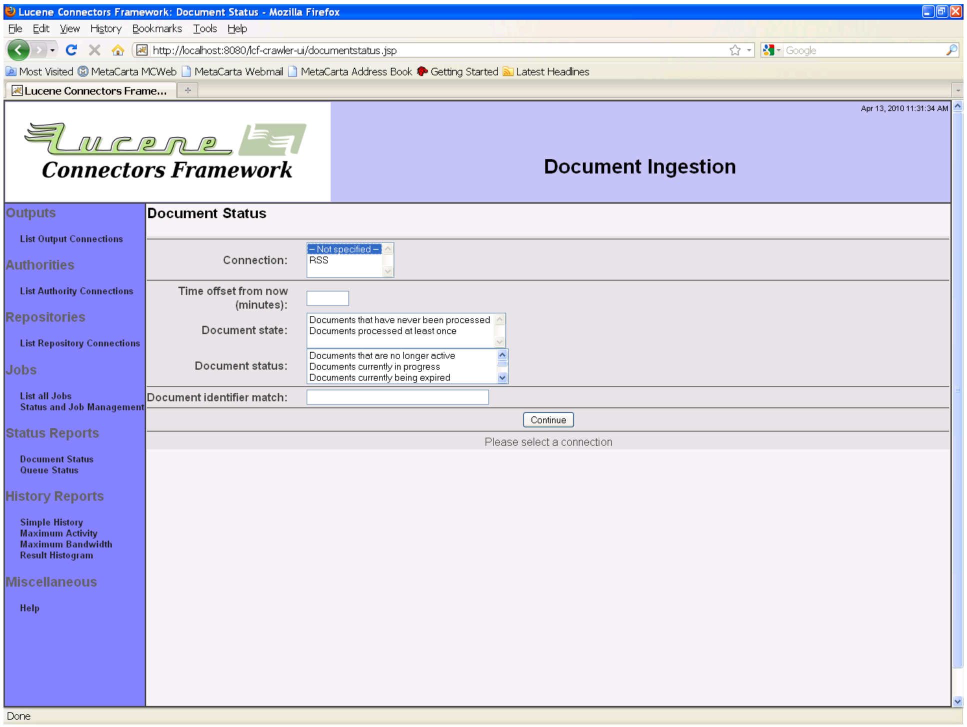
Each status report allows you to select what documents you are interested in from a job's queue based on the following information:

- The job
- The document identifier
- The document's status and state
- When the document is scheduled to be processed next

1.6.1 Document Status

A document status report simply lists all matching documents from within the queue, along with their state, status, and planned future activity. You might use this report if you were trying to figure out (for example) whether a specific document had been processed yet during a job run.

Click on the "Document Status" link on the left-hand menu. You will see a screen that looks something like this:


Lucene Connectors Framework: Document Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/documentstatus.jsp
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 13, 2010 11:31:34 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Continue
Done

Document Status

Connection: -- Not specified -- / RSS

Time offset from now (minutes):

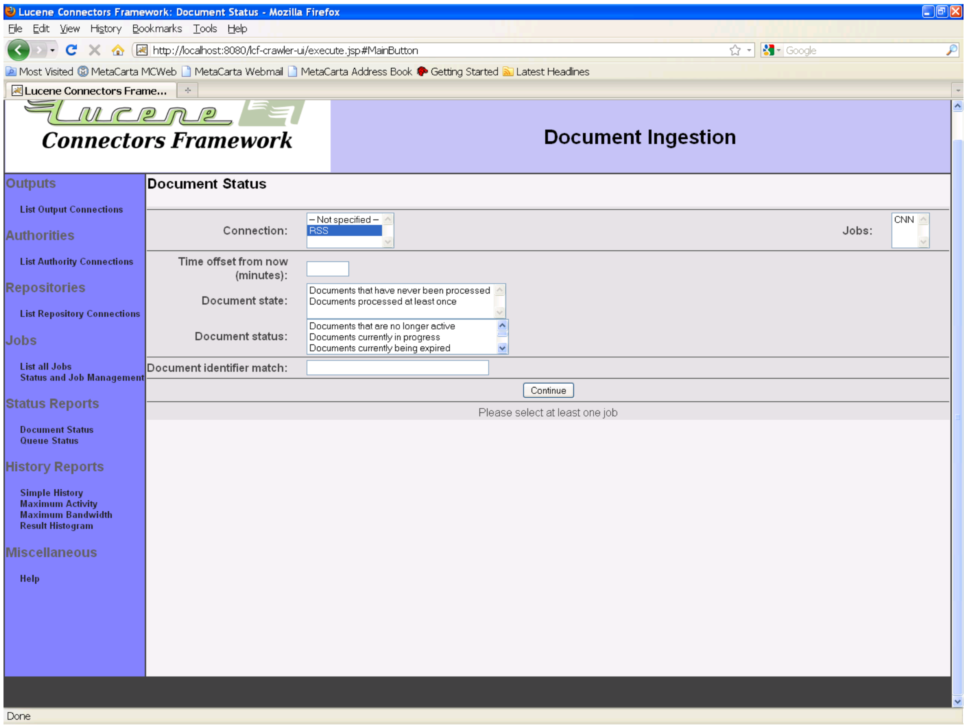
Document state: Documents that have never been processed / Documents processed at least once

Document status: Documents that are no longer active / Documents currently in progress / Documents currently being expired

Document identifier match:

Please select a connection

Select the desired connection. You may also select the desired document state and status, as well as specify a regular expression for the document identifier, if you want. Then, click the "Continue" button:

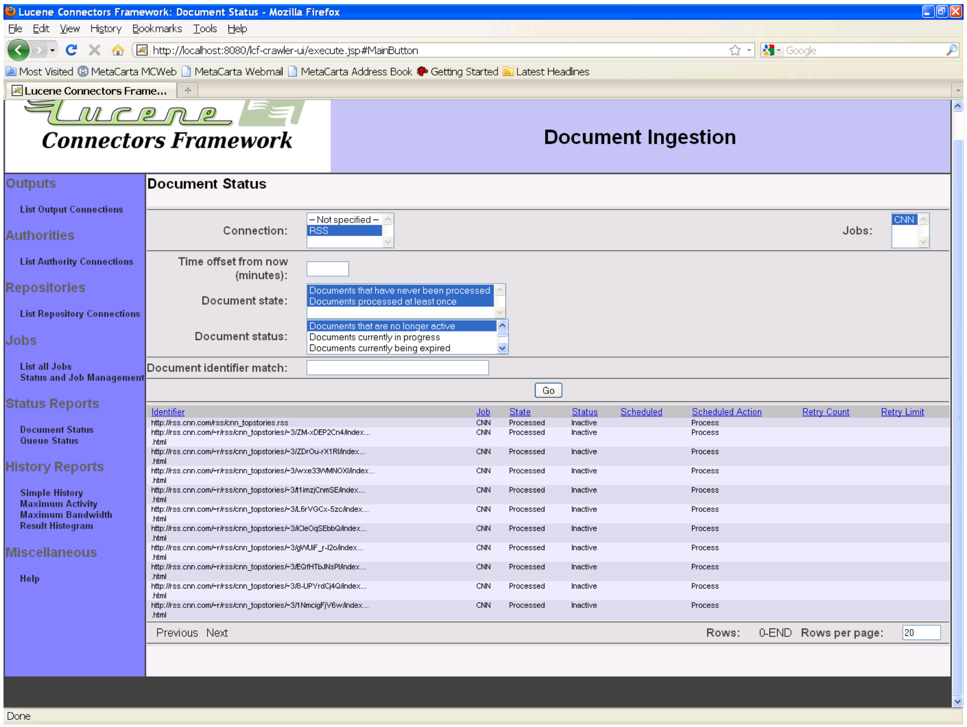

Lucene Connectors Framework: Document Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp#MainButton
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Continue
Done

Document Status

Connection:	-- Not specified -- RSS
Jobs:	CNN
Time offset from now (minutes):	
Document state:	Documents that have never been processed Documents processed at least once
Document status:	Documents that are no longer active Documents currently in progress Documents currently being expired
Document identifier match:	

Please select at least one job

Select the job whose documents you want to see, and click "Continue" once again. The results will display:


Lucene Connectors Framework: Document Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp#MainButton
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

Document Status

Connection: -- Not specified -- / RSS (selected)

Jobs: CNN (selected)

Time offset from now (minutes):

Document state: Documents that have never been processed (selected), Documents processed at least once (selected)

Document status: Documents that are no longer active (selected), Documents currently in progress, Documents currently being expired

Document identifier match:

Go

Identifier	Job	State	Status	Scheduled	Scheduled Action	Retry Count	Retry Limit
http://rss.cnn.com/rss/cnn_topstories.rss	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/ZM-xDEP2Cn4/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/ZDrOu-rX1RI/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/wxe33WMNOXI/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/t1imzjCnmSE/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/L6rVGCx-5zc/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/iCleOqSEbbQ/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/gWUiF_r-I2o/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/EQfHTbJNsPI/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/8-UPYrdCj4Q/index... .html	CNN	Processed	Inactive		Process		
http://rss.cnn.com/~r/rss/cnn_topstories/~3/1NmcigFjV6w/index... .html	CNN	Processed	Inactive		Process		

Previous Next Rows: 0-END Rows per page: 20

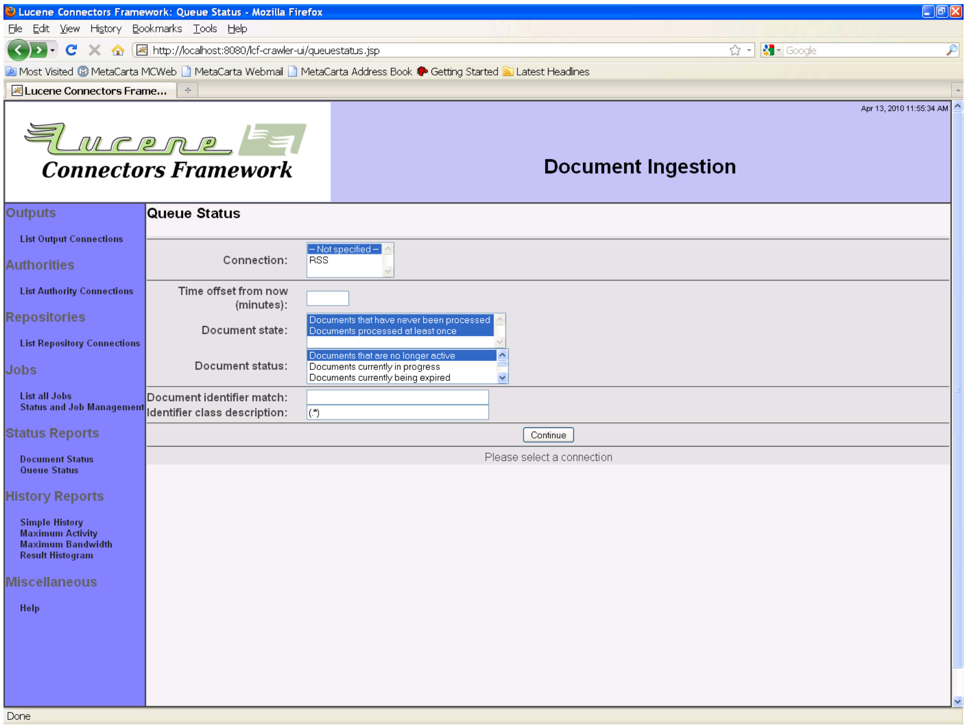
You may alter the criteria, and click "Go" again, if you so choose. Or, you can alter the number of result rows displayed at a time, and click "Go" to redisplay. Finally, you can page up and down through the results using the "Prev" and "Next" links.

1.6.2 Queue Status

A queue status report is an aggregate report that counts the number of occurrences of documents in specified classes. The classes are specified as a grouping within a regular expression, which is matched against all specified document identifiers. The results that are displayed are counts of documents. There will be a column for each combination of document state and status.

For example, a class specification of "()" will produce exactly one result row, and will provide a count of documents that are in each state/status combination. A class description of "(.*)", on the other hand, will create one row for each document identifier, and will put a "1" in the column representing state and status of that document, with a "0" in all other column positions.

Click the "Queue Status" link on the left-hand menu. You will see a screen that looks like this:


Lucene Connectors Framework: Queue Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/queuestatus.jsp
Google
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Apr 13, 2010 11:55:34 AM
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

Queue Status

Connection: – Not specified – / RSS

Time offset from now (minutes):

Document state: Documents that have never been processed / Documents processed at least once

Document status: Documents that are no longer active / Documents currently in progress / Documents currently being expired

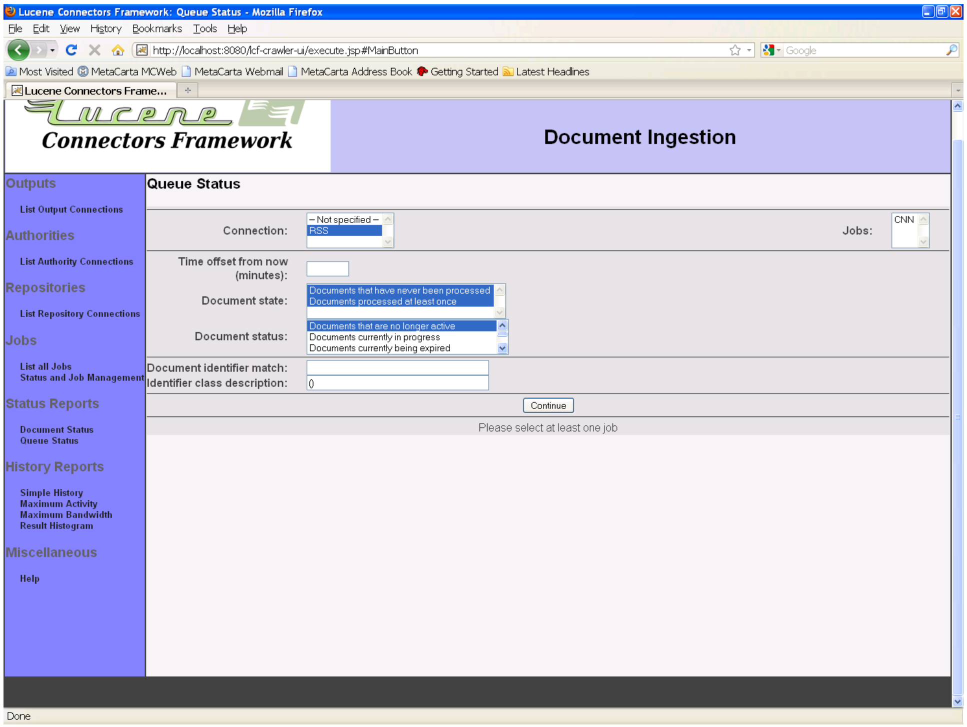
Document identifier match:

Identifier class description: (.*)

Continue

Please select a connection

Select the desired connection. You may also select the desired document state and status, as well as specify a regular expression for the document identifier, if you want. You will probably want to change the document identifier class from its default value of "(.*)". Then, click the "Continue" button:

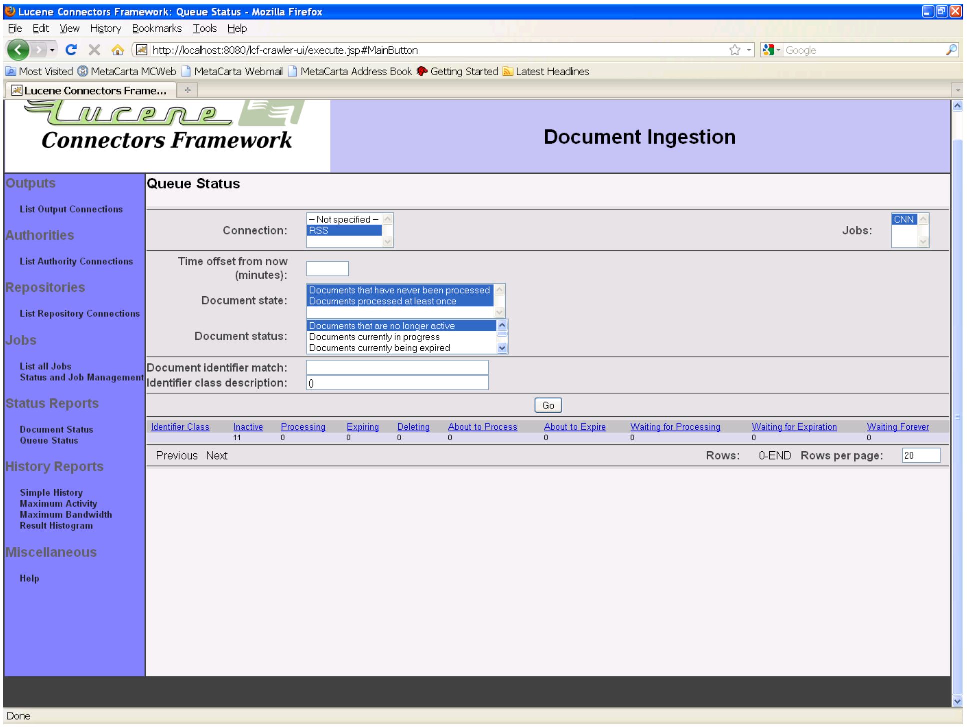


Queue Status

Connection:	-- Not specified -- RSS	Jobs:	CNN
Time offset from now (minutes):			
Document state:	Documents that have never been processed Documents processed at least once		
Document status:	Documents that are no longer active Documents currently in progress Documents currently being expired		
Document identifier match:			
Identifier class description:	()		

Please select at least one job

Select the job whose documents you want to see, and click "Continue" once again. The results will display:


Lucene Connectors Framework: Queue Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp#MainButton
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Go
Done

Queue Status

Connection: – Not specified – / RSS (selected)

Jobs: CNN (selected)

Time offset from now (minutes):

Document state: Documents that have never been processed (selected) / Documents processed at least once (selected)

Document status: Documents that are no longer active (selected) / Documents currently in progress / Documents currently being expired

Document identifier match:

Identifier class description: ()

Identifier Class	Inactive	Processing	Expiring	Deleting	About to Process	About to Expire	Waiting for Processing	Waiting for Expiration	Waiting Forever
	11	0	0	0	0	0	0	0	0

Previous Next Rows: 0-END Rows per page: 20

You may alter the criteria, and click "Go" again, if you so choose. Or, you can alter the number of result rows displayed at a time, and click "Go" to redisplay. Finally, you can page up and down through the results using the "Prev" and "Next" links.

1.7 History Reports

For every repository connection, ManifoldCF keeps a history of what has taken place involving that connection. This history includes both events that the framework itself logs, as well as events that a repository connection or output connection will log. These individual events are categorized by "activity type". Some of the kinds of activity types that exist are:

- Job start
- Job end
- Job abort
- Various connection-type-specific read or access operations
- Various connection-type-specific output or indexing operations

This history can be enormously helpful in understand how your system is behaving, and whether or not it is working properly. For this reason, the

Framework UI has the ability to generate several canned reports which query this history data and display the results.

All history reports allow you to specify what history records you are interested in including. These records are selected using the following criteria:

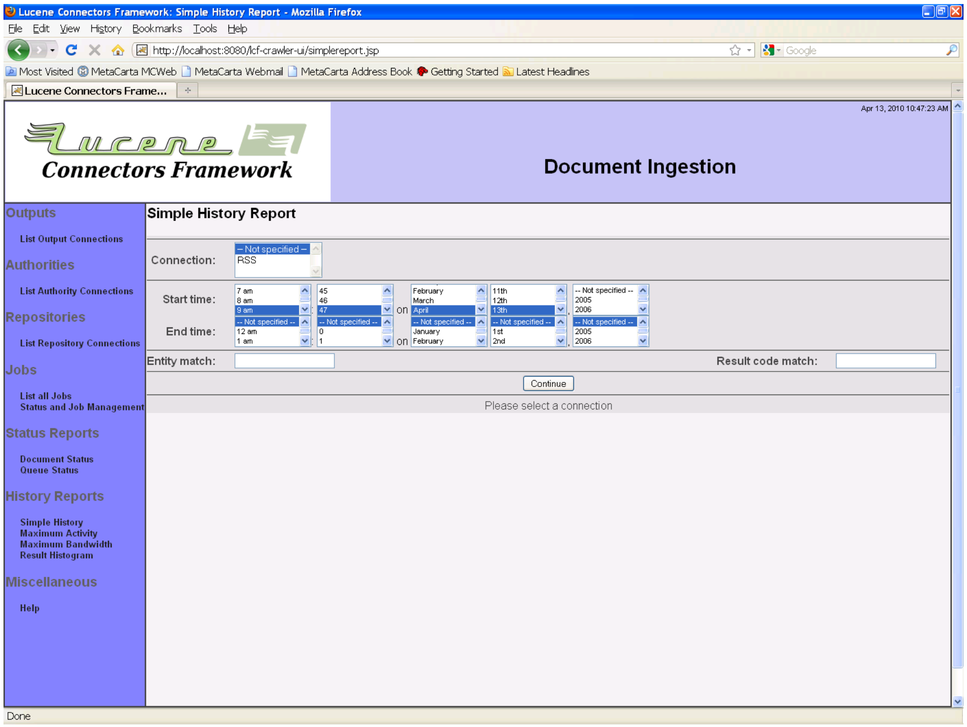
- The repository connection name
- The activity type(s) desired
- The start time desired
- The end time desired
- The identifier(s) involved, specified as a regular expression
- The result(s) produced, specified as a regular expression

The actual reports available are designed to be useful for diagnosing both access issues, and performance issues. See below for a summary of the types available.

1.7.1 Simple History Reports

As the name suggests, a simple history report does not attempt to aggregate any data, but instead just lists matching records from the repository connection's history. These records are initially presented in most-recent-first order, and include columns for the start and end time of the event, the kind of activity represented by the event, the identifier involved, the number of bytes involved, and the results of the event. Once displayed, you may choose to display more or less data, or reorder the display by column, or page through the data.

To get started, click on the "Simple History" link on the left-hand menu. You will see a screen that looks like this:



Simple History Report

Connection: -- Not specified -- | RSS

Start time: 9 am : 47 on April 13th , -- Not specified --

End time: -- Not specified -- : -- Not specified -- on -- Not specified -- -- Not specified -- , -- Not specified --

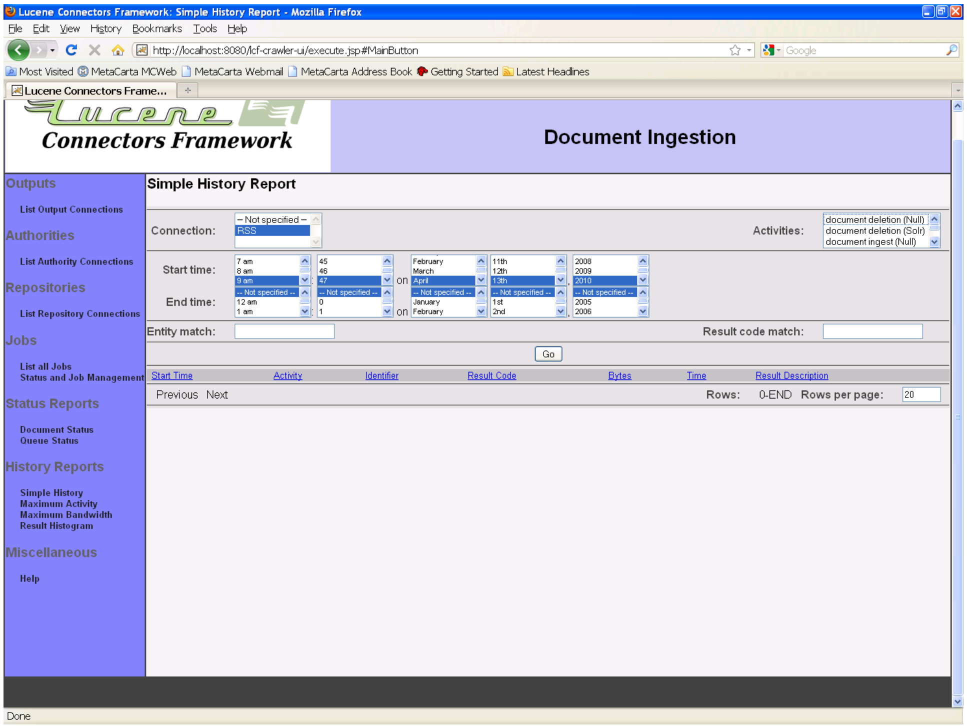
Entity match:

Result code match:

Please select a connection

Now, select the desired repository connection from the pulldown in the upper left hand corner. If you like, you can also change the specified date/time range, or specify an identifier regular expression or result code regular expression. By default, the date/time range selects all events within the last hour, while the identifier regular expression and result code regular expression matches all identifiers and result codes.

Next, click the "Continue" button. A list of pertinent activities should then appear in a pulldown in the upper right:


Lucene Connectors Framework: Simple History Report - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp#MainButton
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

Document Ingestion

Simple History Report

Connection: -- Not specified -- / RSS (selected)

Activities: document deletion (Null), document deletion (Solr), document ingest (Null)

Start time: 9 am : 47 on April 13th, 2010

End time: -- Not specified -- : -- Not specified -- on -- Not specified -- -- Not specified --, -- Not specified --

Entity match:

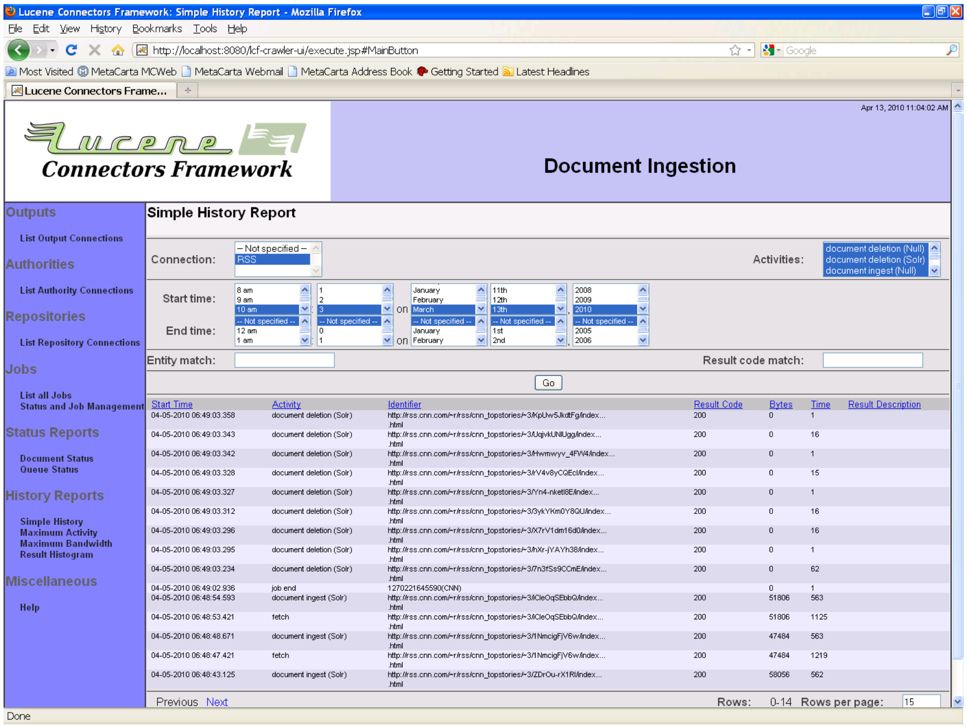
Result code match:

Go

Start Time	Activity	Identifier	Result Code	Bytes	Time	Result Description

Previous Next Rows: 0-END Rows per page: 20

You may select one or more activities that you would like a report on. When you are done, click the "Go" button. The results will appear, ordered by time, most recent event first:


Lucene Connectors Framework: Simple History Report - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp#MainButton
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Apr 13, 2010 11:04:02 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Go
Previous Next
Done

Simple History Report

Connection: – Not specified – / RSS (selected)

Activities: document deletion (Null), document deletion (Solr), document ingest (Null) (selected)

Start time: 10 am : 3 on March 13th, 2010

End time: -- Not specified -- : -- Not specified -- on -- Not specified -- -- Not specified --, -- Not specified --

Entity match:

Result code match:

Start Time	Activity	Identifier	Result Code	Bytes	Time	Result Description
04-05-2010 06:49:03.358	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/KpUw5JkdtFg/index... .html	200	0	1	
04-05-2010 06:49:03.343	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/UqjvkUNlUgg/index... .html	200	0	16	
04-05-2010 06:49:03.342	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/Hwmwyv_4FW4/index... .html	200	0	1	
04-05-2010 06:49:03.328	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/rV4v8yCQEcI/index... .html	200	0	15	
04-05-2010 06:49:03.327	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/Yn4-nketI8E/index... .html	200	0	1	
04-05-2010 06:49:03.312	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/3ykYKm0Y8QU/index... .html	200	0	16	
04-05-2010 06:49:03.296	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/X7rV1dm16d0/index... .html	200	0	16	
04-05-2010 06:49:03.295	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/hXr-jYAYh38/index... .html	200	0	1	
04-05-2010 06:49:03.234	document deletion (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/7n3fSs9CCmE/index... .html	200	0	62	
04-05-2010 06:49:02.936	job end	1270221645590(CNN)		0	1	
04-05-2010 06:48:54.593	document ingest (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/iCleOqSEbbQ/index... .html	200	51806	563	
04-05-2010 06:48:53.421	fetch	http://rss.cnn.com/~r/rss/cnn_topstories/~3/iCleOqSEbbQ/index... .html	200	51806	1125	
04-05-2010 06:48:48.671	document ingest (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/1NmcigFjV6w/index... .html	200	47484	563	
04-05-2010 06:48:47.421	fetch	http://rss.cnn.com/~r/rss/cnn_topstories/~3/1NmcigFjV6w/index... .html	200	47484	1219	
04-05-2010 06:48:43.125	document ingest (Solr)	http://rss.cnn.com/~r/rss/cnn_topstories/~3/ZDrOu-rX1RI/index... .html	200	58056	562	

Rows: 0-14 Rows per page: 15

You may alter the criteria, and click "Go" again, if you so choose. Or, you can alter the number of result rows displayed at a time, and click "Go" to redisplay. Finally, you can page up and down through the results using the "Prev" and "Next" links.

Please bear in mind that the report redisplays whatever matches each time you click "Go". So, if your time interval goes from an hour beforehand to "now", and you have activity happening, you will see different results each time "Go" is clicked.

1.7.2 Maximum Activity Reports

A maximum activity report is an aggregate report used primarily to display the maximum rate that events occur within a specified time interval. MHL

1.7.3 Maximum Bandwidth Reports

A maximum bandwidth report is an aggregate report used primarily to display the maximum byte rate that pertains to events occurring within a specified time interval. MHL

1.7.4 Result Histogram Reports

A result histogram report is an aggregate report is used to count the occurrences of each kind of matching result for all matching events. MHL

1.8 A Note About Credentials

If any of your selected connection types require credentials, you may find it necessary to approach your system administrator to obtain an appropriate set. System administrators are often reluctant to provide accounts and credentials that have any more power than is utterly necessary, and sometimes not even that. Great care has been taken in the development of all connection types to be sure they require no more privilege than is utterly necessary. If a security-related warning appears when you view a connection's status, you must inform the system administrator that the credentials are inadequate to allow the connection to accomplish its task, and work with him/her to correct the problem.

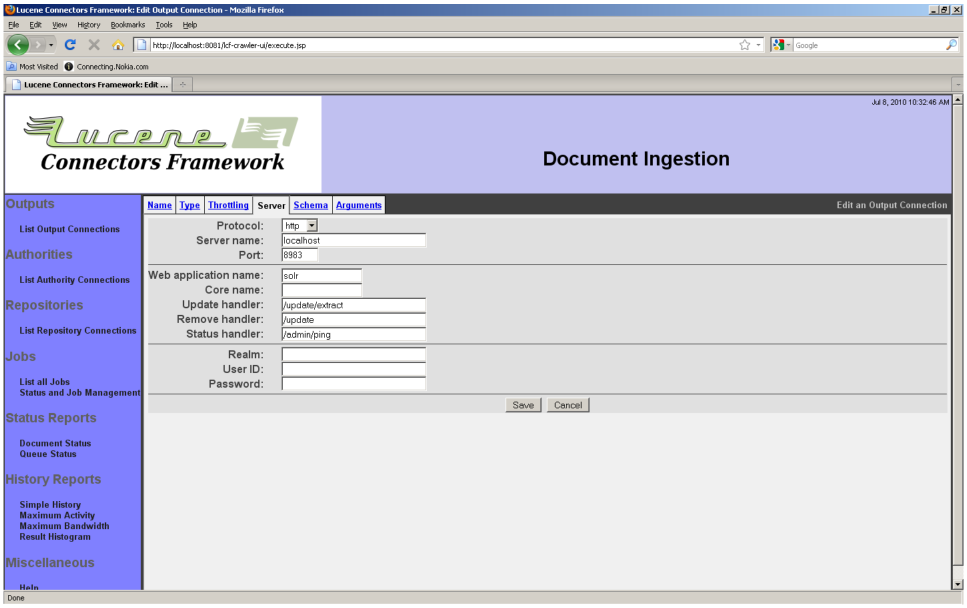
2 Output Connection Types

2.1 Solr Output Connection

The Solr output connection type is designed to allow ManifoldCF to submit documents to an appropriate Solr pipeline, via the Solr HTTP ingestion API. The configuration parameters are set to the default Solr values, which can be changed (since Solr's configuration can be changed). The Solr output connection type furthermore makes no judgment as to whether a given document is indexable or not - it accepts everything, and passes all documents on to the pipeline, where presumably the configured pipeline will decide if a document should be rejected or not. (All of that happens without a Solr connection being aware of it in any way.)

Unfortunately, this lack of specificity comes at a cost. Unless you take care to filter documents properly in each job, large movie files or other opaque content may well be picked up and sent to Solr for indexing, which will greatly increase the dead load on the overall system. It is therefore a good idea to review all crawls done through a Solr connection while they are underway, to be sure there isn't a misconfiguration of this kind.

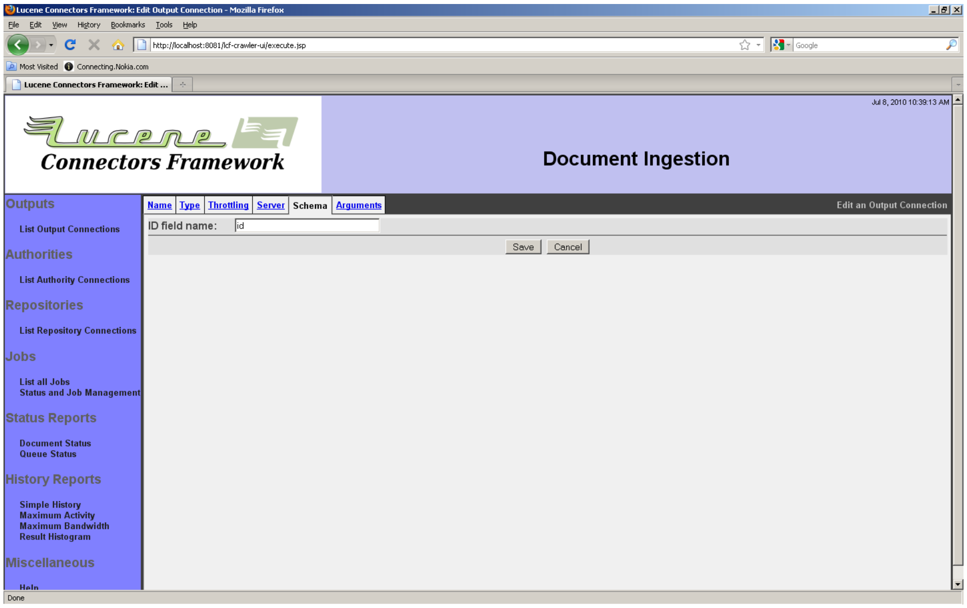
When you create a Solr output connection, five configuration tabs appear. The "Server" tab allows you to configure the HTTP target of the connection:



Protocol: http
Server name: localhost
Port: 8983
Web application name: solr
Core name:
Update handler: /update/extract
Remove handler: /update
Status handler: /admin/ping
Realm:
User ID:
Password:

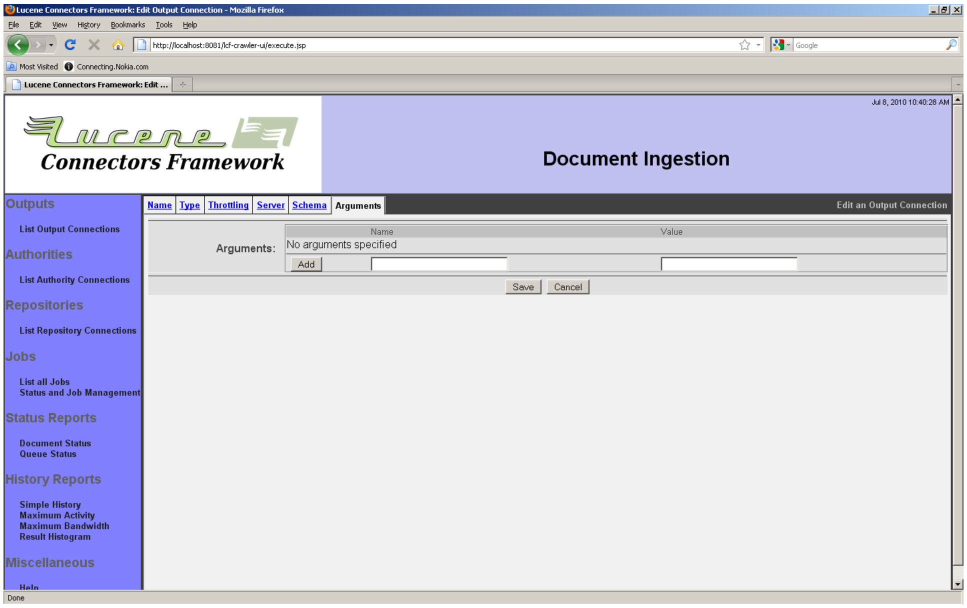
Fill in the fields according to your Solr configuration. The Solr connection type supports only basic authentication at this time; if you have this enabled, supply the credentials as requested on the bottom part of the form.

The second tab is the "Schema" tab, which allows you to specify the name of the Solr field to use as a document identifier. The Solr connection type will treat this field as being a unique key for locating the indexed document for further modification or deletion:



ID field name: id

The third tab is the "Arguments" tab, which allows you to specify arbitrary arguments to be sent to Solr. All valid Solr update request parameters can be specified here. You can for instance add update.chain=myChain to select the document processing pipeline/chain to use for processing documents in Solr. See the Solr documentation for more valid arguments. The tab looks like:

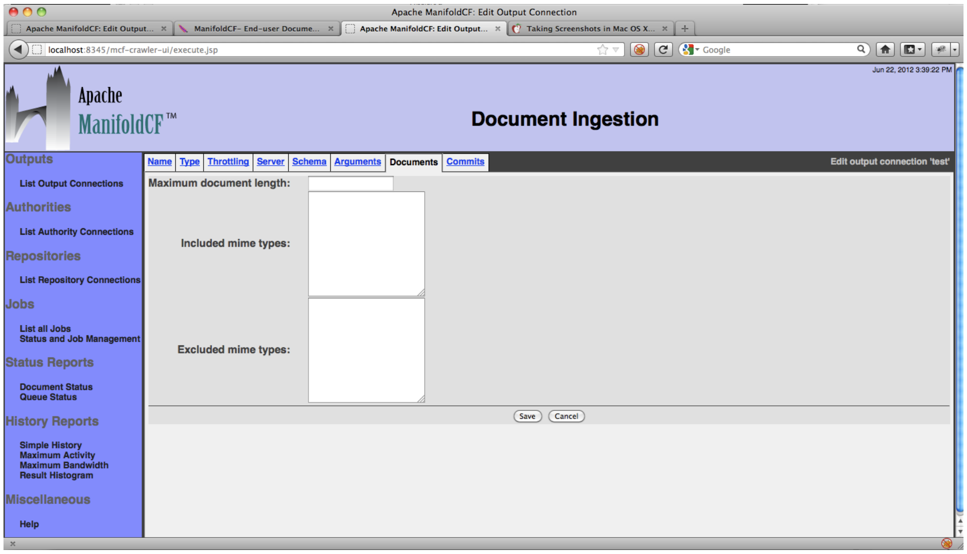


Arguments:

Name	Value
No arguments specified	

Fill in the argument name and value, and click the "Add" button. Bear in mind that if you add an argument with the same name as an existing one, it will replace the existing one with the new specified value. You can delete existing arguments by clicking the "Delete" button next to the argument you want to delete.

The fourth tab is the "Documents" tab, which allows you to do document filtering based on size and mime types. By specifying a maximum document length in bytes, you can filter out documents which exceed that size (e.g. 10485760 which is equivalent to 10 MB). If you only want to add documents with specific mime types, you can enter them into the "included mime types" field (e.g. "text/html" for filtering out all documents but HTML). The "excluded mime types" field is for excluding documents with specific mime types (e.g. "image/jpeg" for filtering out JPEG images). The tab looks like:



Document Ingestion

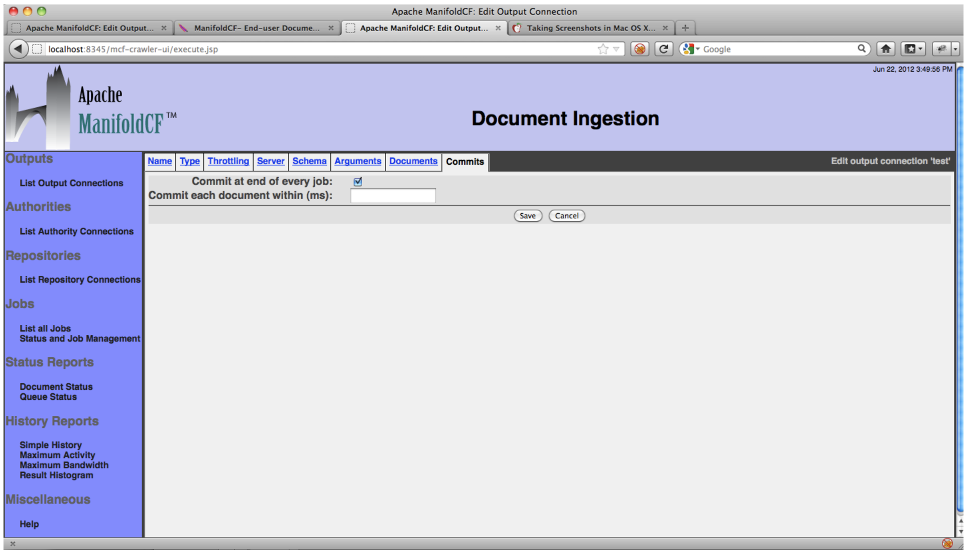
Edit output connection 'test'

Maximum document length:

Included mime types:

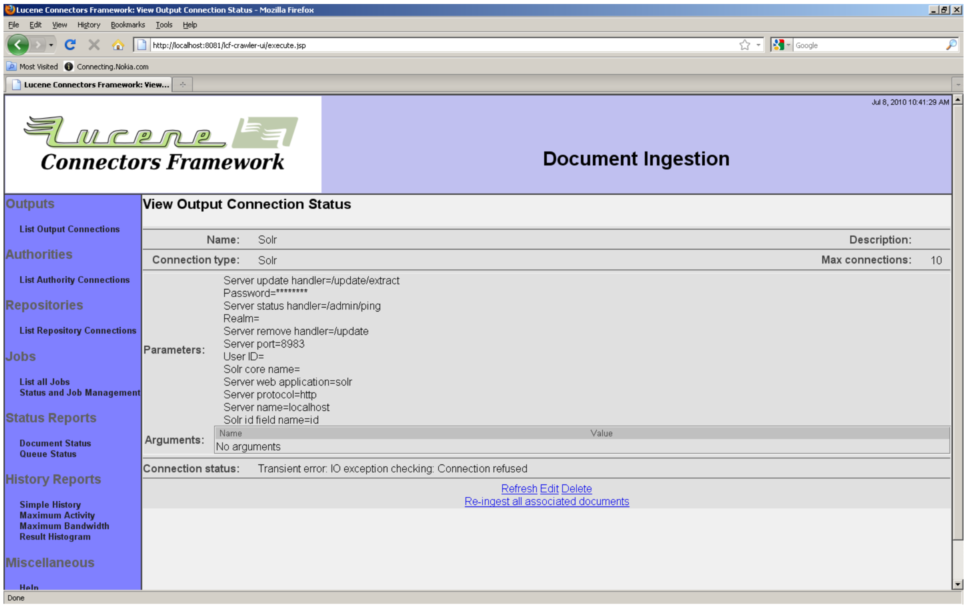
Excluded mime types:

The fifth tab is the "Commits" tab, which allows you to control the commit strategies. As well as committing documents at the end of every job, an option which is enabled by default, you may also commit each document within a certain time in milliseconds (e.g. "10000" for committing within 10 seconds). The commit within strategy will leave the responsibility to Solr instead of ManifoldCF. The tab looks like:



Commit at end of every job: ☑
Commit each document within (ms):

When you are done, don't forget to click the "Save" button to save your changes! When you do, a connection summary and status screen will be presented, which may look something like this:



View Output Connection Status

Name: Solr
Description:
Connection type: Solr
Max connections: 10

Parameters:
Server update handler=/update/extract
Password=********
Server status handler=/admin/ping
Realm=
Server remove handler=/update
Server port=8983
User ID=
Solr core name=
Server web application=solr
Server protocol=http
Server name=localhost
Solr id field name=id

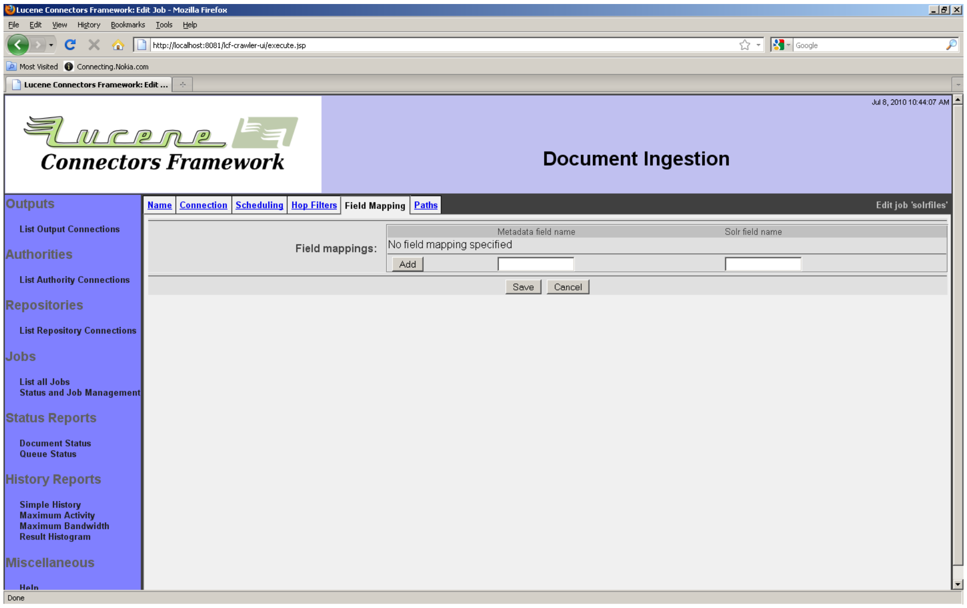
Arguments:

Name	Value
No arguments	

Connection status: Transient error: IO exception checking: Connection refused

Note that in this example, the Solr connection is not responding, which is leading to an error status message instead of "Connection working".

When you configure a job to use a Solr-type output connection, the Solr connection type provides a tab called "Field Mapping". The purpose of this tab is to allow you to map metadata fields as fetched by the job's connection type to fields that Solr is set up to receive. This is necessary because the names of the metadata items are often determined by the repository, with no alignment to fields defined in the Solr schema. You may also suppress specific metadata items from being sent to the index using this tab. The tab looks like this:



Lucene Connectors Framework: Edit Job - Mozilla Firefox

http://localhost:8081/lcf-crawler-ui/execute.jsp

Jul 8, 2010 10:44:07 AM

Document Ingestion

Edit job 'solrfiles'

Field mappings:

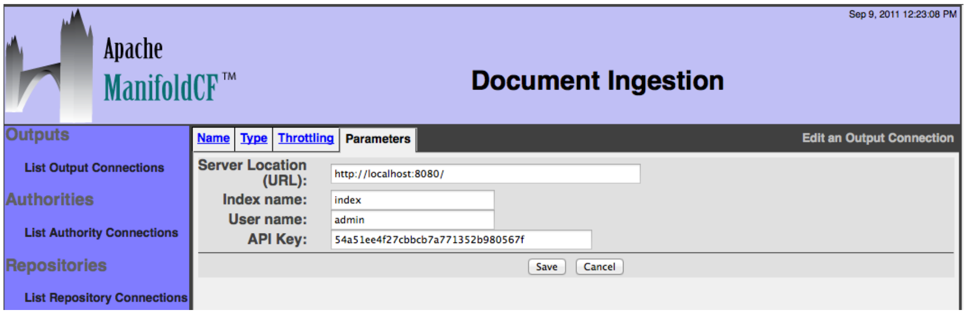
Metadata field name	Solr field name
No field mapping specified	

Add a new mapping by filling in the "source" with the name of the metadata item from the repository, and "target" as the name of the output field in Solr, and click the "Add" button. Leaving the "target" field blank will result in all metadata items of that name not being sent to Solr.

2.2 OpenSearchServer Output Connection

The OpenSearchServer Output Connection allow ManifoldCF to submit documents to an OpenSearchServer instance, via the XML over HTTP API. The connector has been designed to be as easy to use as possible.

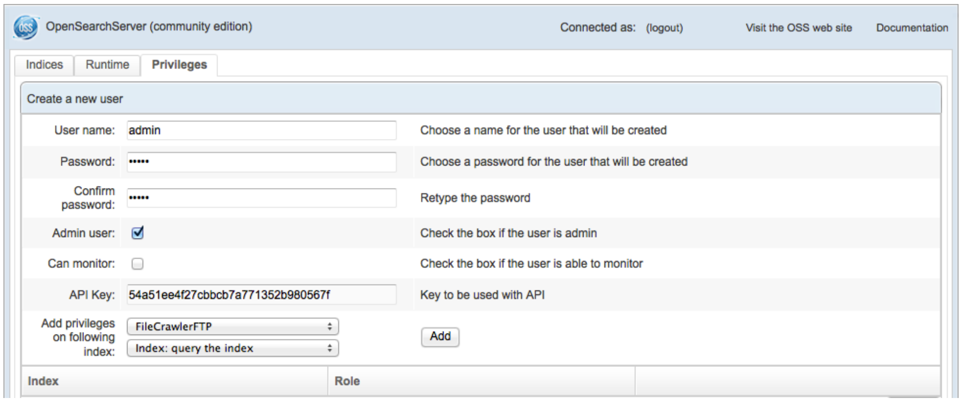
After creating an OpenSearchServer ouput connection, you have to populate the parameters tab. Fill in the fields according your OpenSearchServer configuration. Each OpenSearchServer output connector instance works with one index. To work with muliple indexes, just create one output connector for each index.



Server Location (URL): http://localhost:8080/
Index name: index
User name: admin
API Key: 54a51ee4f27cbbcb7a771352b980567f

The parameters are:

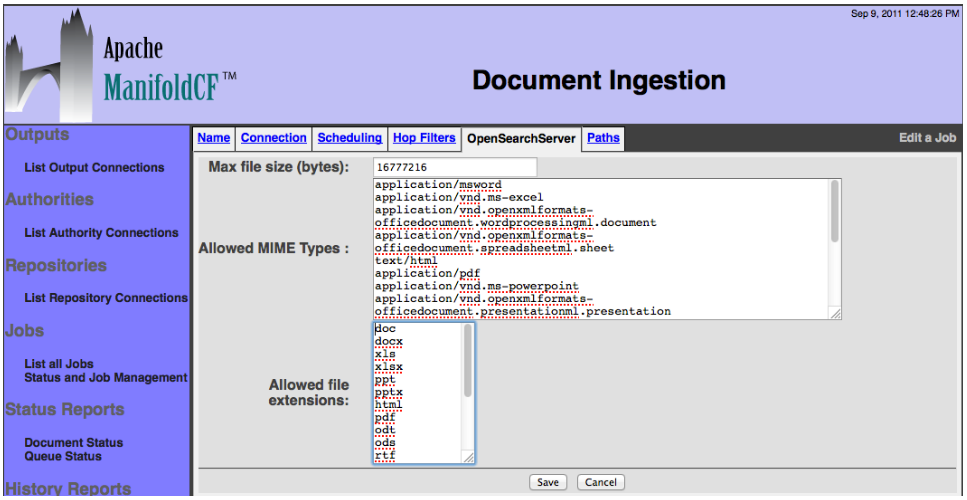
- Server location: An URL that references your OpenSearchServer instance. The default value (http://localhost:8080) is valid if your OpenSearchServer instance runs on the same server than the ManifoldCF instance.
- Index name: The connector will populate the index defined here.
- User name and API Key: The credentials required to connect to the OpenSearchServer instance. It can be left empty if no user has been created. The next figure shows where to find the user's informations in the OpenSearchServer user interface.



User name:	admin	Choose a name for the user that will be created
Password:	•••••	Choose a password for the user that will be created
Confirm password:	•••••	Retype the password
Admin user:	☑	Check the box if the user is admin
Can monitor:	☐	Check the box if the user is able to monitor
API Key:	54a51ee4f27cbbcb7a771352b980567f	Key to be used with API
Add privileges on following index:	FileCrawlerFTP Index: query the index	Add

Once you created a new job, having selected the OpenSearchServer output connector, you will have the OpenSearchServer tab. This tab let you:

- Fix the maximum size of a document before deciding to index it. The value is in bytes. The default value is 16MB.
- The allowed mime types. Warning it does not work with all repository connectors.
- The allowed file extensions. Warning it does not work with all repository connectors.



Max file size (bytes): 16777216

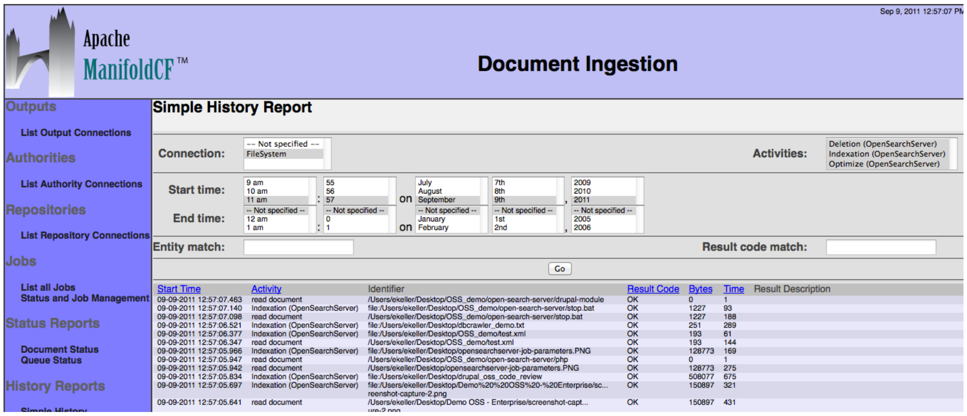
Allowed MIME Types :

```
application/msword
application/vnd.ms-excel
application/vnd.openxmlformats-
officedocument.wordprocessingml.document
application/vnd.openxmlformats-
officedocument.spreadsheetml.sheet
text/html
application/pdf
application/vnd.ms-powerpoint
application/vnd.openxmlformats-
officedocument.presentationml.presentation
```

Allowed file extensions:

```
doc
docx
xls
xlsx
ppt
pptx
html
pdf
odt
ods
rtf
```

In the history report you will be able to monitor all the activites. The connector supports three activites: Document ingestion (Indexation), document deletion and index optimization. The targeted index is automatically optimized when the job is ending.



Simple History Report

Connection: -- Not specified -- / FileSystem

Activities: Deletion (OpenSearchServer) / Indexation (OpenSearchServer) / Optimize (OpenSearchServer)

Start time: 11 am : 57 on September 9th , 2011

End time: -- Not specified -- : -- Not specified -- on -- Not specified -- -- Not specified -- , -- Not specified --

Entity match:

Result code match:

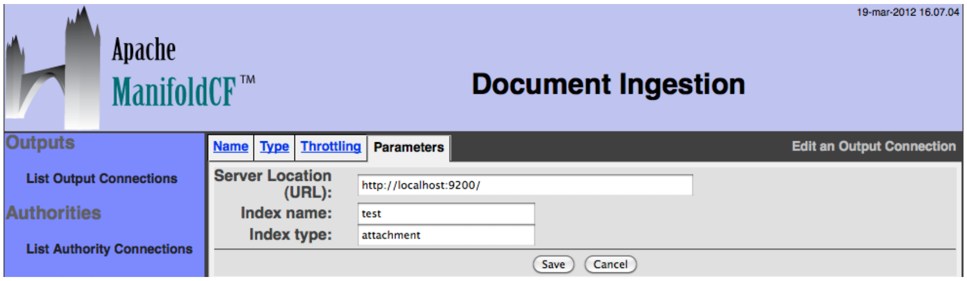
Start Time	Activity	Identifier	Result Code	Bytes	Time	Result Description
09-09-2011 12:57:07.463	read document	/Users/ekeller/Desktop/OSS_demo/open-search-server/drupal-module	OK	0	1	
09-09-2011 12:57:07.140	Indexation (OpenSearchServer)	file:/Users/ekeller/Desktop/OSS_demo/open-search-server/stop.bat	OK	1227	93	
09-09-2011 12:57:07.098	read document	/Users/ekeller/Desktop/OSS_demo/open-search-server/stop.bat	OK	1227	188	
09-09-2011 12:57:06.521	Indexation (OpenSearchServer)	file:/Users/ekeller/Desktop/dbcrawler_demo.txt	OK	251	289	
09-09-2011 12:57:06.377	Indexation (OpenSearchServer)	file:/Users/ekeller/Desktop/OSS_demo/test.xml	OK	193	61	
09-09-2011 12:57:06.347	read document	/Users/ekeller/Desktop/OSS_demo/test.xml	OK	193	144	
09-09-2011 12:57:05.966	Indexation (OpenSearchServer)	file:/Users/ekeller/Desktop/opensearchserver-job-parameters.PNG	OK	128773	169	
09-09-2011 12:57:05.947	read document	/Users/ekeller/Desktop/OSS_demo/open-search-server/php	OK	0	1	
09-09-2011 12:57:05.942	read document	/Users/ekeller/Desktop/opensearchserver-job-parameters.PNG	OK	128773	275	
09-09-2011 12:57:05.834	Indexation (OpenSearchServer)	file:/Users/ekeller/Desktop/drupal_oss_code_review	OK	508077	675	
09-09-2011 12:57:05.697	Indexation (OpenSearchServer)	file:/Users/ekeller/Desktop/Demo%20%20OSS%20-%20Enterprise/sc...reenshot-capture-2.png	OK	150897	321	
09-09-2011 12:57:05.641	read document	/Users/ekeller/Desktop/Demo OSS - Enterprise/screenshot-capt...ure-2.png	OK	150897	431	

You may also refer to the OpenSearchServer's user documentation.

2.3 ElasticSearch Output Connection

The ElasticSearch Output Connection allow ManifoldCF to submit documents to an ElasticSearch instance, via the XML over HTTP API. The connector has been designed to be as easy to use as possible.

After creating an ElasticSearch ouput connection, you have to populate the parameters tab. Fill in the fields according your ElasticSearch configuration. Each ElasticSearch output connector instance works with one index. To work with multiple indexes, just create one output connector for each index.



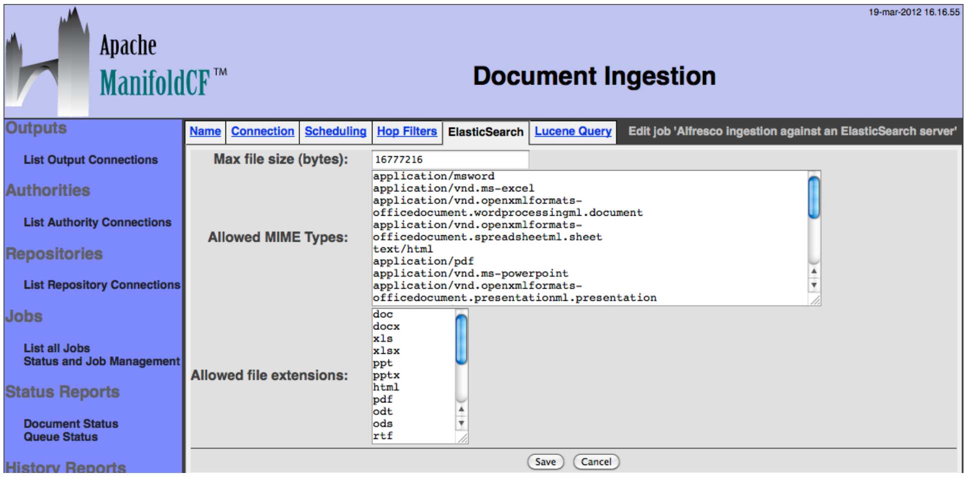
Server Location (URL): http://localhost:9200/
Index name: test
Index type: attachment

The parameters are:

- Server location: An URL that references your ElasticSearch instance. The default value (http://localhost:9200) is valid if your ElasticSearch instance runs on the same server than the ManifoldCF instance.
- Index name: The connector will populate the index defined here.

Once you created a new job, having selected the ElasticSearch output connector, you will have the ElasticSearch tab. This tab let you:

- Fix the maximum size of a document before deciding to index it. The value is in bytes. The default value is 16MB.
- The allowed mime types. Warning it does not work with all repository connectors.
- The allowed file extensions. Warning it does not work with all repository connectors.



Max file size (bytes): 16777216

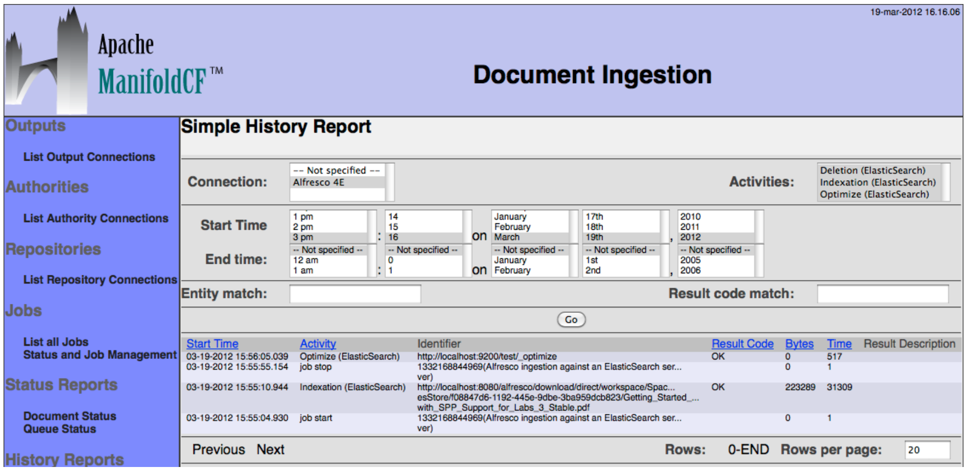
Allowed MIME Types:

```
application/msword
application/vnd.ms-excel
application/vnd.openxmlformats-
officedocument.wordprocessingml.document
application/vnd.openxmlformats-
officedocument.spreadsheetml.sheet
text/html
application/pdf
application/vnd.ms-powerpoint
application/vnd.openxmlformats-
officedocument.presentationml.presentation
```

Allowed file extensions:

```
doc
docx
xls
xlsx
ppt
pptx
html
pdf
odt
ods
rtf
```

In the history report you will be able to monitor all the activites. The connector supports three activites: Document ingestion (Indexation), document deletion and index optimization. The targeted index is automatically optimized when the job is ending.



Simple History Report

Connection: -- Not specified --
Alfresco 4E

Activities: Deletion (ElasticSearch)
Indexation (ElasticSearch)
Optimize (ElasticSearch)

Start Time 1 pm / 2 pm / 3 pm : 14 / 15 / 16 on January / February / March 17th / 18th / 19th , 2010 / 2011 / 2012

End time: -- Not specified -- / 12 am / 1 am : -- Not specified -- / 0 / 1 on -- Not specified -- / January / February -- Not specified -- / 1st / 2nd , -- Not specified -- / 2005 / 2006

Entity match:

Result code match:

Start Time	Activity	Identifier	Result Code	Bytes	Time	Result Description
03-19-2012 15:56:05.039	Optimize (ElasticSearch)	http://localhost:9200/test/_optimize	OK	0	517	
03-19-2012 15:55:55.154	job stop	1332168844969(Alfresco ingestion against an ElasticSearch ser...ver)		0	1	
03-19-2012 15:55:10.944	Indexation (ElasticSearch)	http://localhost:8080/alfresco/download/direct/workspace/Spac...esStore/f08847d6-1192-445e-9dbe-3ba959dcb823/Getting_Started_...with_SPP_Support_for_Labs_3_Stable.pdf	OK	223289	31309	
03-19-2012 15:55:04.930	job start	1332168844969(Alfresco ingestion against an ElasticSearch ser...ver)		0	1	

Rows: 0-END Rows per page: 20

You may also refer to ElasticSearch's user documentation.

2.4 MetaCarta GTS Output Connection

The MetaCarta GTS output connection type is designed to allow ManifoldCF to submit documents to an appropriate MetaCarta GTS search appliance, via the appliance's HTTP Ingestion API.

The connection type implicitly understands that GTS can only handle text, HTML, XML, RTF, PDF, and Microsoft Office documents. All other document types will be considered to be unindexable. This helps prevent jobs based on a GTS-type output connection from fetching data that is large, but of no particular relevance.

When you configure a job to use a GTS-type output connection, two additional tabs will be presented to the user: "Collections" and "Document Templates". These tabs allow per-job specification of these GTS-specific features.

More here later

2.5 Null Output Connection

The null output connection type is meant primarily to function as an aid for people writing repository connection types. It is not expected to be useful in practice.

The null output connection type simply logs indexing and deletion requests, and does nothing else. It does not have any special configuration tabs, nor does it contribute tabs to jobs defined that use it.

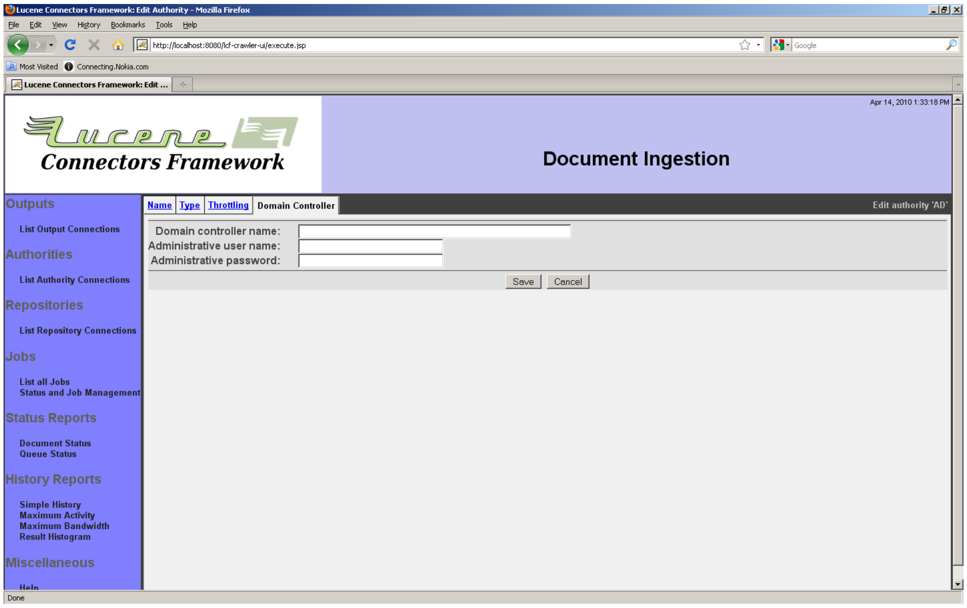
3 Authority Connection Types

3.1 Active Directory Authority Connection

An active directory authority connection is essential for enforcing security for documents from Windows shares, Microsoft SharePoint, and IBM FileNet repositories. This connection type needs to be provided with information about how to log into an appropriate Windows domain controller, with a user that has sufficient privileges to be able to look up any user's ID and group relationships. While the connection type has some known limitations, it should function well for most straightforward Windows security architecture situations. The cases in which it may not be adequate include:

- when child domains are present
- when the expected number of requests per second is fairly high

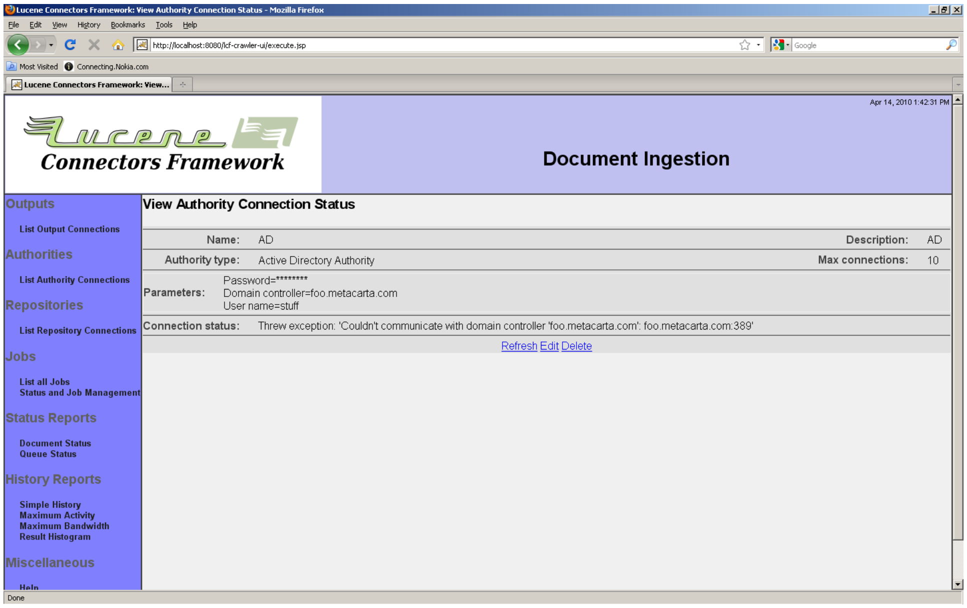
An active directory authority connection type has a single special tab in the authority connection editing screen: the "Domain Controller" tab:



Domain controller name:
Administrative user name:
Administrative password:

Fill in the requested values. Note that the "Administrative user name" field usually requires no domain suffix, but depending on the details of how the domain controller is configured, may sometimes only accept the "name@domain" format.

When you are done, click the "Save" button. When you do, a connection summary and status screen will be presented, which may look something like this:



View Authority Connection Status

Name: AD
Description: AD
Authority type: Active Directory Authority
Max connections: 10
Parameters: Password=********
Domain controller=foo.metacarta.com
User name=stuff
Connection status: Threw exception: 'Couldn't communicate with domain controller 'foo.metacarta.com': foo.metacarta.com:389'

Refresh Edit Delete

Note that in this example, the Active Directory connection is not responding, which is leading to an error status message instead of "Connection working".

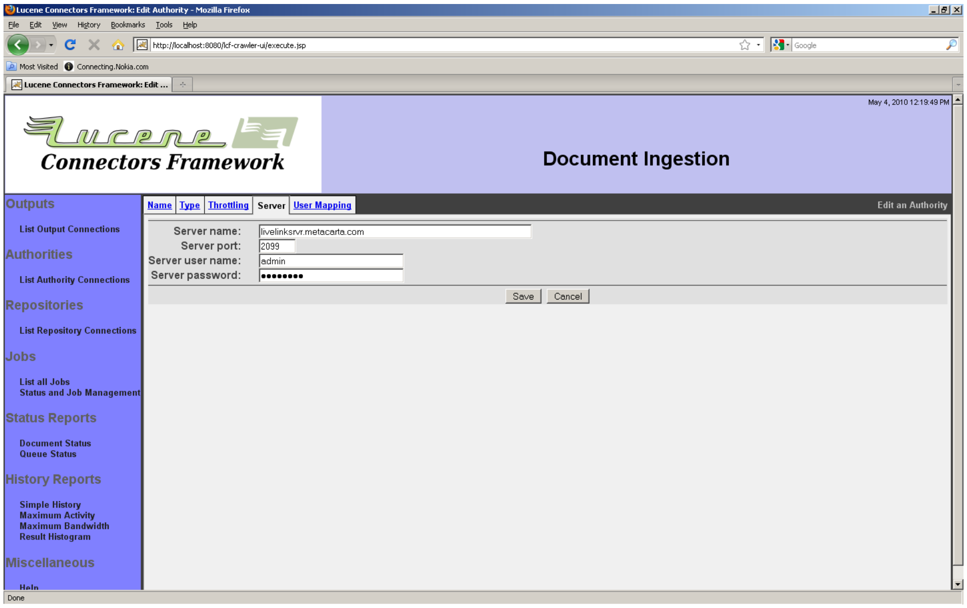
3.2 OpenText LiveLink Authority Connection

A LiveLink authority connection is needed to enforce security for documents retrieved from LiveLink repositories.

In order to function, this connection type needs to be provided with information about the name of the LiveLink server, and credentials appropriate for retrieving a user's ACLs from that machine. Since LiveLink operates with its own list of users, you may also want to specify a rule-based mapping between an Active Directory user and the corresponding LiveLink user. The authority type allows you to specify such a mapping using regular expressions.

A LiveLink authority connection has two special tabs you will need to configure: the "Server" tab, and the "User Mapping" tab.

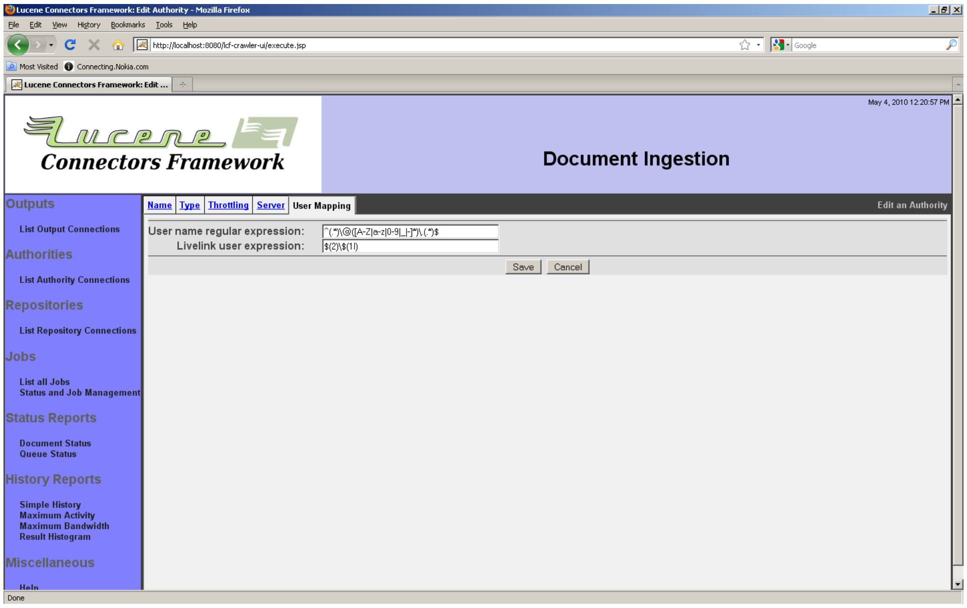
The "Server" tab looks like this:



Server name: livelinksrvr.metacarta.com
Server port: 2099
Server user name: admin
Server password: •••••••••

Enter the name of the desired LiveLink server, the LiveLink port, and the LiveLink credentials.

The "User Mapping" tab looks like this:

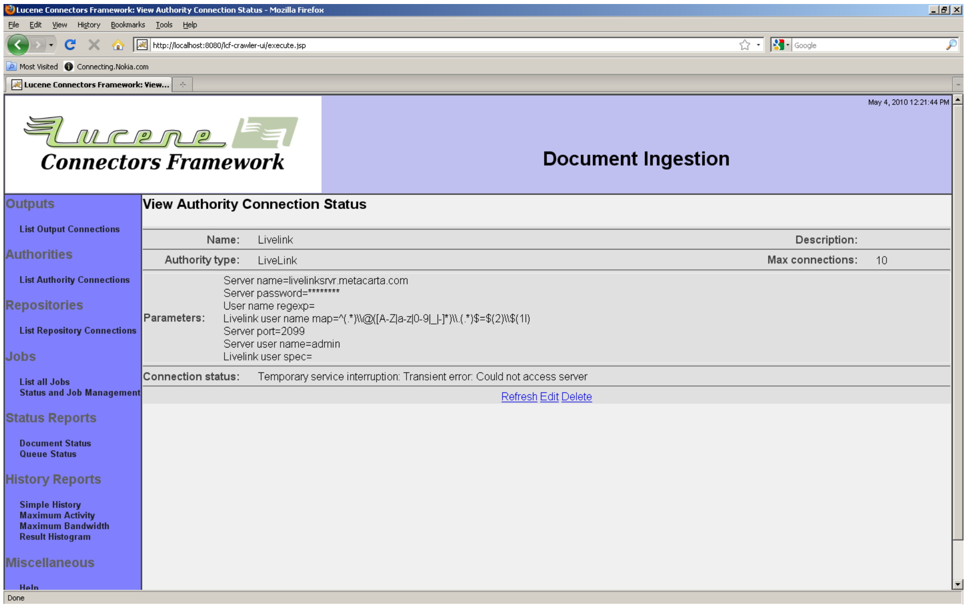


User name regular expression: ^(.*)\@([A-Z|a-z|0-9|_|-]*)\.(.*)$
Livelink user expression: $(2)\$(1l)

The purpose of the "User Mapping" tab is to allow you to map the incoming user name and domain (usually from Active Directory) to its LiveLink equivalent. The mapping consists of a match expression, which is a regular expression where parentheses ("(" and ")") mark sections you are interested in, and a replace string. The sections marked with parentheses are called "groups" in regular expression parlance. The replace string consists of constant text plus substitutions of the groups from the match, perhaps modified. For example, "$(1)" refers to the first group within the match, while "$(1l)" refers to the first match group mapped to lower case. Similarly, "$(1u)" refers to the same characters, but mapped to upper case.

For example, a match expression of `^(.*)\@([A-Z|a-z|0-9|_|-]*)\.(.*)$` with a replace string of `$(2)\$(1l)` would convert an AD username of `MyUserName@subdomain.domain.com` into the LiveLink user name `subdomain\myusername`.

When you are done, click the "Save" button. You will then see a summary and status for the authority connection:



View Authority Connection Status

Name: Livelink
Description:
Authority type: LiveLink
Max connections: 10

Parameters:
Server name=livelinksrvr.metacarta.com
Server password=********
User name regexp=
Livelink user name map=^(.*)\\@([A-Z|a-z|0-9|_|-]*)\\.(.*)$=$(2)\\$(1l)
Server port=2099
Server user name=admin
Livelink user spec=

Connection status: Temporary service interruption: Transient error: Could not access server

We suggest that you examine the status carefully and correct any reported errors before proceeding. Note that in this example, the LiveLink server would not accept connections, which is leading to an error status message instead of "Connection working".

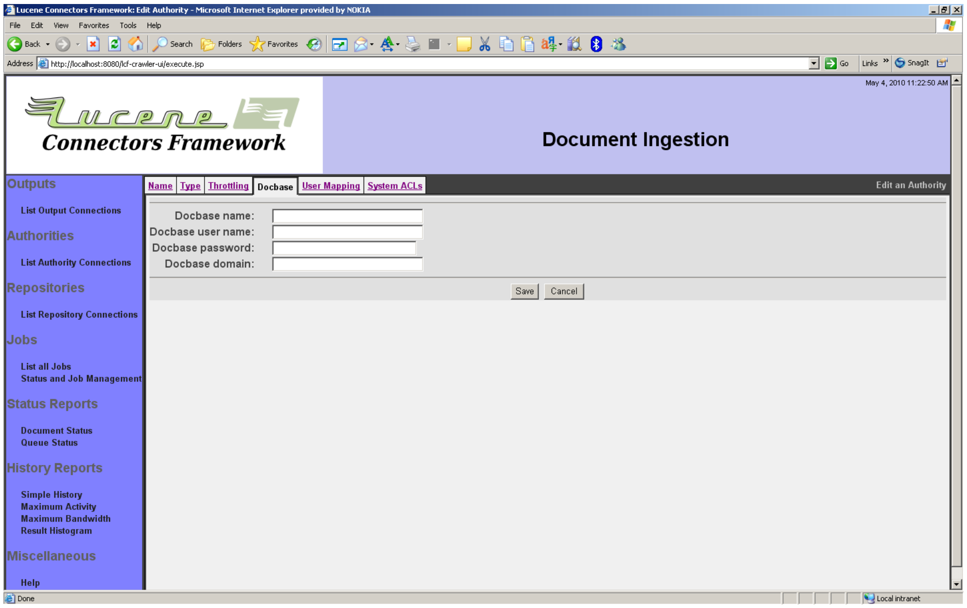
3.3 EMC Documentum Authority Connection

A Documentum authority connection is required for enforcing security for documents retrieved from Documentum repositories.

This connection type needs to be provided with information about what Content Server to connect to, and the credentials that should be used to retrieve a user's ACLs from that machine. In addition, you can also specify whether or not you wish to include auto-generated ACLs in every user's list. Auto-generated ACLs are created within Documentum for every folder object. Because there are often a very large number of folders, including these ACLs can bloat the number of ManifoldCF access tokens returned for a user to tens of thousands, which can negatively impact perfomance. Even more notably, few Documentum installations make any real use of these ACLs in any way. Since Documentum's ACLs are purely additive (that is, there are no mechanisms for 'deny' semantics), the impact of a missing ACLs is only to block a user from seeing something they otherwise could see. It is thus safe, and often desirable, to simply ignore the existence of these auto-generated ACLs.

A Documentum authority connection has three special tabs you will need to configure: the "Docbase" tab, the "User Mapping" tab, and the "System ACLs" tab.

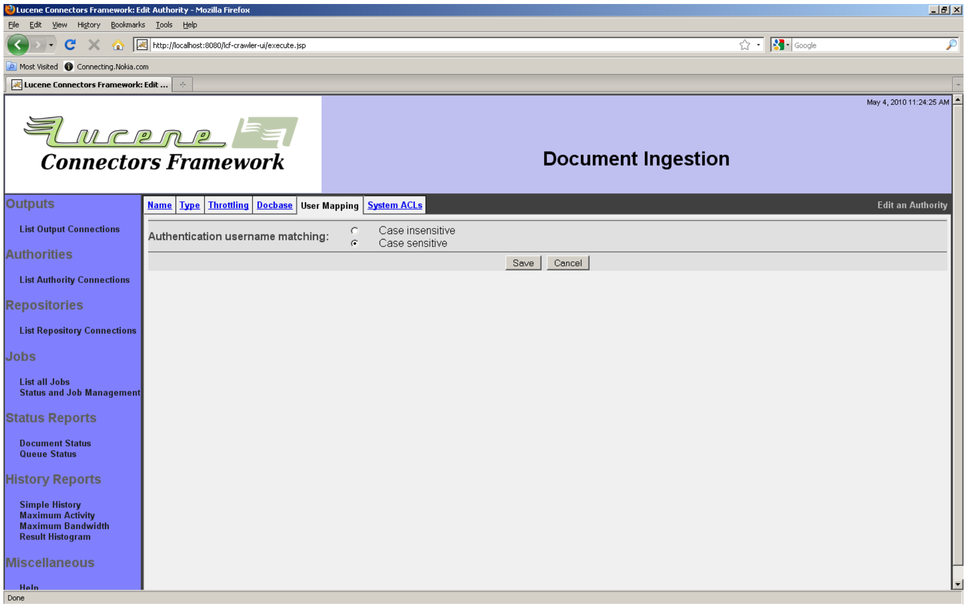
The "Docbase" tab looks like this:



Docbase name:
Docbase user name:
Docbase password:
Docbase domain:

Enter the desired Content Server docbase name, and enter the appropriate credentials. You may leave the "Domain" field blank if the Content Server you specify does not have Active Directory support enabled.

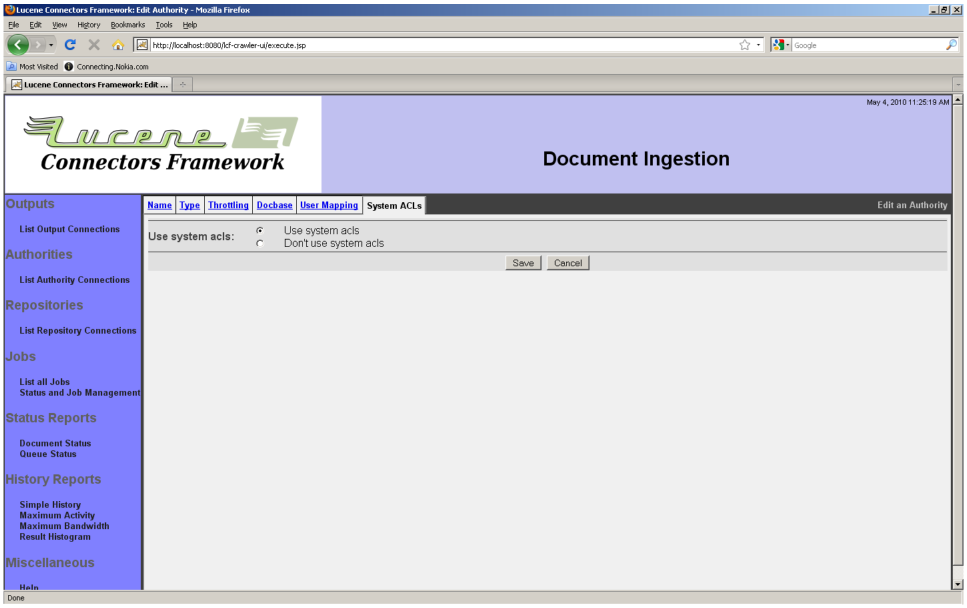
The "User Mapping" tab looks like this:



Authentication username matching: ○ Case insensitive ● Case sensitive

Here you can specify whether the mapping between incoming user names and Content Server user names is case sensitive or case insensitive. No other mappings are currently permitted. Typically, Documentum instances operate in conjunction with Active Directory, such that Documentum user names are either the same as the Active Directory user names, or are the Active Directory user names mapped to all lower case characters. You may need to consult with your Documentum system administrator to decide what the correct setting should be for this option.

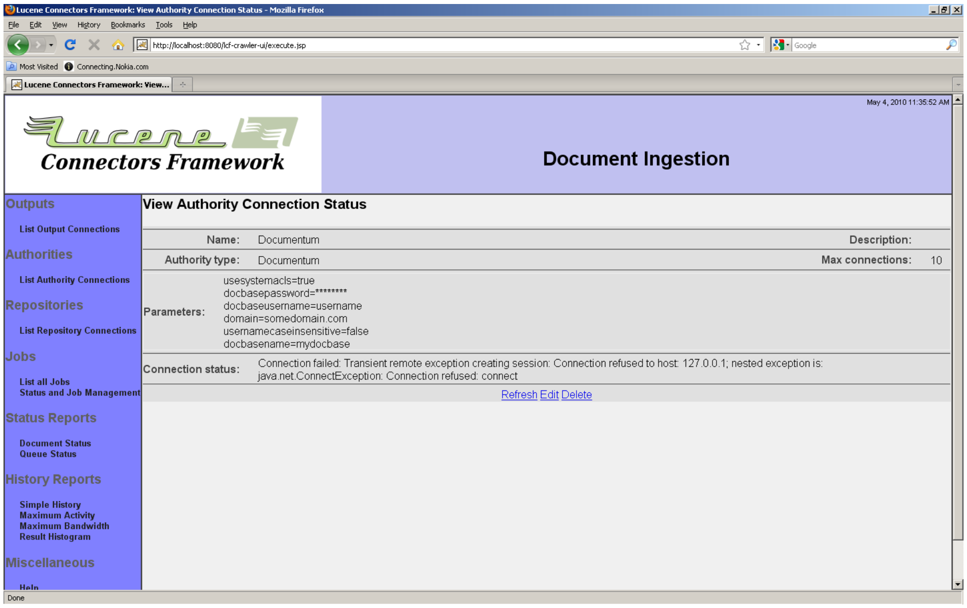
The "System ACLs" tab looks like this:


Lucene Connectors Framework: Edit Authority - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited Connecting.Nokia.com
Lucene Connectors Framework: Edit ...
May 4, 2010 11:25:19 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Throttling
Docbase
User Mapping
System ACLs
Edit an Authority
Save
Cancel
Done

Use system acls: ● Use system acls ○ Don't use system acls

Here, you can choose to ignore all auto-generated ACLs associated with a user. We recommend that you try ignoring such ACLs, and only choose the default if you have reason to believe that your Documentum content is protected in a significant way by the use of auto-generated ACLs. Your may need to consult with your Documentum system administrator to decide what the proper setting should be for this option.

When you are done, click the "Save" button. When you do, a connection summary and status screen will be presented:


Lucene Connectors Framework: View Authority Connection Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Google
Most Visited
Connecting.Nokia.com
Lucene Connectors Framework: View...
Lucene Connectors Framework
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

May 4, 2010 11:35:52 AM

Document Ingestion

View Authority Connection Status

Name:	Documentum	Description:	
Authority type:	Documentum	Max connections:	10

Parameters:	usesystemacls=true
	docbasepassword=********
	docbaseusername=username
	domain=somedomain.com
	usernamecaseinsensitive=false
	docbasename=mydocbase		

Connection status:	Connection failed: Transient remote exception creating session: Connection refused to host: 127.0.0.1; nested exception is: java.net.ConnectException: Connection refused: connect		

Refresh Edit Delete

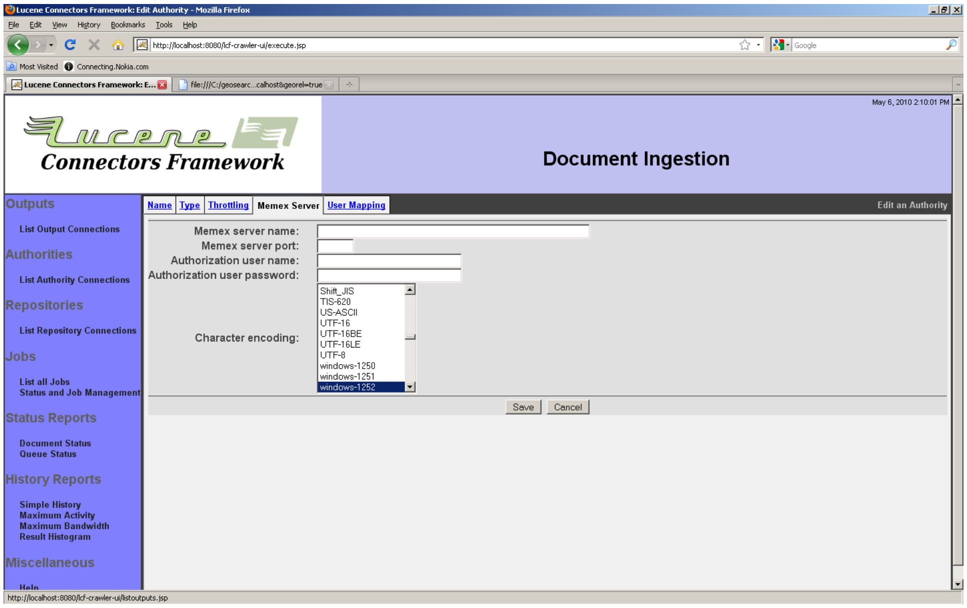
Pay careful attention to the status, and be prepared to correct any problems that are displayed.

3.4 Memex Patriarch Authority Connection

A Memex authority connection is required for enforcing security for documents retrieved from Memex repositories.

This connection type needs to be provided with information about what Memex Server to connect to, and what user mapping to perform. Also needed are the Memex credentials that should be used to retrieve a user's permissions from the Memex server.

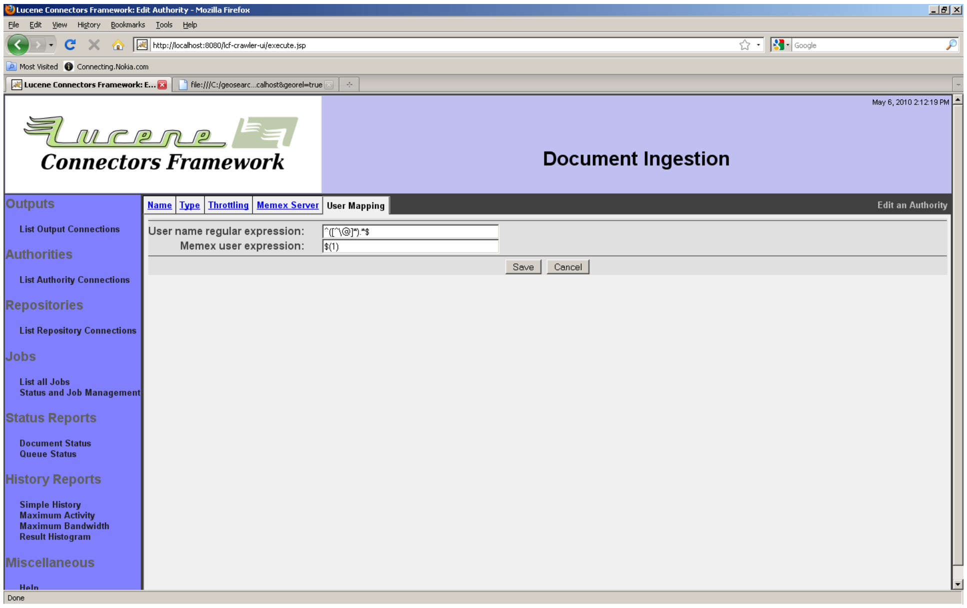
A Memex authority connection has the following special tabs you will need to configure: the "Memex Server" tab, and the "User Mapping" tab. The "Memex Server" tab looks like this:



Memex server name:
Memex server port:
Authorization user name:
Authorization user password:
Character encoding: Shift_JIS, TIS-620, US-ASCII, UTF-16, UTF-16BE, UTF-16LE, UTF-8, windows-1250, windows-1251, windows-1252 (selected)

You must supply the name of your Memex server, and the connection port, along with the Memex credentials for a user that has sufficient permissions to retrieve Memex user information. You must also select the Memex server's character encoding. If you do not know the encoding, consult your Memex system administrator.

The "User Mapping" tab looks like this:

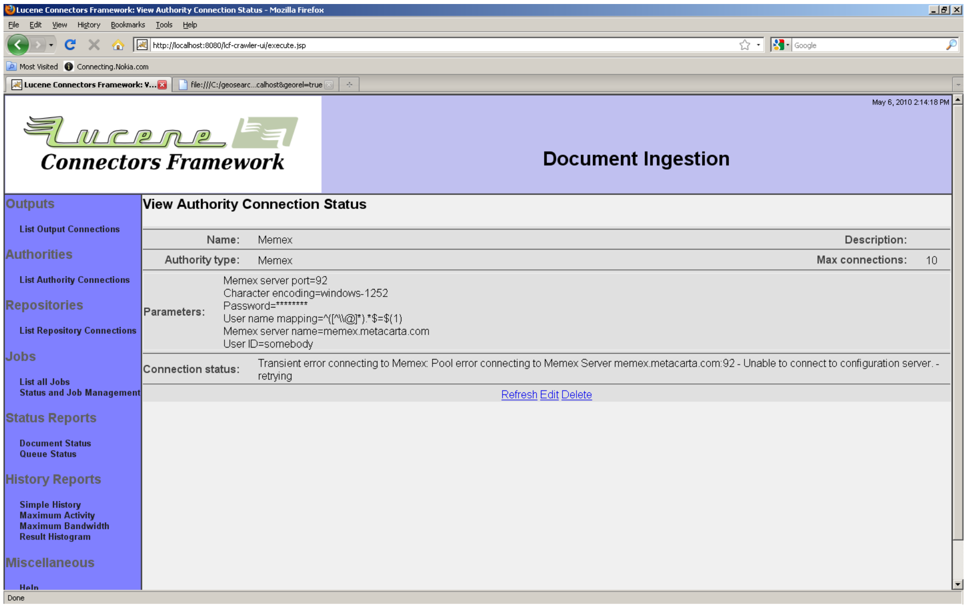


User name regular expression: ^([^\@]*).*$
Memex user expression: $(1)

The purpose of the "User Mapping" tab is to allow you to map the incoming user name and domain (usually from Active Directory) to its Memex equivalent. The mapping consists of a match expression, which is a regular expression where parentheses ("(" and ")") mark sections you are interested in, and a replace string. The sections marked with parentheses are called "groups" in regular expression parlance. The replace string consists of constant text plus substitutions of the groups from the match, perhaps modified. For example, "$(1)" refers to the first group within the match, while "$(1l)" refers to the first match group mapped to lower case. Similarly, "$(1u)" refers to the same characters, but mapped to upper case.

For example, a match expression of `^(.*)\@([A-Z|a-z|0-9|_|-]*)\.(.*)$` with a replace string of `$(2)\$(1l)` would convert an AD username of `MyUserName@subdomain.domain.com` into the Memex user name `subdomain\myusername`.

When you are done, click the "Save" button. You will then see a summary and status for the authority connection:



View Authority Connection Status

Name: Memex
Description:
Authority type: Memex
Max connections: 10

Parameters:
Memex server port=92
Character encoding=windows-1252
Password=********
User name mapping=^([^\\@]*).*$=$(1)
Memex server name=memex.metacarta.com
User ID=somebody

Connection status: Transient error connecting to Memex: Pool error connecting to Memex Server memex.metacarta.com:92 - Unable to connect to configuration server. - retrying

We suggest that you examine the status carefully and correct any reported errors before proceeding. Note that in this example, the Memex server has a license error, which is leading to an error status message instead of "Connection working".

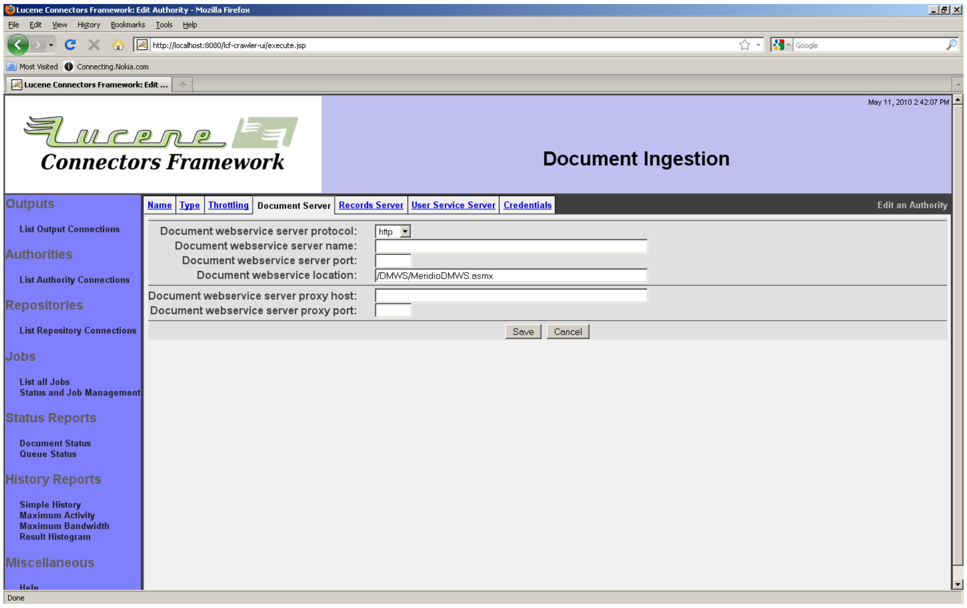
3.5 Autonomy Meridio Authority Connection

A Meridio authority connection is required for enforcing security for documents retrieved from Meridio repositories.

This connection type needs to be provided with information about what Document Server to connect to, what Records Server to connect to, and what User Service Server to connect to. Also needed are the Meridio credentials that should be used to retrieve a user's ACLs from those machines.

Note that the User Service is part of the Meridio Authority, and must be installed somewhere in the Meridio system in order for the Meridio Authority to function correctly. If you do not know whether this has yet been done, or on what server, please ask your system administrator.

A Meridio authority connection has the following special tabs you will need to configure: the "Document Server" tab, the "Records Server" tab, the "User Service Server" tab, and the "Credentials" tab. The "Document Server" tab looks like this:

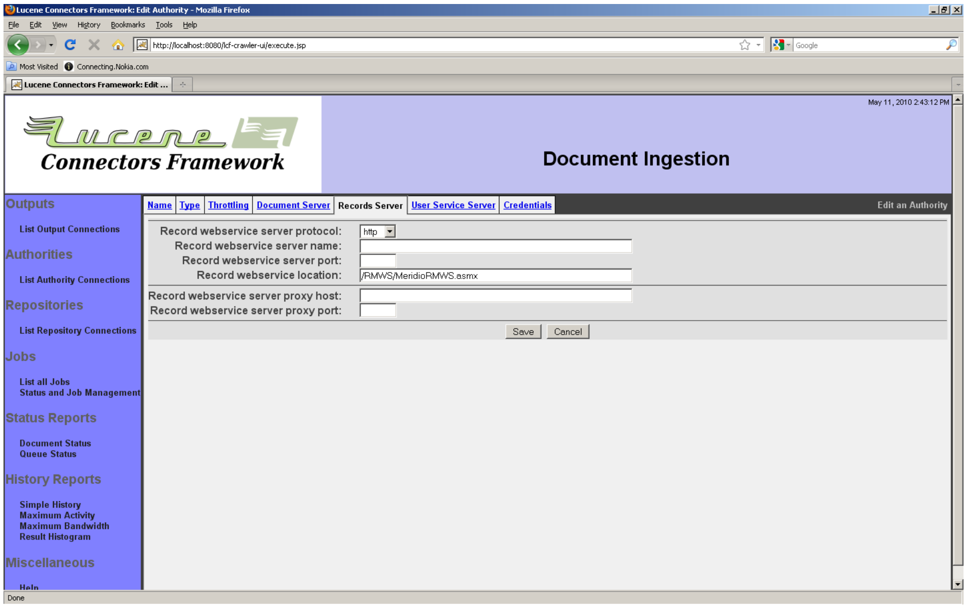


Document webservice server protocol: http
Document webservice server name:
Document webservice server port:
Document webservice location: /DMWS/MeridioDMWS.asmx
Document webservice server proxy host:
Document webservice server proxy port:

Select the correct protocol, and enter the correct server name, port, and location to reference the Meridio document server services. If a proxy is involved, enter the proxy host and port. Authenticated proxies are not supported by this connection type at this time.

Note that, in the Meridio system, while it is possible that different services run on different servers, this is not typically the case. The connection type, on the other hand, makes no assumptions, and permits the most general configuration.

The "Records Server" tab looks like this:



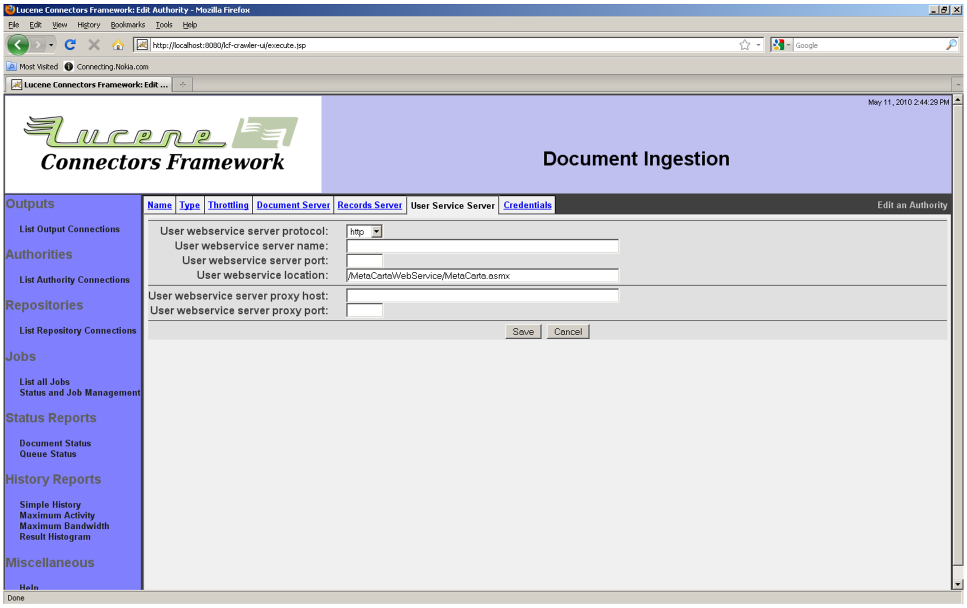
Record webservice server protocol: http
Record webservice server name:
Record webservice server port:
Record webservice location: /RMWS/MeridioRMWS.asmx

Record webservice server proxy host:
Record webservice server proxy port:

Select the correct protocol, and enter the correct server name, port, and location to reference the Meridio records server services. If a proxy is involved, enter the proxy host and port. Authenticated proxies are not supported by this connection type at this time.

Note that, in the Meridio system, while it is possible that different services run on different servers, this is not typically the case. The connection type, on the other hand, makes no assumptions, and permits the most general configuration.

The "User Service Server" tab looks like this:



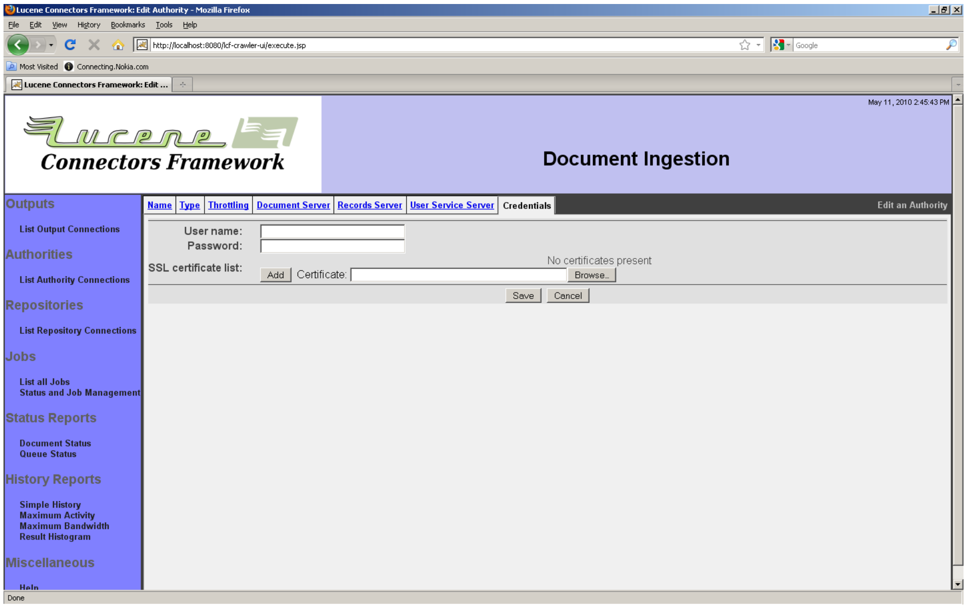
User webservice server protocol:	http
User webservice server name:	
User webservice server port:	
User webservice location:	/MetaCartaWebService/MetaCarta.asmx
User webservice server proxy host:	
User webservice server proxy port:	

You will require knowledge of where the special Meridio Authority extensions have been installed in order to fill out this tab.

Select the correct protocol, and enter the correct server name, port, and location to reference the Meridio user service server services. If a proxy is involved, enter the proxy host and port. Authenticated proxies are not supported by this connection type at this time.

Note that, in the Meridio system, while it is possible that different services run on different servers, this is not typically the case. The connection type, on the other hand, makes no assumptions, and permits the most general configuration.

The "Credentials" tab looks like this:



User name:

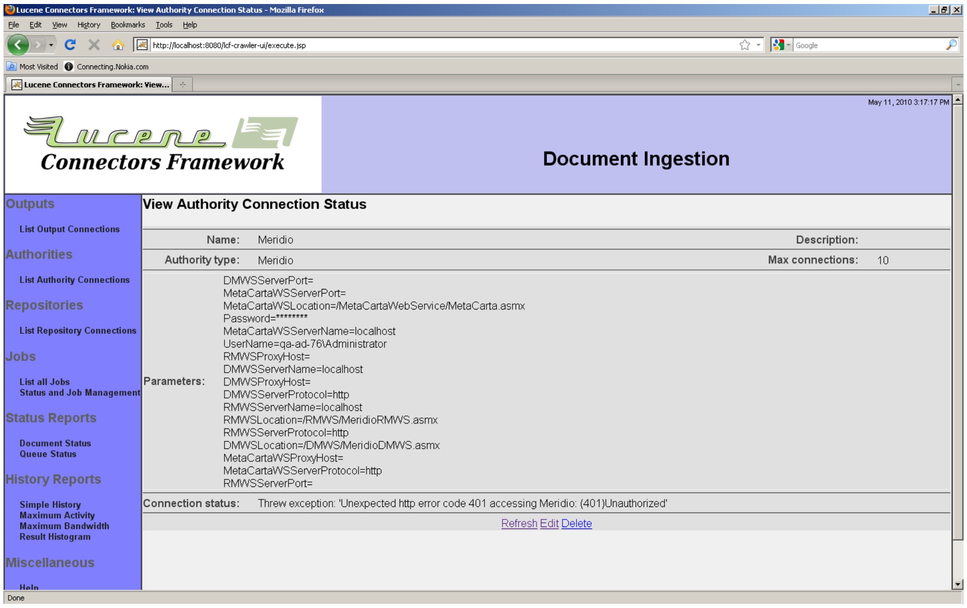
Password:

SSL certificate list: No certificates present

Certificate:

Enter the Meridio server credentials needed to access the Meridio system.

When you are done, click the "Save" button. You will then see a screen looking something like this:


Lucene Connectors Framework: View Authority Connection Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited
Connecting.Nokia.com
Lucene Connectors Framework: View...
Lucene Connectors Framework
May 11, 2010 3:17:17 PM
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Refresh
Edit
Delete
Done

View Authority Connection Status

Name: Meridio Description:

Authority type: Meridio Max connections: 10

Parameters:

```
DMWSServerPort=
MetaCartaWSServerPort=
MetaCartaWSLocation=/MetaCartaWebService/MetaCarta.asmx
Password=********
MetaCartaWSServerName=localhost
UserName=qa-ad-76\Administrator
RMWSProxyHost=
DMWSServerName=localhost
DMWSProxyHost=
DMWSServerProtocol=http
RMWSServerName=localhost
RMWSLocation=/RMWS/MeridioRMWS.asmx
RMWSServerProtocol=http
DMWSLocation=/DMWS/MeridioDMWS.asmx
MetaCartaWSProxyHost=
MetaCartaWSServerProtocol=http
RMWSServerPort=
```

Connection status: Threw exception: 'Unexpected http error code 401 accessing Meridio: (401)Unauthorized'

In this example, logon has not succeeded because the server on which the Meridio Authority is running is unknown to the Windows domain under which Meridio is running. This results in an error message, instead of the "Connection working" message that you would see if the authority was working properly.

Since Meridio uses Windows IIS for authentication, there are many ways in which the configuration of either IIS or the Windows domain under which Meridio runs can affect the correct functioning of the Meridio Authority. It is beyond the scope of this manual to describe the kinds of analysis and debugging techniques that might be required to diagnose connection and authentication problems. If you have trouble, you will almost certainly need to involve your Meridio IT personnel. Debugging tools may include (but are not limited to):

- Windows security event logs
- ManifoldCF logs (see below)
- Packet captures (using a tool such as WireShark)

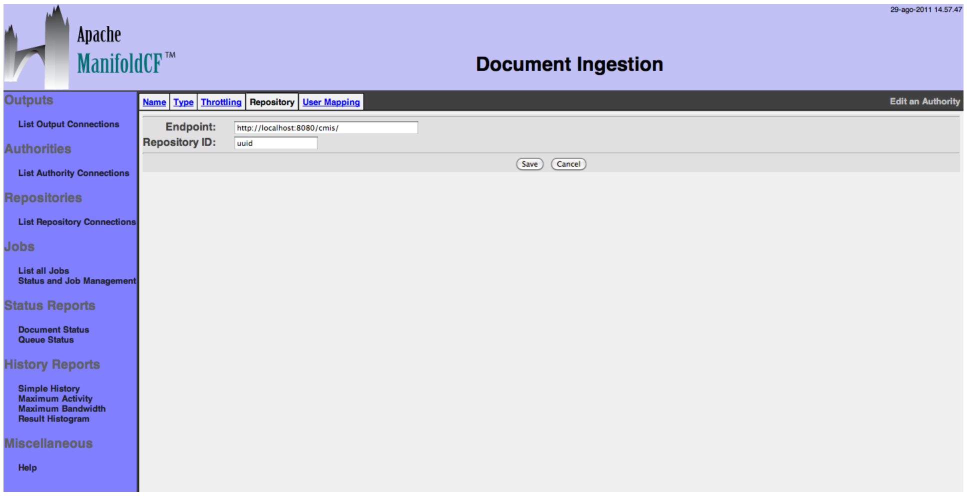
If you need specific ManifoldCF logging information, contact your system integrator.

3.6 CMIS Authority Connection

A CMIS authority connection is required for enforcing security for documents retrieved from CMIS repositories.

The CMIS specification includes the concept of authorities only depending on a specific document, this authority connector is only based on a regular expression comparator.

A CMIS authority connection has the following special tabs you will need to configure: the "Repository" tab and the "User Mapping" tab. The "Repository" tab looks like this:

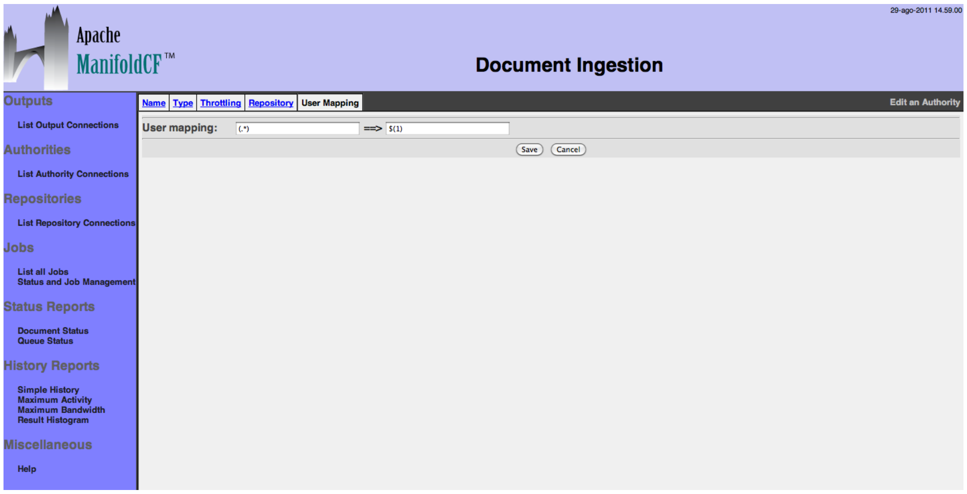


Endpoint: http://localhost:8080/cmis/
Repository ID: uuid

The repository configuration will be only used to track an ID for a specific CMIS repository. No calls will be performed against the CMIS repository.

The second tab that you need to configure is the "User Mapping" tab that allows you to define a regular expression to specify the user mapping.

The "User Mapping" tab looks like the following:

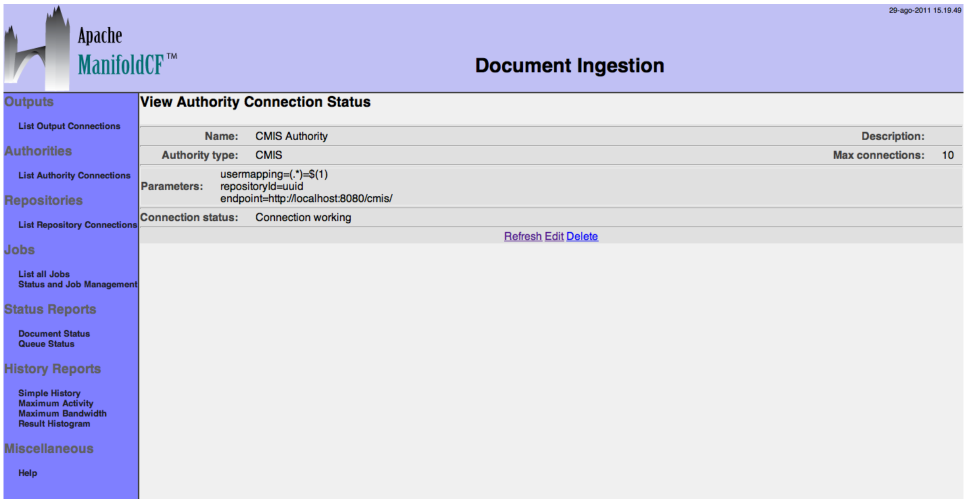


User mapping: (.*) ==> $(1)

The purpose of the "User Mapping" tab is to allow you to map the incoming user name and domain (usually from Active Directory) to its CMIS user equivalent. The mapping consists of a match expression, which is a regular expression where parentheses ("(" and ")") mark sections you are interested in, and a replace string. The sections marked with parentheses are called "groups" in regular expression parlance. The replace string consists of constant text plus substitutions of the groups from the match, perhaps modified. For example, "$(1)" refers to the first group within the match, while "$(1l)" refers to the first match group mapped to lower case. Similarly, "$(1u)" refers to the same characters, but mapped to upper case.

For example, a match expression of `^(.*)\@([A-Z|a-z|0-9|_|-]*)\.(.*)$` with a replace string of `$(2)\$(1l)` would convert an AD username of `MyUserName@subdomain.domain.com` into the LiveLink user name `subdomain\myusername`.

When you are done, click the "Save" button. You will then see a summary and status for the authority connection:



View Authority Connection Status

Name: CMIS Authority
Description:
Authority type: CMIS
Max connections: 10
Parameters:
usermapping=(.*)=$(1)
repositoryId=uuid
endpoint=http://localhost:8080/cmis/
Connection status: Connection working

4 Repository Connection Types

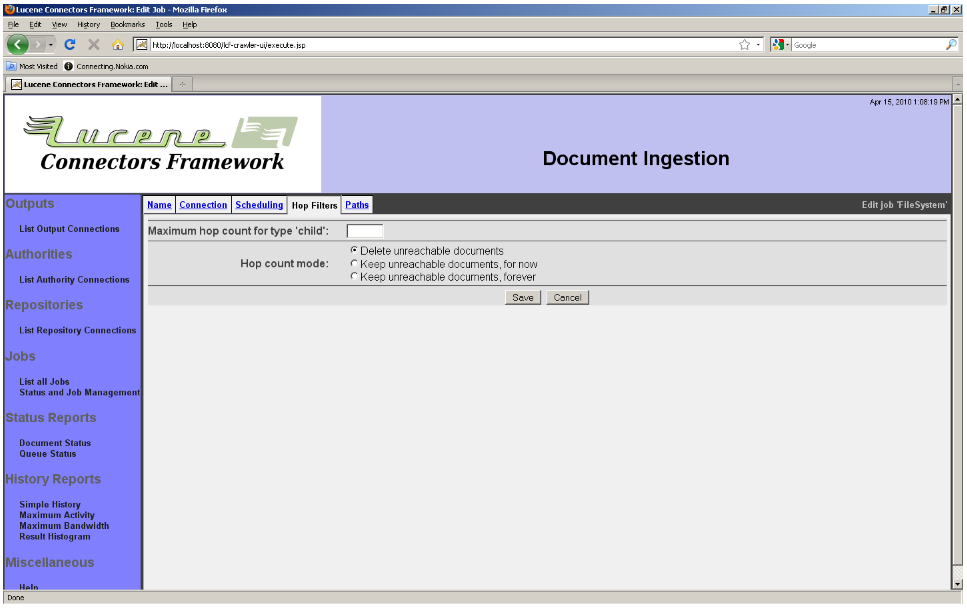
4.1 Generic File System Repository Connection

The generic file system repository connection type was developed primarily as an example, demonstration, and testing tool, although it can potentially be useful for indexing local files that exist on the same machine that ManifoldCF is running on. Bear in mind that there is no support in this connection type for any kind of security, and the options are somewhat limited.

The file system repository connection type provides no configuration tabs beyond the standard ones. However, please consider setting a "Maximum connections per JVM" value on the "Throttling" tab to at least one per worker thread, or 30, for best performance.

Jobs created using a file-system-type repository connection have two tabs in addition to the standard repertoire: the "Hop Filters" tab, and the "Paths" tab.

The "Hop Filters" tab allows you to restrict the document set by the number of child hops from the path root. While this is not terribly interesting in the case of a file system, the same basic functionality is also used in the Web connection type, where it is a more important feature. The file system connection type gives you a way to see how this feature works, in a more predictable environment:



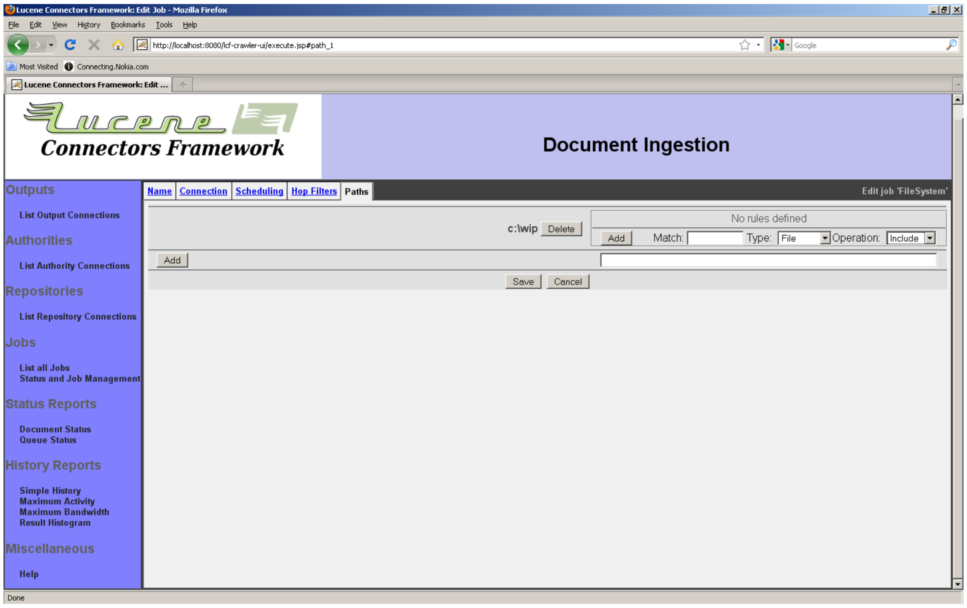
Maximum hop count for type 'child':

Hop count mode:
● Delete unreachable documents
○ Keep unreachable documents, for now
○ Keep unreachable documents, forever

In the case of the file system connection type, there is only one variety of relationship between documents, which is called a "child" relationship. If you want to restrict the document set by how far away a document is from the path root, enter the maximum allowed number of hops in the text box. Leaving the box blank indicates that no such filtering will take place.

On this same tab, you can tell the Framework what to do should there be changes in the distance from the root to a document. The choice "Delete unreachable documents" requires the Framework to recalculate the distance to every potentially affected document whenever a change takes place. This may require expensive bookkeeping, however, so you also have the option of ignoring such changes. There are two varieties of this latter option - you can ignore the changes for now, with the option of turning back on the aggressive bookkeeping at a later time, or you can decide not to ever allow changes to propagate, in which case the Framework will discard the necessary bookkeeping information permanently.

The "Paths" tab looks like this:



Lucene Connectors Framework: Edit Job - Mozilla Firefox

http://localhost:8080/lcf-crawler-ui/execute.jsp#path_1

Edit job 'FileSystem'

c:\wip

No rules defined

Match: Type: File Operation: Include

This tab allows you to type in a set of paths which function as the roots of the crawl. For each desired path, type in the path and click the "Add" button to add it to the list. The form of the path you type in obviously needs to be meaningful for the operating system the Framework is running on.

Each root path has a set of rules which determines whether a document is included or not in the set for the job. Once you have added the root path to the list, you may then add rules to it. Each rule has a match expression, an indication of whether the rule is intended to match files or directories, and an action (include or exclude). Rules are evaluated from top to bottom, and the first rule that matches the file name is the one that is chosen. To add a rule, select the desired pulldowns, type in a match file specification (e.g. "*.txt"), and click the "Add" button.

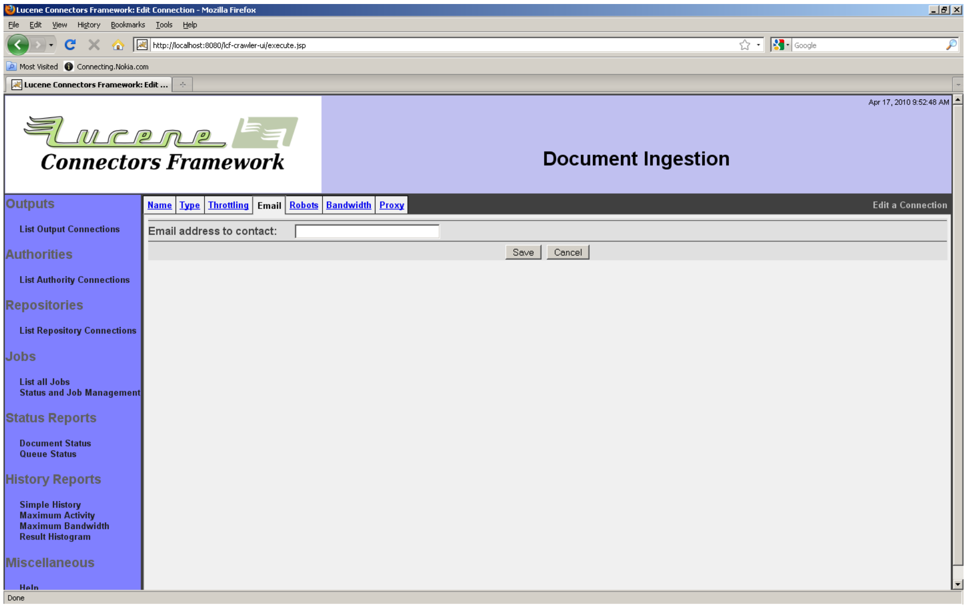
4.2 Generic RSS Repository Connection

The RSS connection type is specifically designed to crawl RSS feeds. While the Web connection type can also extract links from RSS feeds, the RSS connection type differs in the following ways:

- Links are only extracted from feeds
- Feeds themselves are not indexed
- There is fine-grained control over how often feeds are refetched, and they are treated distinctly from documents in this regard
- The RSS connection type knows how to carry certain data down from the feeds to individual documents, as metadata

Many users of the RSS connection type set up their jobs to run continuously, configuring their jobs to never refetch documents, but rather to expire them after some 30 days. This model works reasonably well for news, which is what RSS is often used for.

An RSS connection has the following special tabs: "Email", "Robots", "Bandwidth", and "Proxy". The "Email" tab looks like this:

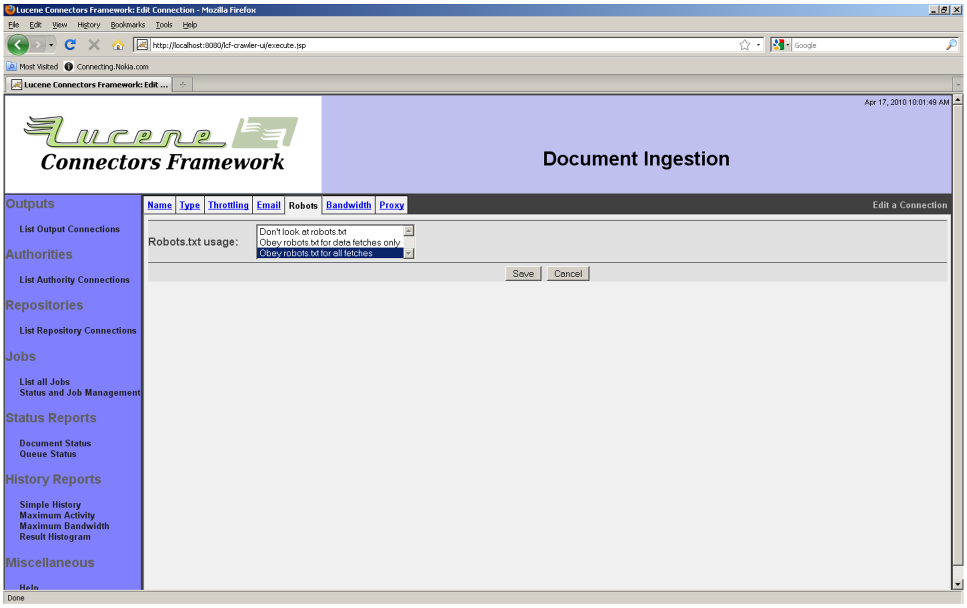

Lucene Connectors Framework: Edit Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited Connecting.Nokia.com
Lucene Connectors Framework: Edit ...
Apr 17, 2010 9:52:48 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name Type Throttling Email Robots Bandwidth Proxy
Edit a Connection
Save Cancel
Done

Email address to contact:

Enter an email address. This email address will be included in all requests made by the RSS connection, so that webmasters can report any difficulties that their sites experience as the result of improper throttling, etc.

This field is mandatory. While an RSS connection makes no effort to validate the correctness of the email field, you will probably want to remain a good web citizen and provide a valid email address. Remember that it is very easy for a webmaster to block access to a crawler that does not seem to be behaving in a polite manner.

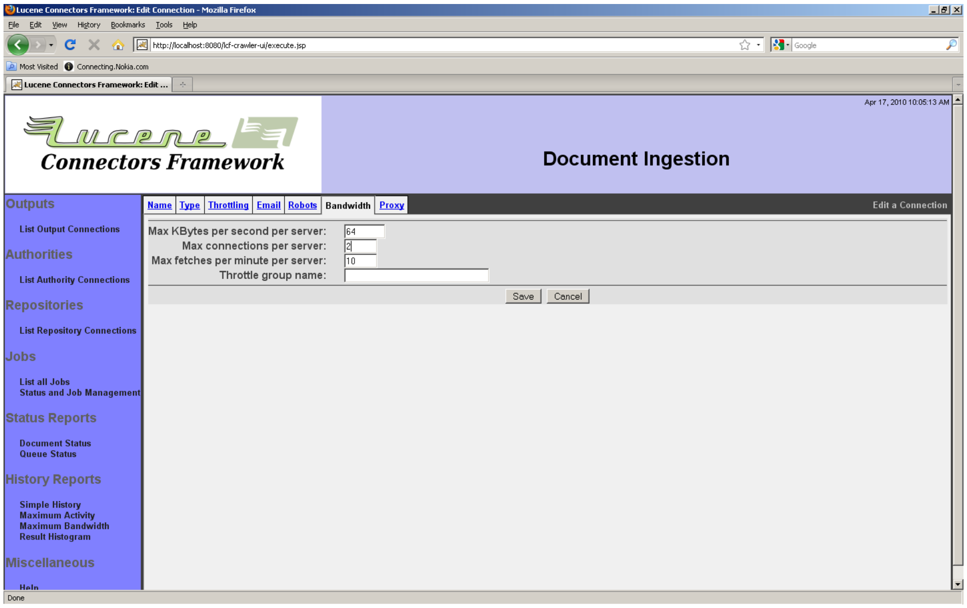
The "Robots" tab looks like this:



Robots.txt usage: Don't look at robots.txt / Obey robots.txt for data fetches only / [Obey robots.txt for all fetches]

Select how the connection will interpret robots.txt. Remember that you have an interest in crawling people's sites as politely as is possible.

The "Bandwidth" tab looks like this:



Max KBytes per second per server: 64
Max connections per server: 2
Max fetches per minute per server: 10
Throttle group name:

This tab allows you to control the maximum rate at which the connection fetches data, on a per-server basis, as well as the maximum fetches per minute, also per-server. Finally, the maximum number of socket connections made per server at any one time is also controllable by this tab.

The screen shot displays parameters that are considered reasonably polite. The default values for this table are all blank, meaning that, by default, there is no throttling whatsoever! Please do not make the mistake of crawling other people's sites without adequate politeness parameters in place.

The "Throttle group" parameter allows you to treat multiple RSS-type connections together, for the purposes of throttling. All RSS-type connections that have the same throttle group name will use the same pool for throttling purposes.

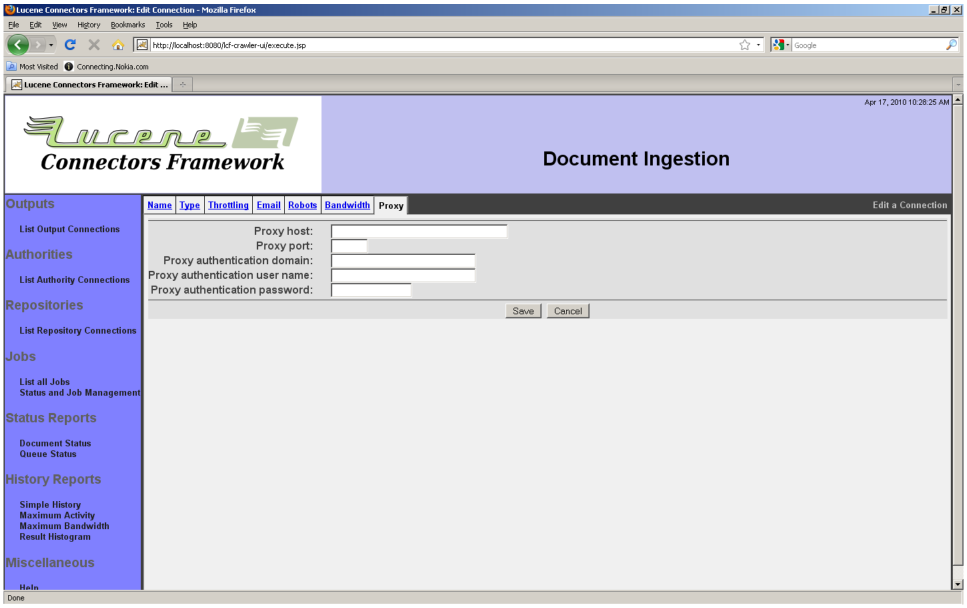
The "Bandwidth" tab is related to the throttles that you can set on the "Throttling" tab in the following ways:

- The "Bandwidth" tab sets the maximum values, while the "Throttling" tab sets the average values.
- The "Bandwidth" tab does not affect how documents are scheduled in the queue; it simply blocks documents until it is safe to go ahead, which will use up a crawler thread for the entire period that both the

wait and the fetch take place. The "Throttling" tab affects how often documents are scheduled, so it does not waste threads.

Because of the above, we suggest that you configure your RSS connection using both the "Bandwidth" and the "Throttling" tabs. Select maximum values on the "Bandwidth" tab, and corresponding average values estimates on the "Throttling" tab. Remember that a document identifier for an RSS connection is the document's URL, and the bin name for that URL is the server name. Also, please note that the "Maximum number of connections per JVM" field's default value of 10 is unlikely to be correct for connections of the RSS type; you should have at least one available connection per worker thread, for best performance. Since the default number of worker threads is 30, you should set this parameter to at least a value of 30 for normal operation.

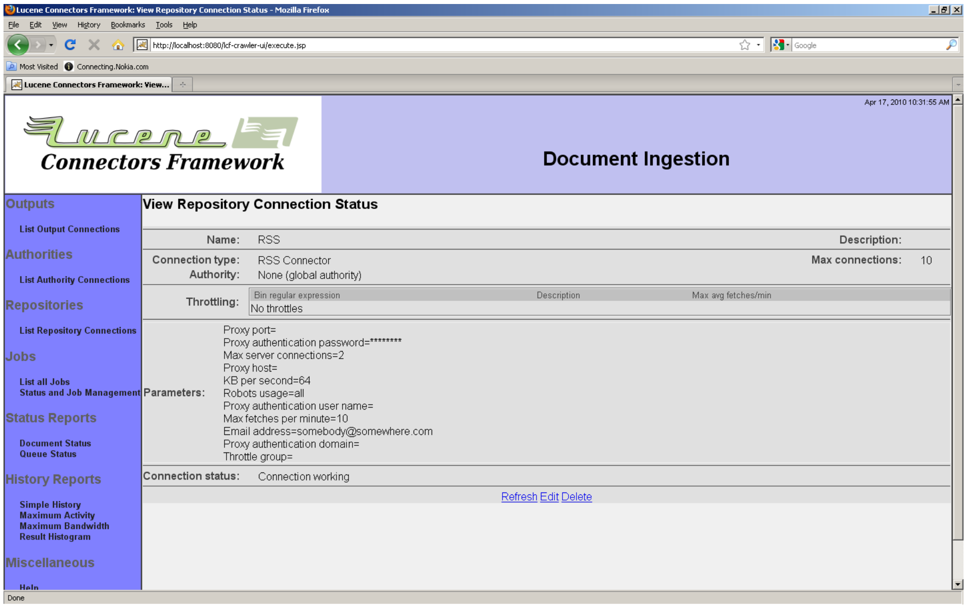
The "Proxy" tab allows you to specify a proxy that you want to crawl through. The RSS connection type supports proxies that are secured with all forms of the NTLM authentication method. This is quite typical of large organizations. The tab looks like this:



Proxy host:
Proxy port:
Proxy authentication domain:
Proxy authentication user name:
Proxy authentication password:

Enter the proxy server you will be proxying through in the "Proxy host" field. Enter the proxy port in the "Proxy port" field. If your server is authenticated, enter the domain, username, and password in the corresponding fields. Leave all fields blank if you want to use no proxy whatsoever.

When you save your RSS connection, you should see a status screen that looks something like this:


Lucene Connectors Framework: View Repository Connection Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Google
Most Visited
Connecting.Nokia.com
Lucene Connectors Framework: View...
Apr 17, 2010 10:31:55 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

View Repository Connection Status

Name: RSS
Description:
Connection type: RSS Connector
Max connections: 10
Authority: None (global authority)

Throttling:

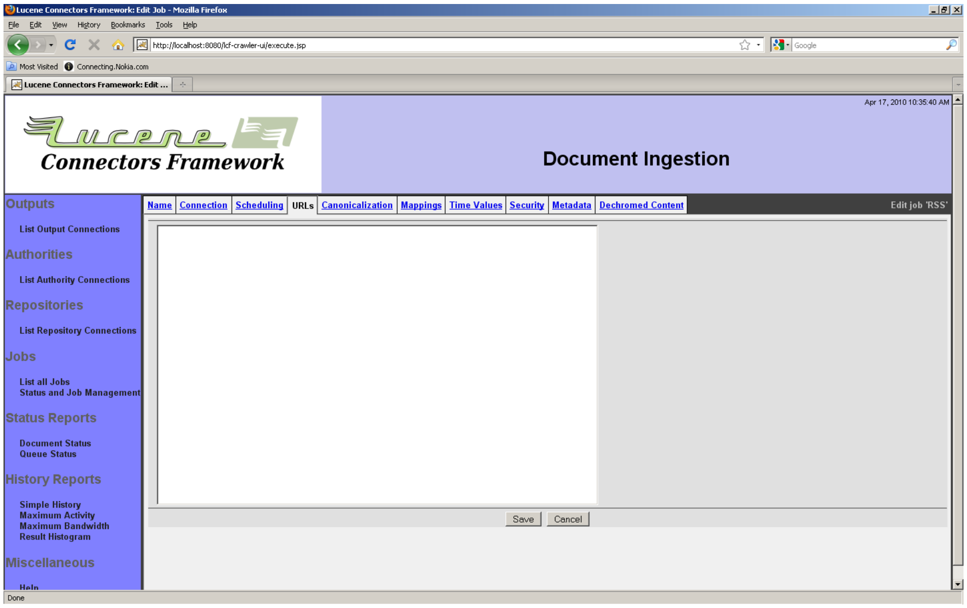
Bin regular expression	Description	Max avg fetches/min
No throttles		

Parameters:
Proxy port=
Proxy authentication password=********
Max server connections=2
Proxy host=
KB per second=64
Robots usage=all
Proxy authentication user name=
Max fetches per minute=10
Email address=somebody@somewhere.com
Proxy authentication domain=
Throttle group=

Connection status: Connection working

Refresh Edit Delete

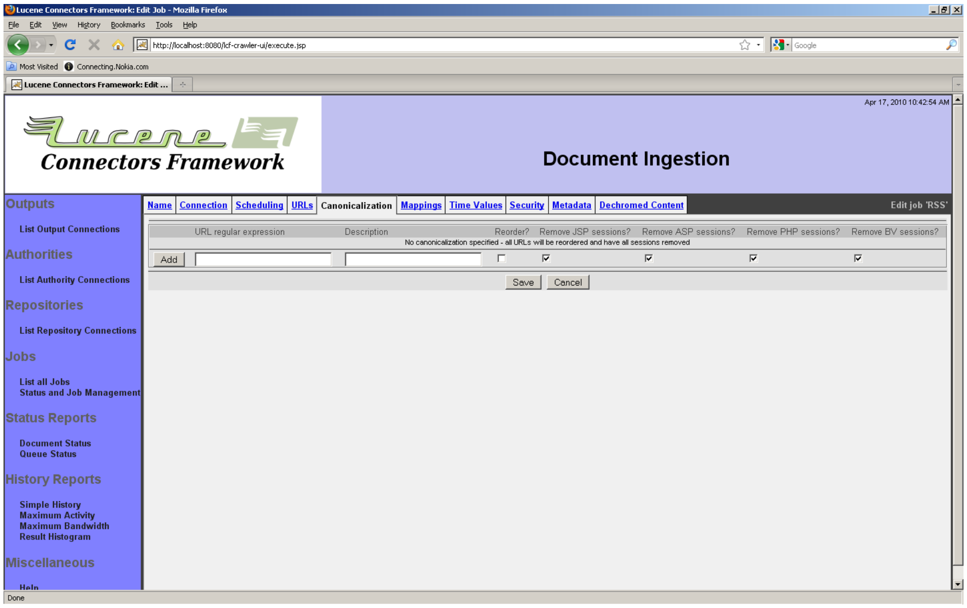
Jobs created using connections of the RSS type have the following additional tabs: "URLs", "Canonicalization", "Mappings", "Time Values", "Security", "Metadata", and "Dechromed Content". The URLs tab is where you describe the feeds that are part of the job. It looks like this:



Enter the list of feed URLs you want to crawl, separated by newlines. You may also have comments by starting lines with ("#") characters.

The "Canonicalization" tab controls how the job handles url canonicalization. Canonicalization refers to the fact that many different URLs may all refer to the same actual resource. For example, arguments in URLs can often be reordered, so that a=1&b=2 is in fact the same as b=2&a=1. Other canonical operations include removal of session cookies, which some dynamic web sites include in the URL.

The "Canonicalization" tab looks like this:



Apr 17, 2010 10:42:54 AM

Document Ingestion

Edit job 'RSS'

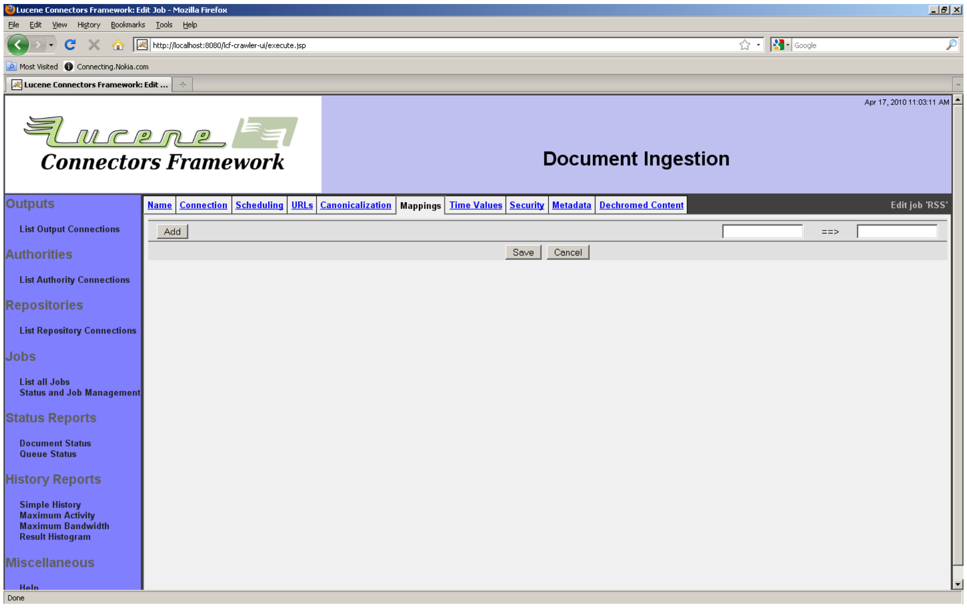
	URL regular expression	Description	Reorder?	Remove JSP sessions?	Remove ASP sessions?	Remove PHP sessions?	Remove BV sessions?
No canonicalization specified - all URLs will be reordered and have all sessions removed							
Add			□	☑	☑	☑	☑

The tab displays a list of canonicalization rules. Each rule consists of a regular expression (which is matched against a document's URL), and some switch selections. The switch selections allow you to specify whether arguments are reordered, or whether certain specific kinds of session cookies are removed. Specific kinds of session cookies that are recognized and can be removed are: JSP (Java applications servers), ASP (.NET), PHP, and Broadvision (BV).

If a URL matches more than one rule, the first matching rule is the one selected.

To add a rule, enter an appropriate regular expression, and make your checkbox selections, then click the "Add" button.

The "Mappings" tab permits you to change the URL under which documents that are fetched will get indexed. This is sometimes useful in an intranet setting because the crawling server might have open access to content, while the users may have restricted access through a somewhat different URL. The tab looks like this:



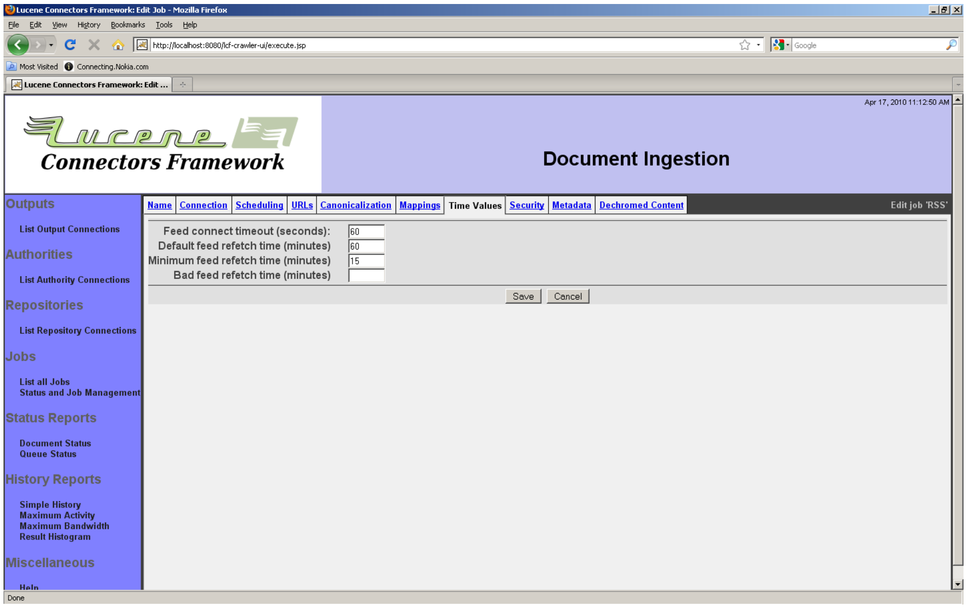
The "Mappings" tab uses the same regular expression/replacement string paradigm as is used by many connection types running under the Framework. The mappings consist of a list of rules. Each rule has a match expression, which is a regular expression where parentheses ("(" and ")") mark sections you are interested in. These sections are called "groups" in regular expression parlance. The replace string consists of constant text plus substitutions of the groups from the match, perhaps modified. For example, "$(1)" refers to the first group within the match, while "$(1l)" refers to the first match group mapped to lower case. Similarly, "$(1u)" refers to the same characters, but mapped to upper case.

For example, suppose you had a rule which had "http://(.*)/(.*)/" as a match expression, and "http://$(2)/" as the replace string. If presented with the path `http://Server/Folder_1/Filename`, it would output the string `http://Folder_1/Filename`.

If more than one rule is present, the rules are all executed in sequence. That is, the output of the first rule is modified by the second rule, etc.

To add a rule, fill in the match expression and output string, and click the "Add" button.

The "Time Values" tab looks like this:

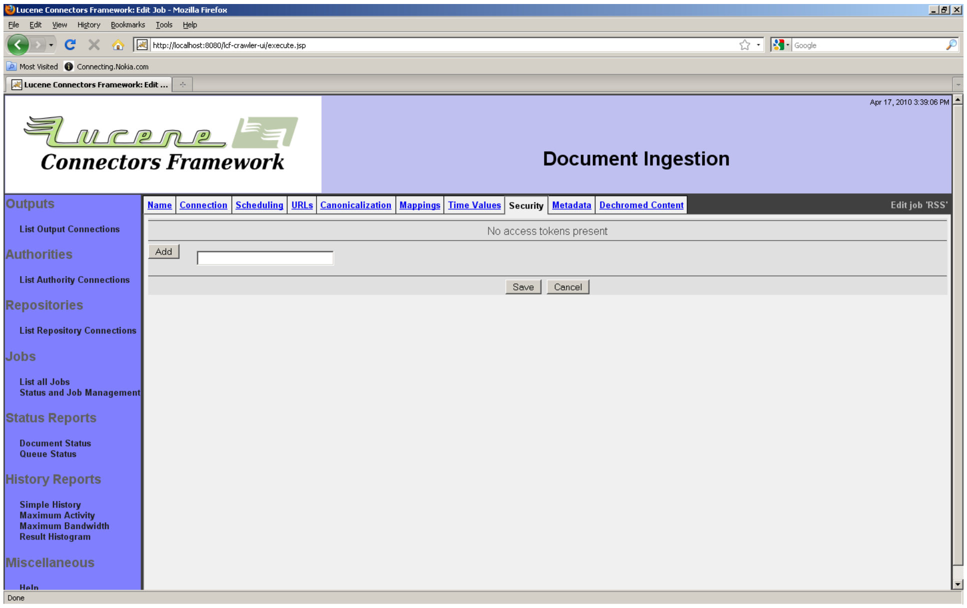


Feed connect timeout (seconds):	60
Default feed refetch time (minutes)	60
Minimum feed refetch time (minutes)	15
Bad feed refetch time (minutes)	

Fill in the desired time values. A description of each value is below.

Value	Description
Feed connect timeout	How long to wait, in seconds, before giving up, when trying to connect to a server
Default feed refetch time	If a feed specifies no refetch time, this is the time to use instead (in minutes)
Minimum feed refetch time	Never refetch feeds faster than this specified time, regardless of what the feed says (in minutes)
Bad feed refetch time	How long to wait before trying to refetch a feed that contains parsing errors (in minutes, empty is infinity)

The "Security" tab allows you to assign access tokens to the documents indexed with this job. In order to use it, you must first decide what authority connection to use to secure these documents, and what the access tokens from that authority connection look like. The tab itself looks like this:



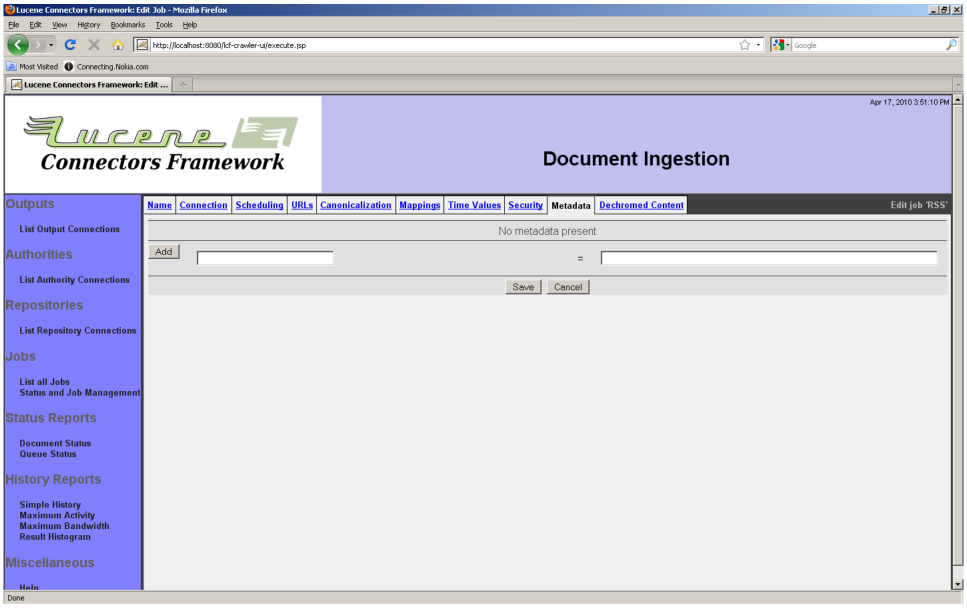
No access tokens present

To add an access token, fill in the text box with the access token value, and click the "Add" button. If there are no access tokens, security will be considered to be "off" for the job.

The "Metadata" tab allows you to specify arbitrary metadata to be indexed along with every document from this job. Documents from connections of the RSS type already receive some metadata having to do with the feed that referenced them. Specifically:

Name	Meaning
PubDate	This contains the document origination time, in milliseconds since Jan 1, 1970. The date is either obtained from the feed, or if it is absent, the date of fetch is included instead.
Source	This is the name of the feed that referred to the document.
Title	This is the title of the document within the feed.
Category	This is the category of the document within the feed.

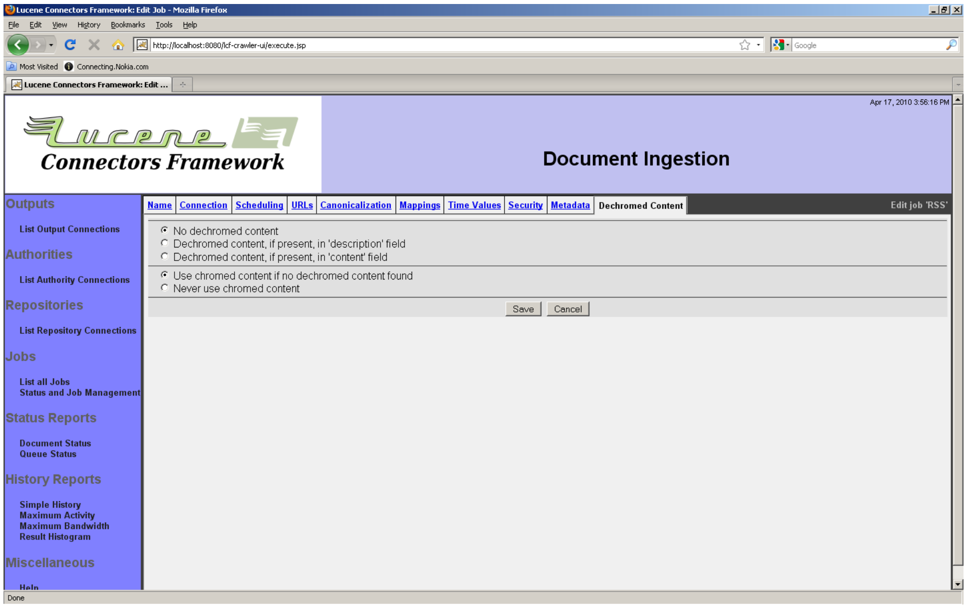
You can add additional metadata to each document using the "Metadata" tab. The tab looks like this:



No metadata present

Enter the name of the metadata item you want on the left, and its desired value on the right, and click the "Add" button to add it to the metadata list.

The "Dechromed Content" tab allows you to index the description of the content from the feed, instead of the document's contents. This is helpful when the description of the documents in the feeds you are crawling is sufficient for indexing purposes, and the actual documents are full of navigation clutter or "chrome". The tab looks like this:



● No dechromed content
○ Dechromed content, if present, in 'description' field
○ Dechromed content, if present, in 'content' field

● Use chromed content if no dechromed content found
○ Never use chromed content

Select the mode you want the connection to operate in.

4.3 Generic Web Repository Connection

The Web connection type is effectively a reasonably full-featured web crawler. It is capable of handling most kinds of authentication (basic, all forms of NTLM, and session-based), and can extract links from the following kinds of documents:

- Text
- HTML
- Generic XML
- RSS feeds

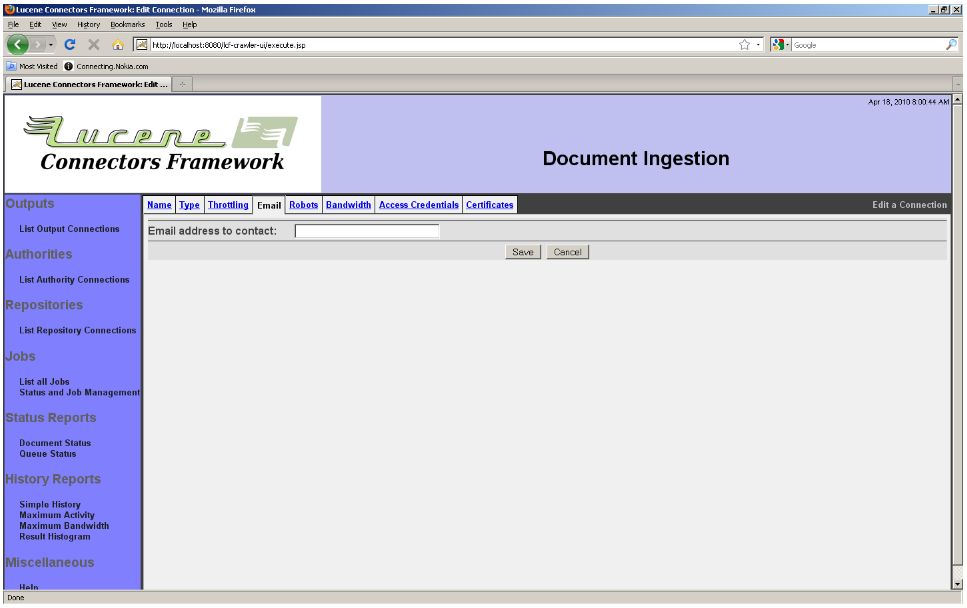
The Web connection type differs from the RSS connection type in the following respects:

- Feeds are indexed, if the output connection accepts them
- Links are extracted from all documents, not just feeds
- Feeds are treated just like any other kind of document - you cannot control how often they refetch independently
- There is support for limiting crawls based on hop count
- There is support for controlling exactly what URLs are considered part of the set, and which are excluded

In other words, the Web connection type is neither as easy to configure, nor as well-targeted in its separation of links and data, as the RSS connection type. For that reason, we strongly encourage you to consider using the RSS connection type for all applications where it might reasonably apply.

Many users of the Web connection type set up their jobs to run continuously, configuring their jobs to occasionally refetch documents, or to not refetch documents ever, and expire them after some period of time.

A Web connection has the following special tabs: "Email", "Robots", "Bandwidth", "Access Credentials", and "Certificates". The "Email" tab looks like this:

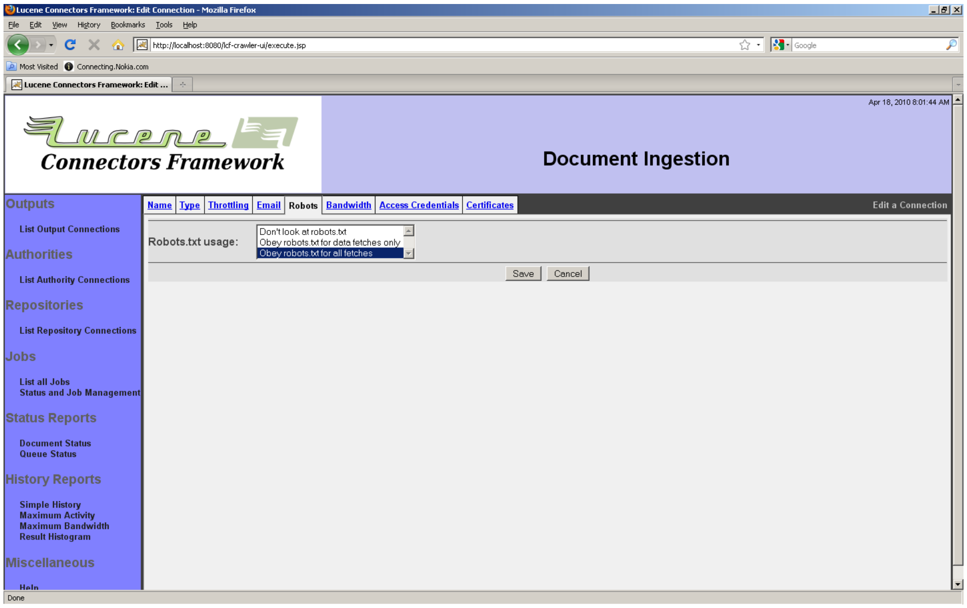


Email address to contact:

Enter an email address. This email address will be included in all requests made by the Web connection, so that webmasters can report any difficulties that their sites experience as the result of improper throttling, etc.

This field is mandatory. While a Web connection makes no effort to validate the correctness of the email field, you will probably want to remain a good web citizen and provide a valid email address. Remember that it is very easy for a webmaster to block access to a crawler that does not seem to be behaving in a polite manner.

The "Robots" tab looks like this:

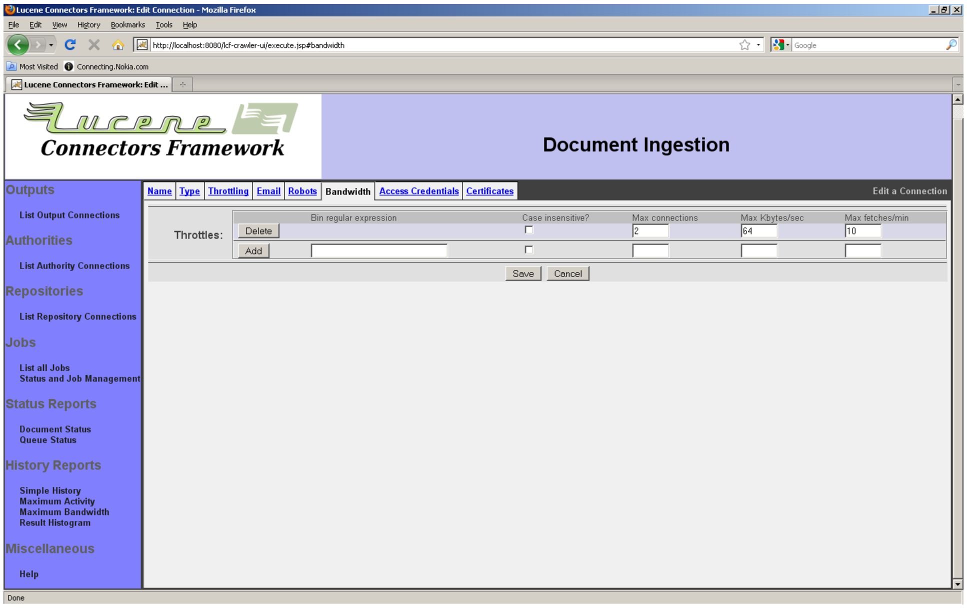


Robots.txt usage:
Don't look at robots.txt
Obey robots.txt for data fetches only
Obey robots.txt for all fetches

Select how the connection will interpret robots.txt. Remember that you have an interest in crawling people's sites as politely as is possible.

The "Bandwidth" tab allows you to specify a list of bandwidth rules. Each rule has a regular expression matched against a URL's throttle bin. Throttle bins, in connections of the Web type, are simply the server name part of the URL. Each rule allows you to select a maximum bandwidth, number of connections, and fetch rate. You can have as many rules as you like; if a URL matches more than one rule, then the most conservative value will be used.

This is what the "Bandwidth" tab looks like:



Throttles:

	Bin regular expression	Case insensitive?	Max connections	Max Kbytes/sec	Max fetches/min
Delete		□	2	64	10
Add		□			

The screen shot shows the tab configured with a setting that is reasonably polite. The default value for this tab is blank, meaning that, by default, there is no throttling whatsoever! Please do not make the mistake of crawling other people's sites without adequate politeness parameters in place.

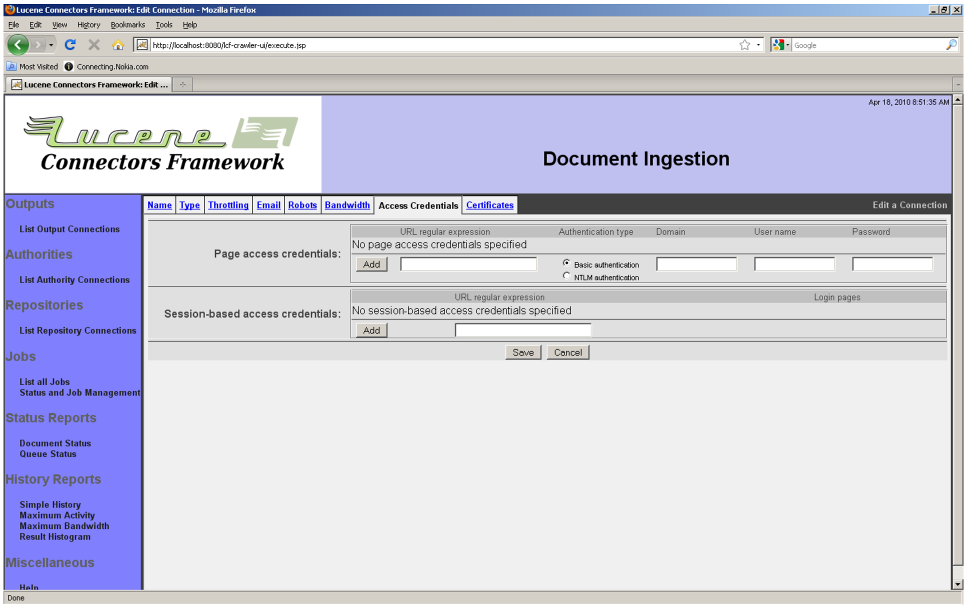
To add a rule, fill in the regular expression and the appropriate rule limit values, and click the "Add" button.

The "Bandwidth" tab is related to the throttles that you can set on the "Throttling" tab in the following ways:

- The "Bandwidth" tab sets the maximum values, while the "Throttling" tab sets the average values.
- The "Bandwidth" tab does not affect how documents are scheduled in the queue; it simply blocks documents until it is safe to go ahead, which will use up a crawler thread for the entire period that both the wait and the fetch take place. The "Throttling" tab affects how often documents are scheduled, so it does not waste threads.

Because of the above, we suggest that you configure your Web connection using both the "Bandwidth" and the "Throttling" tabs. Select maximum values on the "Bandwidth" tab, and corresponding average values estimates on the "Throttling" tab. Remember that a document identifier for a Web connection is the document's URL, and the bin name for that URL is the server name. Also, please note that the "Maximum number of connections per JVM" field's default value of 10 is unlikely to be correct for connections of the Web type; you should have at least one available connection per worker thread, for best performance. Since the default number of worker threads is 30, you should set this parameter to at least a value of 30 for normal operation.

The Web connection's "Access Credentials" tab describes how pages get authenticated. There is support on this tab for both page-based authentication (e.g. basic auth or all forms of NTLM), as well as session-based authentication (which involves the fetch of many pages to establish a logged-in session). The initial appearance of the "Access Credentials" tab shows both kinds of authentication:



Page access credentials: No page access credentials specified

● Basic authentication
○ NTLM authentication

Session-based access credentials: No session-based access credentials specified

Comparing Page and Session Based Authentication:

Authentication Detail	Page Based Authentication	Session Based Authentication
HTTP Return Codes	4xx range, usually 401	Usually 3xx range, often 301 or 302
How it's recognized as a login request	4xx range codes always indicate a challenged response	Recognized by patterns in the URL or content. Manifold must be told what to look for. 3xx range HTTP codes are also used for normal content redirects so there's no built-in way

		for Manifold to tell the difference, that's why it needs regex-based rules.
How Login form is Rendered in normal Web Browser	Standard Browser popup dialog. IE, Firefox, Safari, etc. all have their own specific style.	Server sends custom HTML or Javascript. Might use red text, might not. Might show a login form, or maybe a "click here to login" link. Can be a regular page, or Javascript popup, there's no specific standard.
Login Expiration	Usually doesn't expire, depends on server's policy. If it does expire at all, usually based calendar dates and not related to this specific login.	Often set to several minutes or hours from the the last login in current browser session. A long spider run might need to re-login several times.
HTTP Header Fields	From server: WWW-Authenticate: Basic or NTLM with Realm From client: Authorization: Basic or NTLM	From server: Location: and Set-Cookie: From client: Cookie: Cookie values frequently change.

Each kind of authentication has its own list of rules.

Specifying a page authentication rule requires simply knowing what URLs are protected, and what the proper authentication method and credentials are for those URLs. Enter a regular expression describing the protected URLs, and select the proper authentication method. Fill in the credentials. Click the "Add" button.

Specifying a correct session authentication rule usually requires some research. A single session-authentication rule usually corresponds to a single session-protected site. For that site, you will need to be able to describe the following for session authentication to function:

- The URLs of pages that are protected by this particular site session security
- How to detect when a page fetch is part of the login sequence
- How to fill in the appropriate forms within the login sequence with appropriate login information

A Web connection labels pages that are part of the login sequence "login pages", and pages that are protected site content "content pages". A Web connection will not attempt to index login pages. They are special pages that have but one purpose: establishing an authenticated session.

Remember, the goals of the setup you have to go through are as follows:

- Identify what site, or part of the site, has protected content
- Identify which http/https fetches are not content, but are in fact part of a "login sequence", which a normal person has to go through to get the appropriate cookies

If all this is not complicated enough, your research also has to cover two very different cases: when you are first entering the site anew, and second when you try to fetch a content page and you are no longer logged in, because your session has expired. In both cases, the session authentication rule must be able to properly log in and fetch content, because you cannot control when a page will be fetched or refetched by the Framework.

One key piece of data you will supply is a regular expression that basically describes the set of URLs for which the content is protected, and for which the right cookies have to be in place for you to get at the "real" content. Once you've specified this, then for each protection zone (described by its URL regexp), you need to specify how ManifoldCF should identify whether a given fetch should be considered part of the login sequence or not. It's not enough to just identify the URL of login pages, since (for instance) if your session has expired you may well have a redirection get fetched instead of the content you want. So you specify each class of login page as one of three types, using not only the URL to identify the class (this is where you get the second regexp), but also something about what is on the page: whether it is a redirection to a URL (yes, again described by a URL regexp), whether it has a form with a specified name (described by a regexp), or whether it has a specific link on it (once again, described by a regexp).

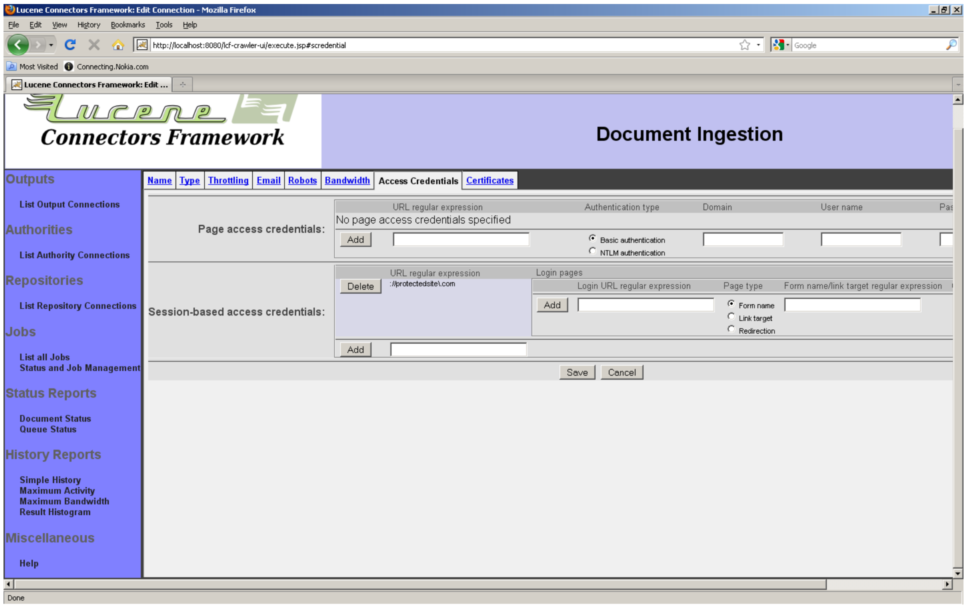
As you can see, you declare a page to be a login page by identifying it both by its URL, and by what the crawler finds on the page when it fetches it. For example, some session-protected sites may redirect you to a login screen when your session expires. So, instead of fetching content, you would be fetching a redirection to a specific page. You do not want either the redirection, or the login screen, to be considered content pages. The correct way to handle such a setup would be to declare one kind of login page to consist of a redirection to the login screen URL, and another kind of login page to consist of the login screen URL with the appropriate form. Furthermore, you would want to supply the correct login data for the form, and allow the form to be submitted, and so the login form's target may also need to be declared as a login page.

The kinds of content that a Web connection can recognize as a login page are the following:

- A redirection to a specific URL, as described by a regular expression
- A page that has a form of a particular name on it, as described by a regular expression
- A page that has a link on it to a specific target, as described by a regular expression

Note that in all three case above that there is an implicit flow through the login sequence that you describe by specifying the pages in the login sequence. For example, if upon session timeout you expect to see a redirection to a link, or family of links (remember, it's a regexp, so you can describe that easily), then as part of identifying the redirection as belonging to the login sequence, the web connector also now has a new link to fetch - the redirection link - which is what it does next. The same applies to forms. If the form name that was specified is found, then the web connector submits that form using values for the form elements that you specify, and using the submission type described in the actual form tag (GET, POST, or multi-part). Any other elements of the form are left in whatever state that the HTML specified; no Javascript is ever evaluated. Thus, if you think a form element's value is being set by Javascript, you have to figure out what it is being set to and enter this value by hand as part of the specification for the "form" type of login page. Typically this amounts to a user name and password.

To add a session authentication rule, fill in a regular expression describing the site pages that are being protected, and click the "Add" button:



Document Ingestion

Page access credentials:

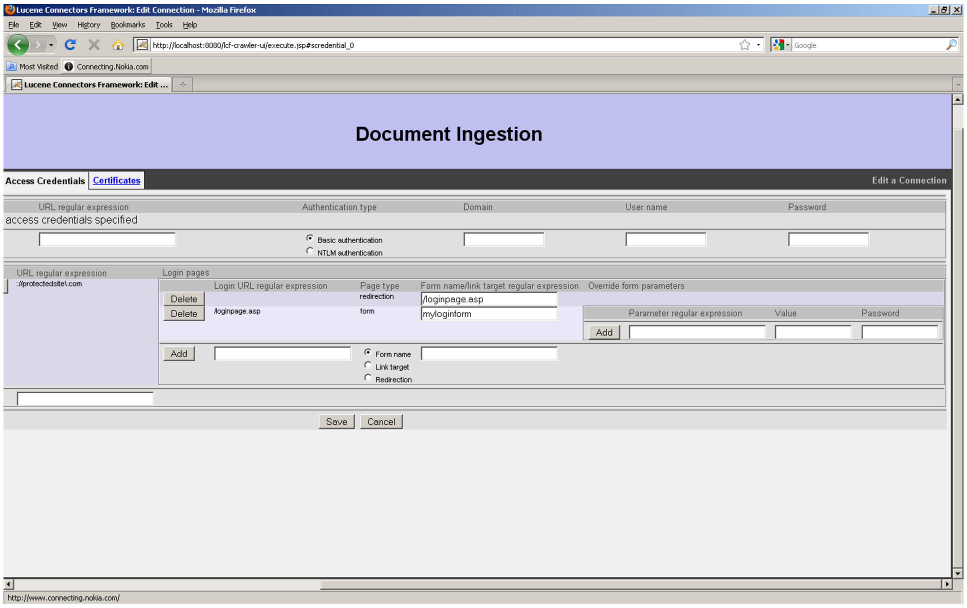
URL regular expression	Authentication type	Domain	User name	Pas
No page access credentials specified				
	● Basic authentication ○ NTLM authentication			

Session-based access credentials:

URL regular expression	Login pages		
	Login URL regular expression	Page type	Form name/link target regular expression
://protectedsite\.com		● Form name ○ Link target ○ Redirection	

Note that you can now add login page descriptions to the newly-created rule. To add a login page description, enter a URL regular expression, a type of login page, a target link or form name regular expression, and click the "Add" button.

When you add a login page of the "form" type, you can then add form fill-in information to the login page, as seen below:


Lucene Connectors Framework: Edit Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp#scredential_0
Most Visited
Connecting.Nokia.com
Lucene Connectors Framework: Edit ...
Access Credentials
Certificates
Edit a Connection
Delete
Add
Save
Cancel
http://www.connecting.nokia.com/

Document Ingestion

URL regular expression	Authentication type	Domain	User name	Password
access credentials specified				
	● Basic authentication ○ NTLM authentication			

URL regular expression: ://protectedsite\.com

Login pages

Login URL regular expression	Page type	Form name/link target regular expression	Override form parameters
	redirection	/loginpage.asp	
/loginpage.asp	form	myloginform	Parameter regular expression / Value / Password
	● Form name ○ Link target ○ Redirection		

Supply a regular expression for the name of the form element you want to set, and also provide a value. If you want the value to not be visible in clear text, fill in the "password" column instead of the "value" column. You can usually figure out the name of the form and its elements by viewing the source of the HTML page in a browser. When you are done, click the "Add" button.

Form data that is not specified will be posted with the default value determined by the HTML of the page. The Web connection type is unable, at this time, to execute Javascript, and therefore you may need to fill out some form values that are filled in by Javascript in order to get the form to post in a useful way. If you have a form that relies heavily on Javascript to post properly, you may need considerable effort and web programming skills to figure out how to get these forms to post properly with a Web connection. Luckily, such obfuscated login screens are still rare.

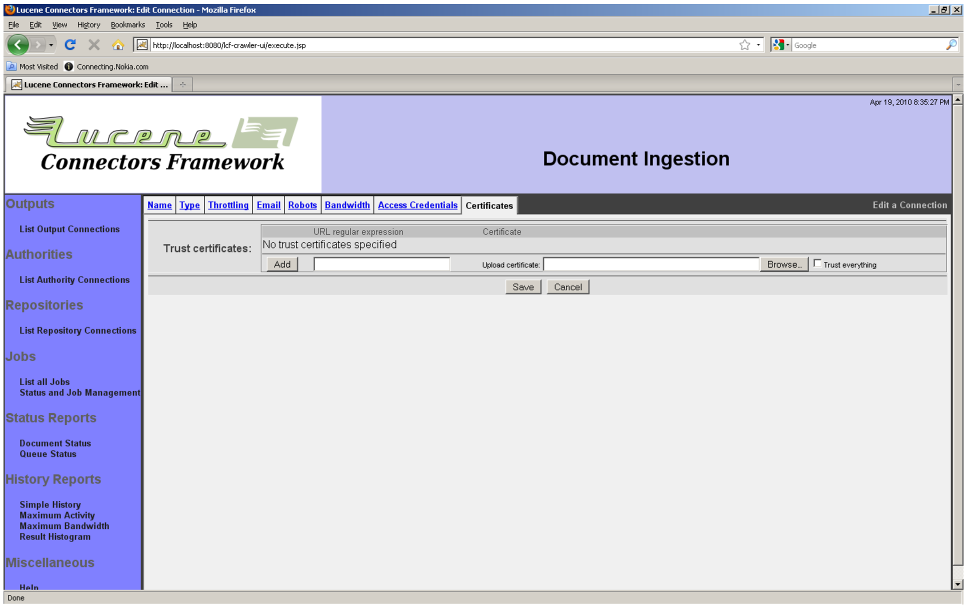
A series of login pages form a "login page sequence" for the site. For each login page, the Web connection decides what page to fetch next by what you specified for the login page criteria. So, for a redirection to a specific URL, the next page to be fetched will be that redirected URL. For a form, the next page fetched will be the action page indicated by the specified form. For a link to a target, the next page fetched will be the target URL. When the login page sequence ends, the next page fetched after that will be the original content page that the Web connection was trying to fetch when the login sequence started.

Debugging session authentication problems is best done by looking at a Simple History report for your Web connection. A Web connection records several types of events which, between them, can give a very strong picture of what is happening. These event types are as follows:

Event type	Meaning
Fetch	This event records the fetch of a URL. The HTTP response is recorded as the response code. In addition, there are several negative code values which the connect generates when the HTTP operation cannot be done or does not complete.
Begin login	This event occurs when the connection detects the transition to a login page sequence. When a login sequence is

	entered, no other pages from that protected site will be fetched until the login sequence is completed.
End login	This event occurs when the connection detects the transition from a login page sequence back to normal content fetching. When this occurs, simultaneous fetching for pages from the site are re-enabled.

The "Certificates" tab is used in conjunction with SSL, and permits you to define independent trust certificate stores for URLs matching specified regular expressions. You can also allow the connection to trust all certificates it sees, if you so choose. The "Certificates" tab looks like this:



Trust certificates:

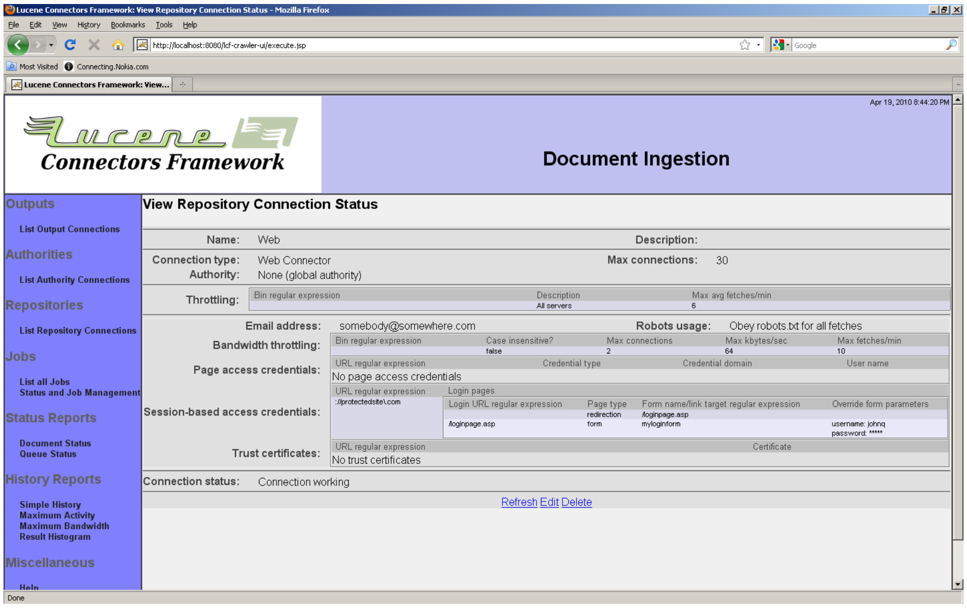
URL regular expression	Certificate
No trust certificates specified	

Upload certificate:

□ Trust everything

Type in a URL regular expression, and either check the "Trust everything" box, or browse for the appropriate certificate authority certificate that you wish to trust. (It will also work to simply trust a server's certificate, but that certificate may change from time to time, as it expires.) Click "Add" to add the certificate rule to the list.

When you are done, and you click the "Save" button, you will see a summary page looking something like this:


Lucene Connectors Framework: View Repository Connection Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited
Connecting.Nokia.com
Lucene Connectors Framework: View...
Lucene Connectors Framework
Apr 19, 2010 8:44:20 PM
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Refresh Edit Delete
Done

Document Ingestion

View Repository Connection Status

Name: Web
Description:
Connection type: Web Connector
Max connections: 30
Authority: None (global authority)

Throttling:

Bin regular expression	Description	Max avg fetches/min
	All servers	6

Email address: somebody@somewhere.com
Robots usage: Obey robots.txt for all fetches

Bandwidth throttling:

Bin regular expression	Case insensitive?	Max connections	Max kbytes/sec	Max fetches/min
	false	2	64	10

Page access credentials:

URL regular expression	Credential type	Credential domain	User name

No page access credentials

Session-based access credentials:

URL regular expression: ://protectedsite\.com

Login pages:

Login URL regular expression	Page type	Form name/link target regular expression	Override form parameters
	redirection	/loginpage.asp	
/loginpage.asp	form	myloginform	username: johnq password: *****

Trust certificates:

URL regular expression	Certificate

No trust certificates

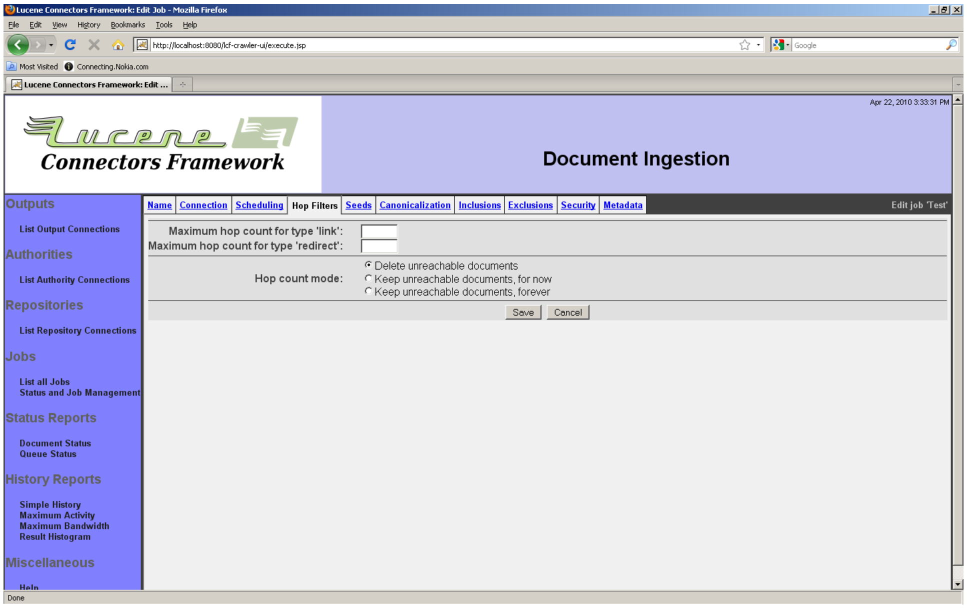
Connection status: Connection working

When you create a job that uses a repository connection of the Web type, the tabs "Hop Filters", "Seeds", "Canonicalization", "Inclusions", "Exclusions", "Security", and "Metadata" will all appear. These tabs allow you to configure the job appropriately for a web crawl.

The "Hop Filters" tab allows you to specify the maximum number of hops from a seed document that a document can be before it is no longer considered to be part of the job. For connections of the Web type, there are two different kinds of hops you can count as well: "link" hops, and "redirection" hops. Each of these represents an independent count and cutoff value. A blank value means no cutoff value at all.

For example, if you specified a maximum "link" hop count of 5, and left the "redirect" hop count blank, then any document that requires more than five links to reach from a seed will be considered out-of-set. If you specified both a maximum "link" hop count of 5, and a maximum "redirect" hop count 2, then any document that requires more than five links to reach from a seed, and more than two redirections, will be considered out-of-set.

The "Hop Filters" tab looks like this:

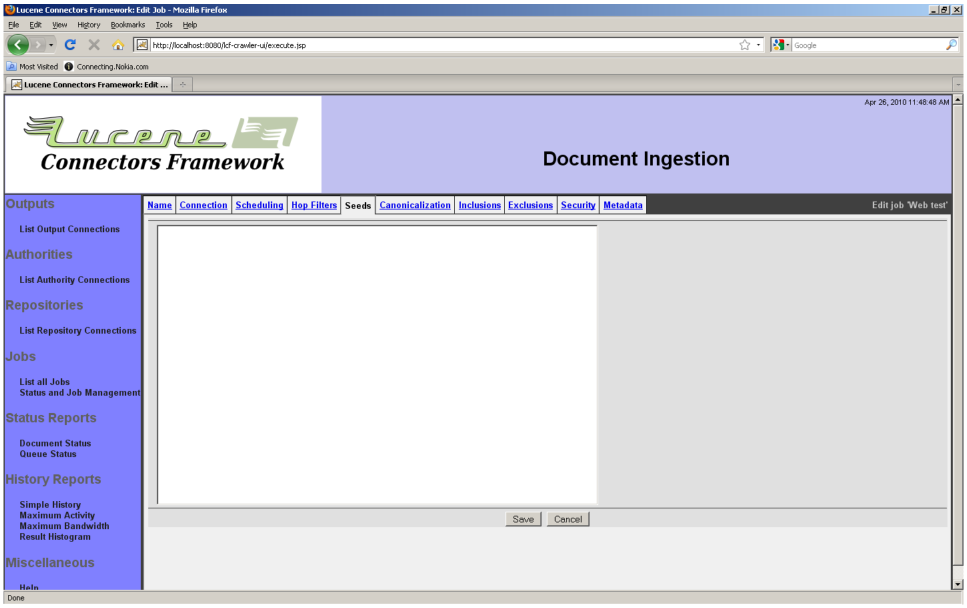


Maximum hop count for type 'link':
Maximum hop count for type 'redirect':

Hop count mode:
● Delete unreachable documents
○ Keep unreachable documents, for now
○ Keep unreachable documents, forever

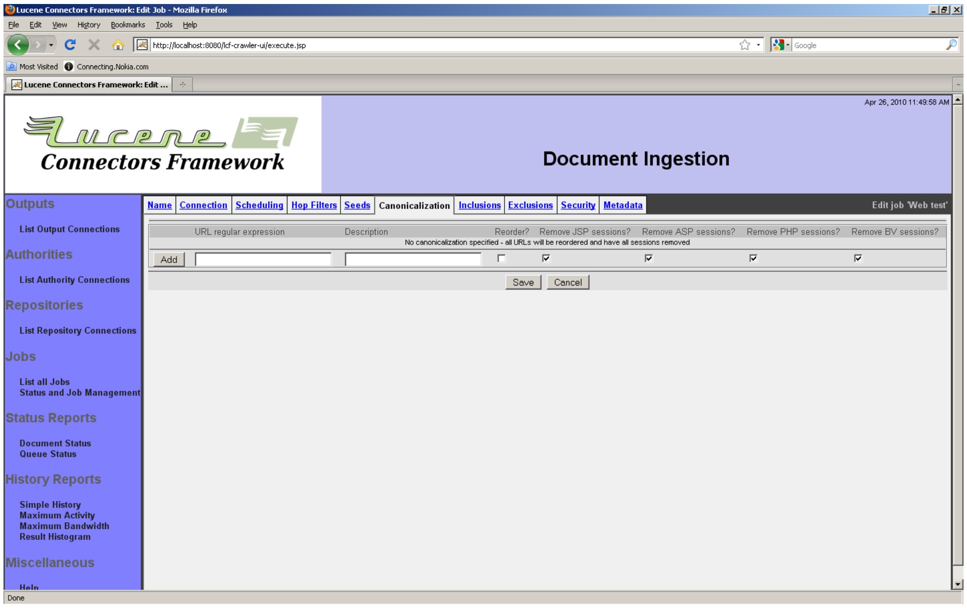
On this same tab, you can tell the Framework what to do should there be changes in the distance from the root to a document. The choice "Delete unreachable documents" requires the Framework to recalculate the distance to every potentially affected document whenever a change takes place. This may require expensive bookkeeping, however, so you also have the option of ignoring such changes. There are two varieties of this latter option - you can ignore the changes for now, with the option of turning back on the aggressive bookkeeping at a later time, or you can decide not to ever allow changes to propagate, in which case the Framework will discard the necessary bookkeeping information permanently. This last option is the most efficient.

The "Seeds" tab is where you enter the starting points for your crawl. It looks like this:



Enter a list of seeds, separated by newline characters. Blank lines, or lines that begin with a "#' character, will be ignored.

The "Canonicalization" tab controls how a web job converts URLs into a standard form. It looks like this:



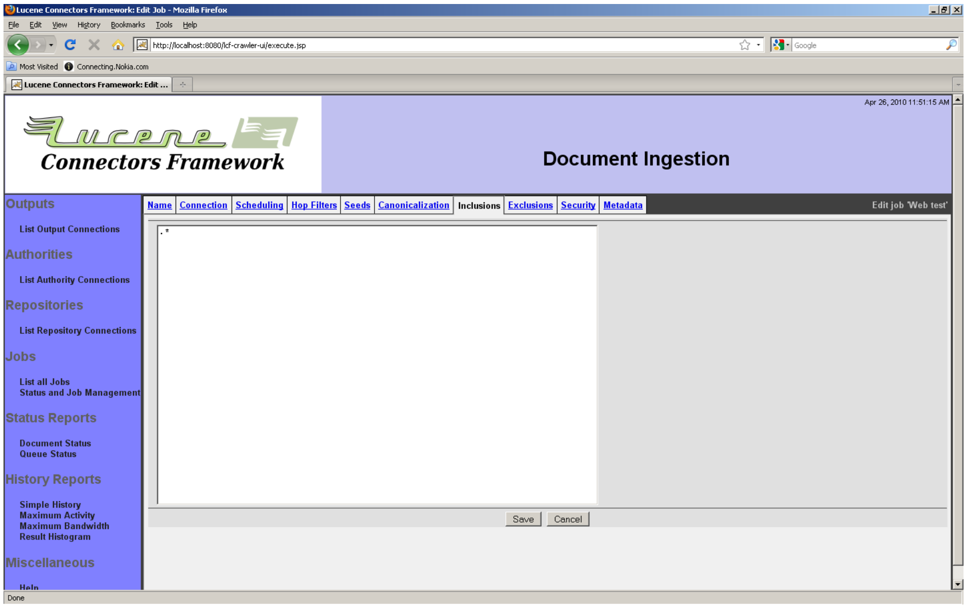
	URL regular expression	Description	Reorder?	Remove JSP sessions?	Remove ASP sessions?	Remove PHP sessions?	Remove BV sessions?
No canonicalization specified - all URLs will be reordered and have all sessions removed							
Add			□	☑	☑	☑	☑

The tab displays a list of canonicalization rules. Each rule consists of a regular expression (which is matched against a document's URL), and some switch selections. The switch selections allow you to specify whether arguments are reordered, or whether certain specific kinds of session cookies are removed. Specific kinds of session cookies that are recognized and can be removed are: JSP (Java applications servers), ASP (.NET), PHP, and Broadvision (BV).

If a URL matches more than one rule, the first matching rule is the one selected.

To add a rule, enter an appropriate regular expression, and make your checkbox selections, then click the "Add" button.

The "Inclusions" tab lets you specify, by means of a set of regular expressions, exactly what URLs will be included as part of the document set for a web job. The tab looks like this:

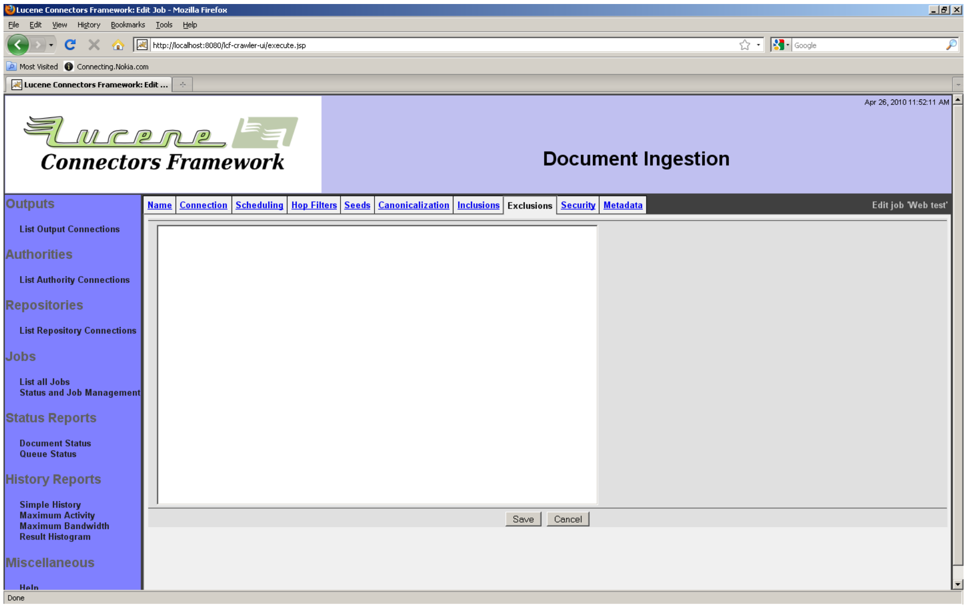


```
.*
```

You will need to provide a series of zero or more regular expressions, separated by newlines.

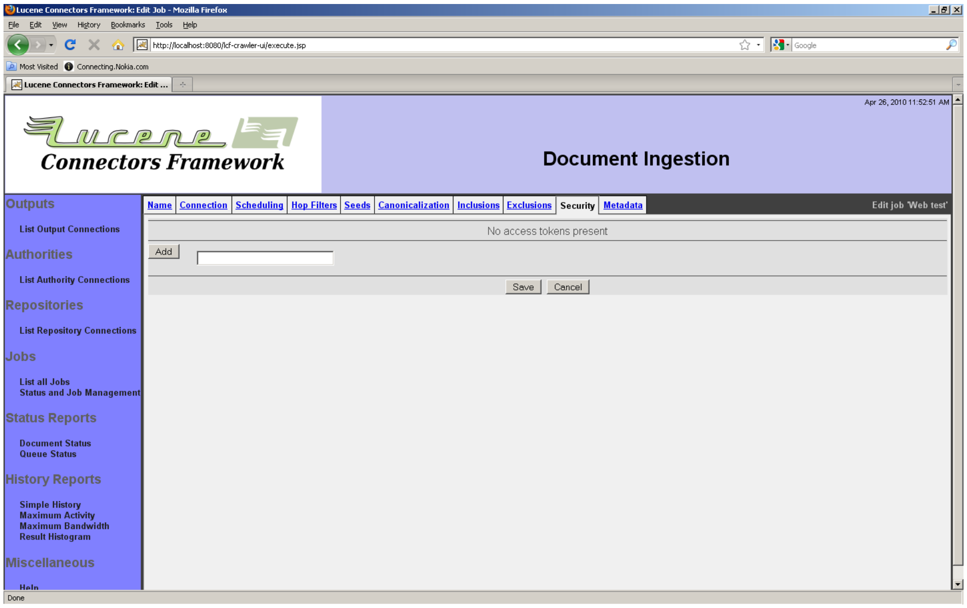
Remember that, by default, a web job includes all documents in the world that are linked to your seeds in any way that the web connection type can determine.

If you wish to restrict which documents are actually processed within your overall set of included documents, you may want to supply some regular expressions on the "Exclusions" tab, which looks like this:



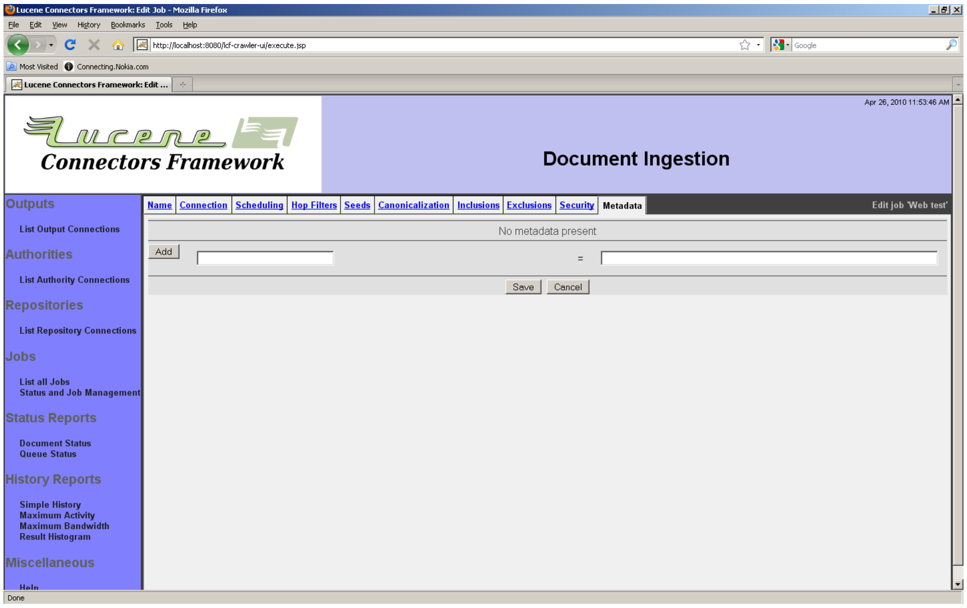
Once again you will need to provide a series of zero or more regular expressions, separated by newlines. It is typical to use the "Exclusions" tab to remove documents from consideration which are suspected to contain content that both has no extractable links, and is not useful to the index you are trying to build, e.g. movie files.

The "Security" tab allows you to specify the access tokens that the documents in the web job get indexed with, and looks like this:


No access tokens present

You will need to know the format of the access tokens for the governing authority before you can add security to your documents in this way. Enter the access token you desire and click the "Add" button.

The "Metadata" tab allows you to include specified metadata along with all documents belonging to a web job. It looks like this:



No metadata present

Enter the name of the desired metadata on the left, and the desired value (if any) on the right, and click the "Add" button.

4.4 Windows Share/DFS Repository Connection

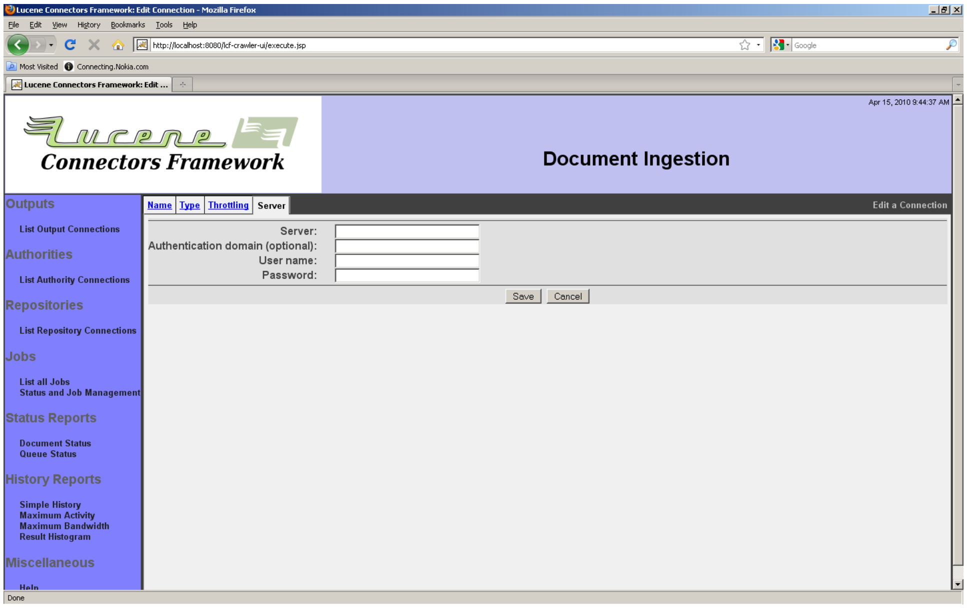
The Windows Share connection type allows you to access content stored on Windows shares, even from non-Windows systems. Also supported are Samba and various third-party Network Attached Storage servers.

DFS nodes and referrals are fully supported, provided the referral machine names can be looked up properly via DNS on the server where the Framework is running. For each document, a Windows Share connection creates an index identifier that can be either a "file:" IRI's, or a mapped "http:" URI's, depending on how it is configured. This allows for a great deal of flexibility in deployment environments, but also may require some work to properly set up. In particular, if you intend to use file IRI's as your identifiers, you should check with your system integrator to be sure these are being handled properly by the search component of your system. When you use a browser such as

Internet Explorer to view a document from a Windows file system called `\\servername\sharename\dir1\filename.txt`, the browser converts that to an IRI that looks something like this: `file://///servername/sharename/dir1/filename.txt`. While this seems simple, major complexities arise when the underlying file name has special characters in it, such as spaces, "#" symbols, or worse still, non-ASCII characters. Unfortunately, every version of Internet Explorer handles these situations somewhat differently, so there is not any fully correct way for the Windows Share connection type to convert file names to IRI's. Instead, the connection always uses a standard canonical form, and expects the search results display system component to know how to properly form the right IRI for the browser or client being used.

If you are interested in enforcing security for documents crawled with a Windows Share repository connection, you will need to first configure an authority connection of the Active Directory type to control access to these documents.

A Windows Share connection has a single special tab on the repository connection editing screen: the "Server" tab:

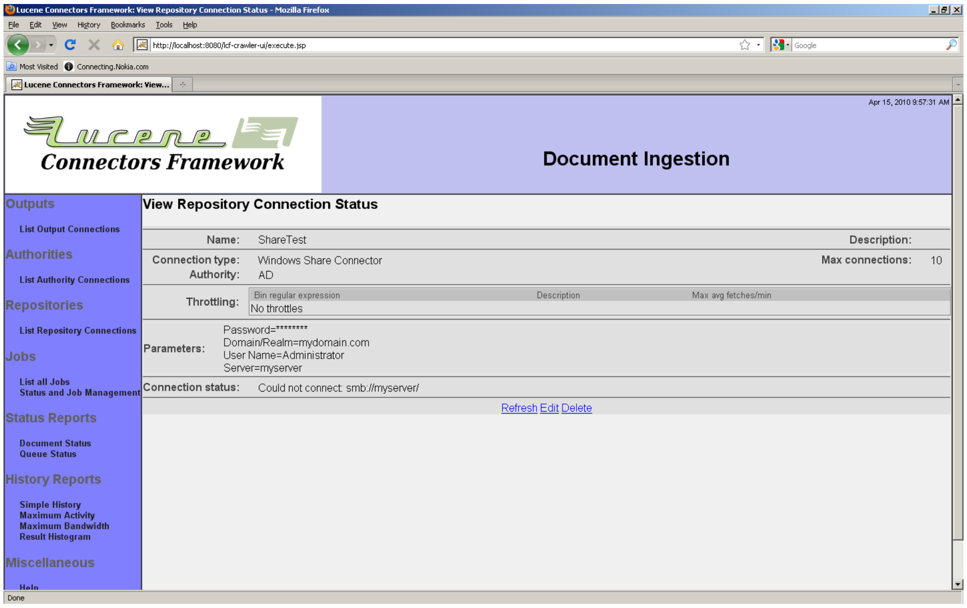


Server:
Authentication domain (optional):
User name:
Password:

You must enter the name of the server to form the connection with in the "Server" field. This can either be an actual machine name, or a domain name (if you intend to connect to a Windows domain-based DFS root). If you supply an actual machine name, it is usually the right thing to do to provide the server name in an unqualified form, and provide a fully-qualified domain name in the "Domain name" field. The user name also should usually be unqualified, e.g. "Administrator" rather than "Administrator@mydomain.com". Sometimes it may work to leave the "Domain name" field blank, and instead supply a fully-qualified machine name in the "Server" field. It never works to supply both a domain name and a fully-qualified server name.

Please note that you should probably set the "Maximum number of connections per JVM" field, on the "Throttling" tab, to a number smaller than the default value of 10, because Windows is not especially good at handling multithreaded file requests. A number less than 5 is likely to perform as well with less chance of causing server-side problems.

After you click the "Save" button, you will see a connection summary screen, which might look something like this:


Lucene Connectors Framework: View Repository Connection Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Google
Most Visited
Connecting.Nokia.com
Lucene Connectors Framework: View...
Lucene Connectors Framework
Apr 15, 2010 9:57:31 AM
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Refresh Edit Delete
Done

Document Ingestion

View Repository Connection Status

Name: ShareTest
Description:
Connection type: Windows Share Connector
Max connections: 10
Authority: AD

Throttling:

Bin regular expression	Description	Max avg fetches/min
No throttles		

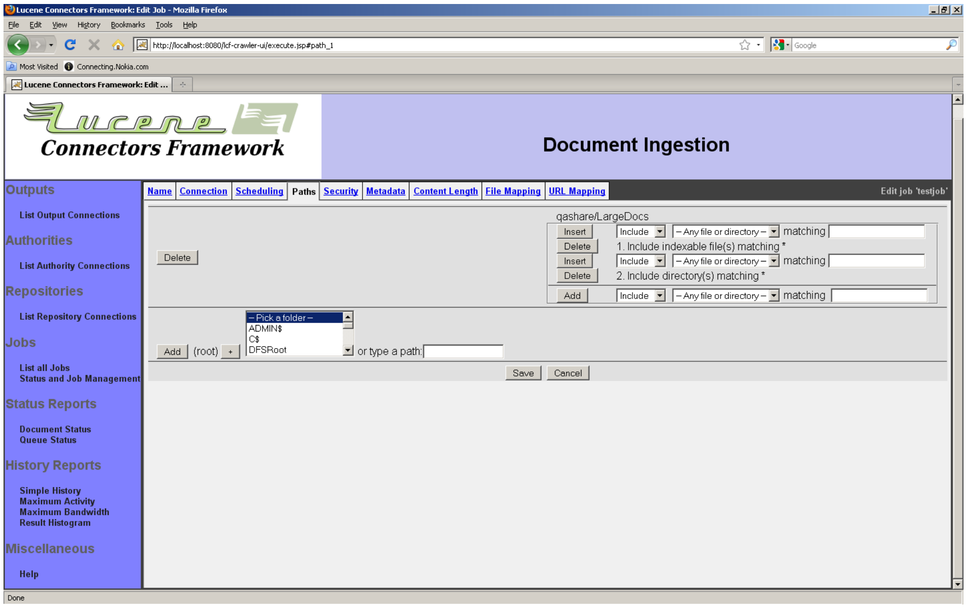
Parameters:
Password=********
Domain/Realm=mydomain.com
User Name=Administrator
Server=myserver

Connection status: Could not connect: smb://myserver/

Note that in this example, the Windows Share connection is not responding, which is leading to an error status message instead of "Connection working".

When you configure a job to use a repository connection of the Windows Share type, several additional tabs are presented. These are, in order, "Paths", "Security", "Metadata", "Content Length", "File Mapping", and "URL Mapping".

The "Paths" tab looks like this:



qashare/LargeDocs

Include – Any file or directory – matching

1. Include indexable file(s) matching *

Include – Any file or directory – matching

2. Include directory(s) matching *

Include – Any file or directory – matching

(root) – Pick a folder – ADMIN$ C$ DFSRoot or type a path:

This tab allows you to construct starting-point paths by drilling down, and then add the constructed paths to a list, or remove existing paths from the list. Without any starting paths, your job includes zero documents.

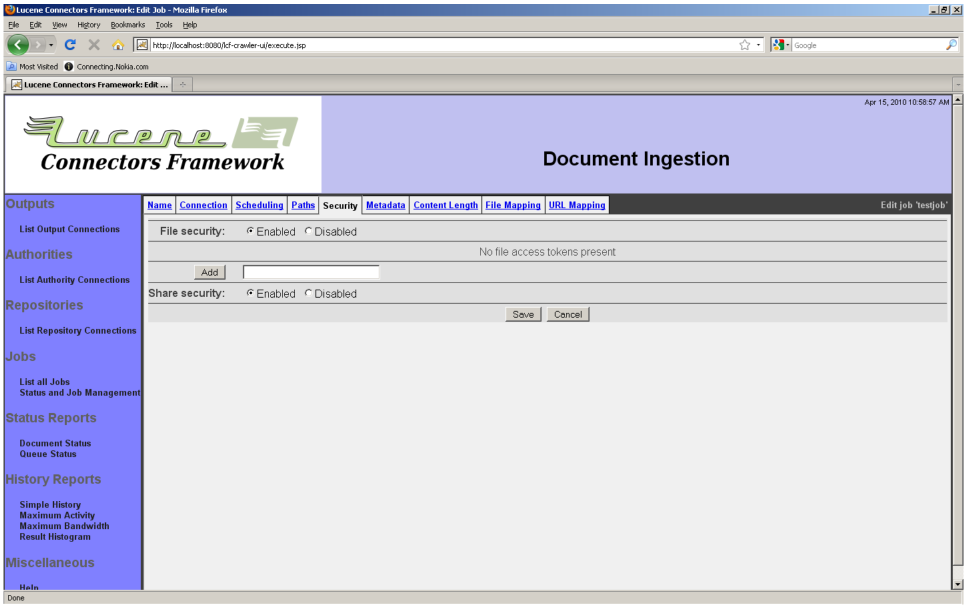
Make sure your connection has a status of "Connection working" before you open this tab, or you will see an error message, and you will not be able to build any paths.

For each included path, a list of rules is displayed which determines what folders and documents get included with the job. These rules will be evaluated from top to bottom, in order. Whichever rule first matches a given path is the one that will be used for that path.

Each rule describes the path matching criteria. This consists of the file specification (e.g. "*.txt"), whether the path is a file or folder name, and whether a file is considered indexable or not by the output connection. The rule also describes the action to take should the rule be matched: include or exclude. The file specification character "*" is a wildcard which matches zero or more characters, while the character "?" matches exactly one character. All other characters must match exactly.

To add a rule for a starting path, select the desired values of all the pulldowns, type in the desired file criteria, and click the "Add" button. You may also insert a new rule above any existing rule, by using one of the "Insert" buttons.

The "Security" tab looks like this:



File security: ● Enabled ○ Disabled

No file access tokens present

Add

Share security: ● Enabled ○ Disabled

Save Cancel

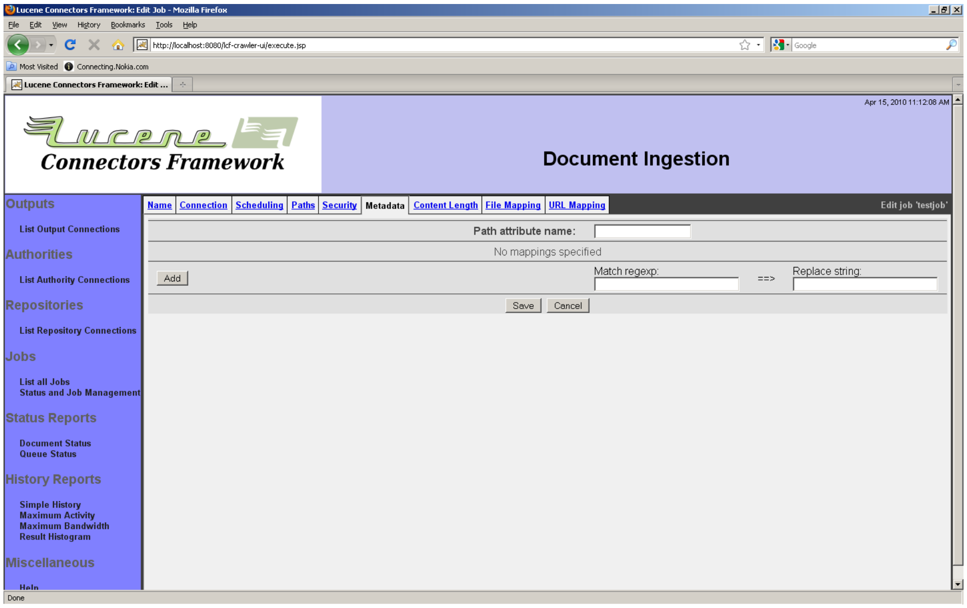
The "Security" tab lets you control three things: File security, share security, and (if security is off) the security tokens attached to all documents indexed by the job.

File security is the security Windows applies to individual files. This kind of security is supported by practically all Windows-compatible NAS-type servers, so you may use this feature without cause for concern.

Share security is the security Windows applies to Windows shares. This is an older kind of security that is no longer prevalent in most enterprise organizations. Many modern NAS systems and Samba also do not support this security model. If you enable this kind of security in your job while crawling against a system that does not support it, your job will not run correctly; the first document access will cause an error, and the job will abort.

If you turn off file security, you have the option of adding index access tokens of your own to all documents crawled by the job. These tokens must, of course, be in a form appropriate for the governing authority connection. Type the token into the box and click the "Add" button. It is unusual to use this feature other than for demonstrations.

The "Metadata" tab looks like this:



Path attribute name:

No mappings specified

Match regexp:

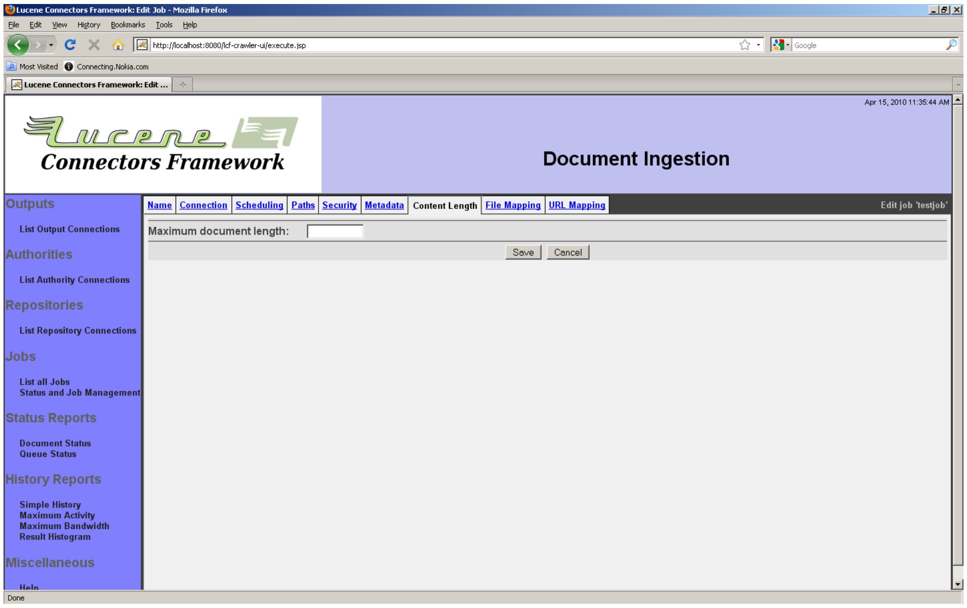
Replace string:

This tab allows you to ingest a document's path, as modified by a set of regular expression rules, as a piece of document metadata. Enter the metadata name you want in the "Path attribute name" field. Then, add the rules you want to the list of rules. Each rule has a match expression, which is a regular expression where parentheses ("(" and ")") mark sections you are interested in. These sections are called "groups" in regular expression parlance. The replace string consists of constant text plus substitutions of the groups from the match, perhaps modified. For example, "$(1)" refers to the first group within the match, while "$(1l)" refers to the first match group mapped to lower case. Similarly, "$(1u)" refers to the same characters, but mapped to upper case.

For example, suppose you had a rule which had ".*/(.*)/(.*)/.*" as a match expression, and "$(1) $(2)" as the replace string. If presented with the path `Project/Folder_1/Folder_2/Filename`, it would output the string `Folder_1 Folder_2`.

If more than one rule is present, the rules are all executed in sequence. That is, the output of the first rule is modified by the second rule, etc.

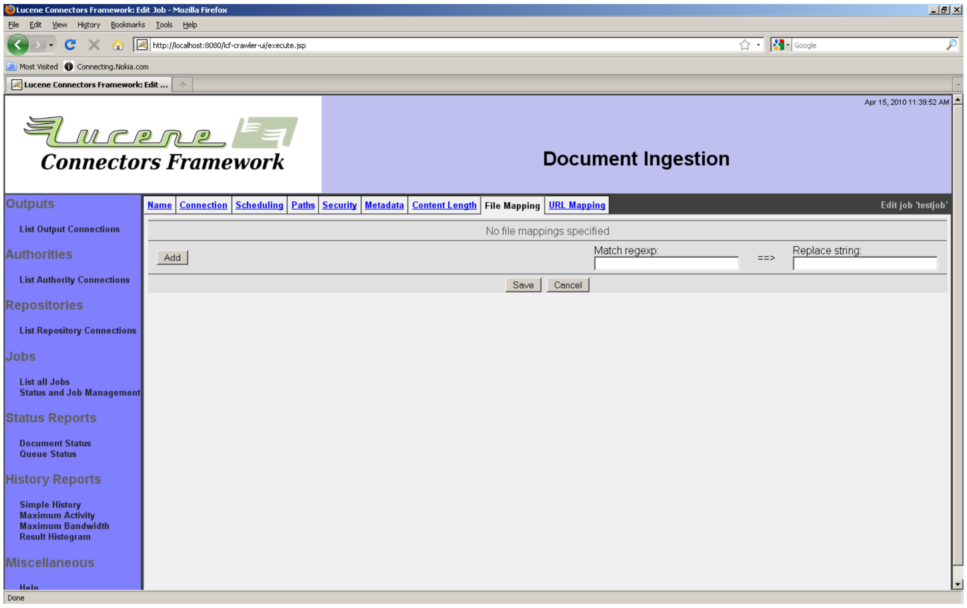
The "Content Length tab looks like this:



Maximum document length:

This tab allows you to set a maximum content length cutoff value, to avoid having the job try to index exceptionally large documents. Enter the desired maximum value. A blank value indicates an unlimited cutoff length.

The "File Mapping" tab looks like this:



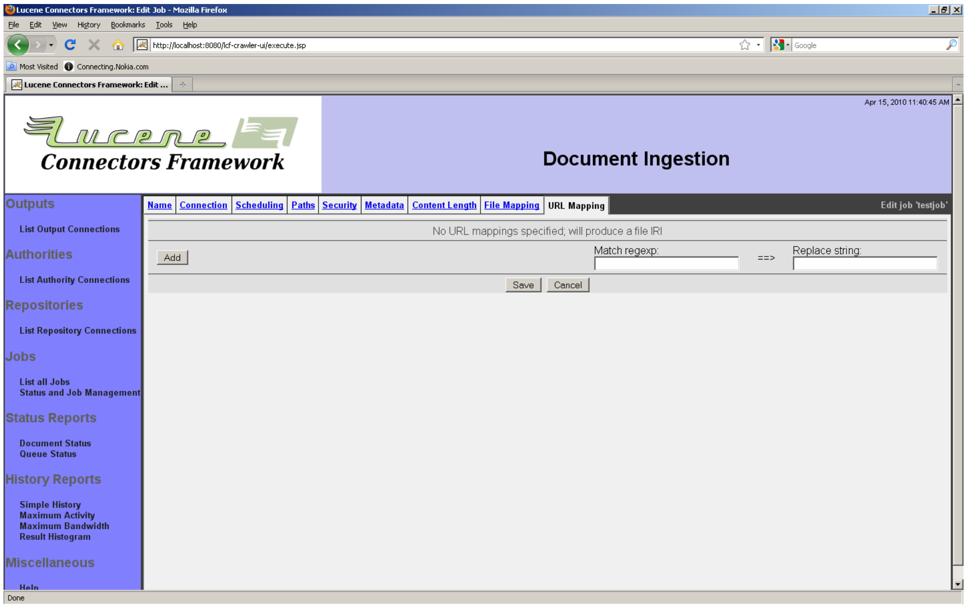
No file mappings specified

Match regexp:

Replace string:

The mappings specified here are similar in all respects to the path attribute mapping setup described above. The mappings are applied to change the actual file path discovered by the crawler into a different file path. This can sometimes be useful if there is some kind of conversion process between raw documents and parallel data files that contain extracted data.

The "URL Mapping" tab looks like this:



No URL mappings specified; will produce a file IRI

Match regexp:

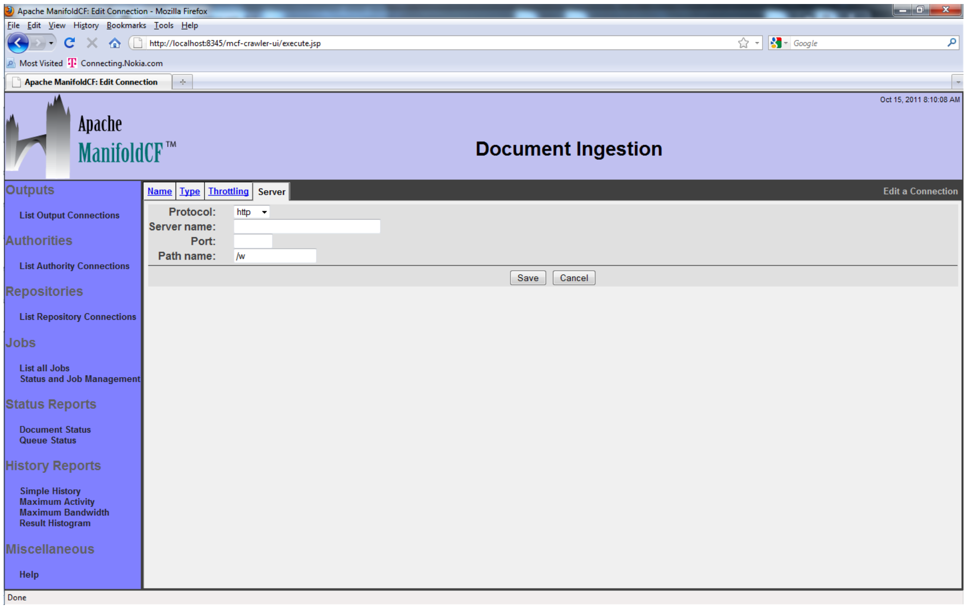
Replace string:

The mappings specified here are similar in all respects to the path attribute mapping setup described above. If no mappings are present, the file path is converted to a canonical file IRI. If mappings are present, the conversion is presumed to produce a valid URL, which can be used to access the document via some variety of Windows Share http server.

4.5 Wiki Repository Connection

The Wiki repository connection type allows you to index content from the main space of a Wiki or MediaWiki site. The connection type uses the Wiki API in order to fetch content. Only publicly visible documents will be indexed, and there is thus no need of an authority for Wiki content.

A Wiki connection has only one special tab on the repository connection editing screen: the "Server" tab. The "Server" tab looks like this:



Protocol: http
Server name:
Port:
Path name: /w

The protocol must be selected in the "Protocol" field. At the moment only the "http" protocol is supported. The server name must be provided in the "Server name" field. The server port must be provided in the "Port" field. Finally, the path part of the Wiki URL must be provided in the "Path name" field and must start with a "/" character.

When you configure a job to use a repository connection of the Wiki type, no additional tabs are currently presented.

4.6 Generic Database Repository Connection

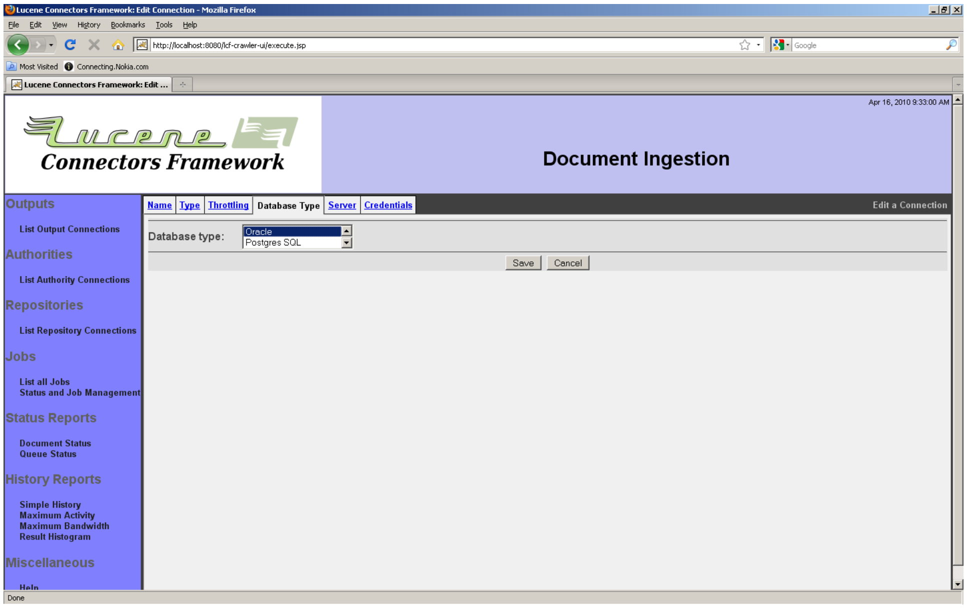
The generic database connection type allows you to index content from a database table, served by one of the following databases:

- Postgresql (via a Postgresql JDBC driver)
- SQL Server (via the JTDS JDBC driver)
- Oracle (via the Oracle JDBC driver)
- Sybase (via the JTDS JDBC driver)

This connection type cannot be configured to work with other databases without software changes. Depending on your particular installation, some of the above options may not be available.

The generic database connection type currently has no per-document notion of security. It is possible to set document security for all documents specified by a given job. Since this form of security requires you to know what the actual access tokens are, you must have detailed knowledge of the authority connection you intend to use, and what sorts of access tokens it produces.

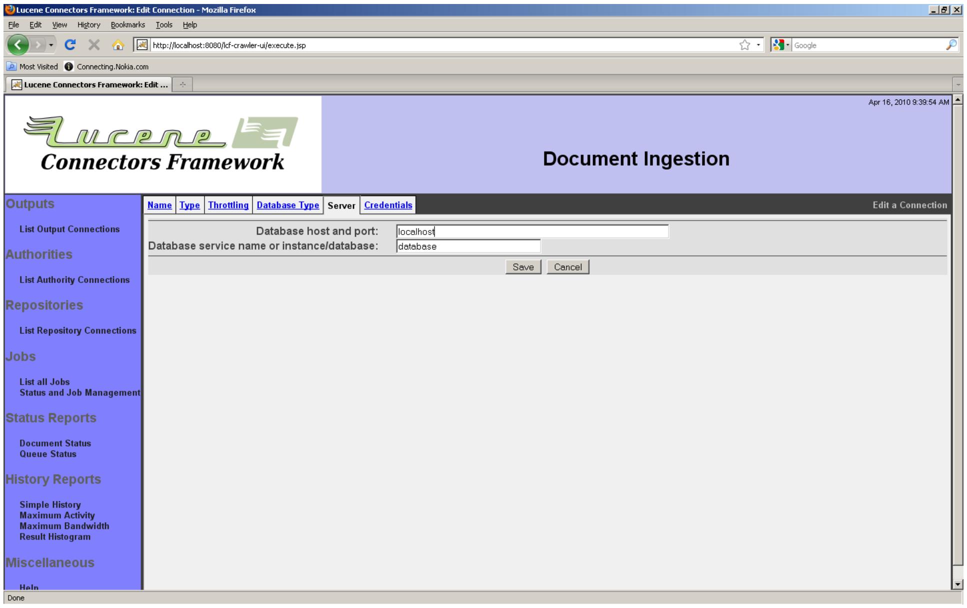
A generic database connection has three special tabs on the repository connection editing screen: the "Database Type" tab, the "Server" tab, and the "Credentials" tab. The "Database Type" tab looks like this:



Database type: Oracle

Select the kind of database you want to connect to, from the pulldown.

The "Server" tab looks like this:

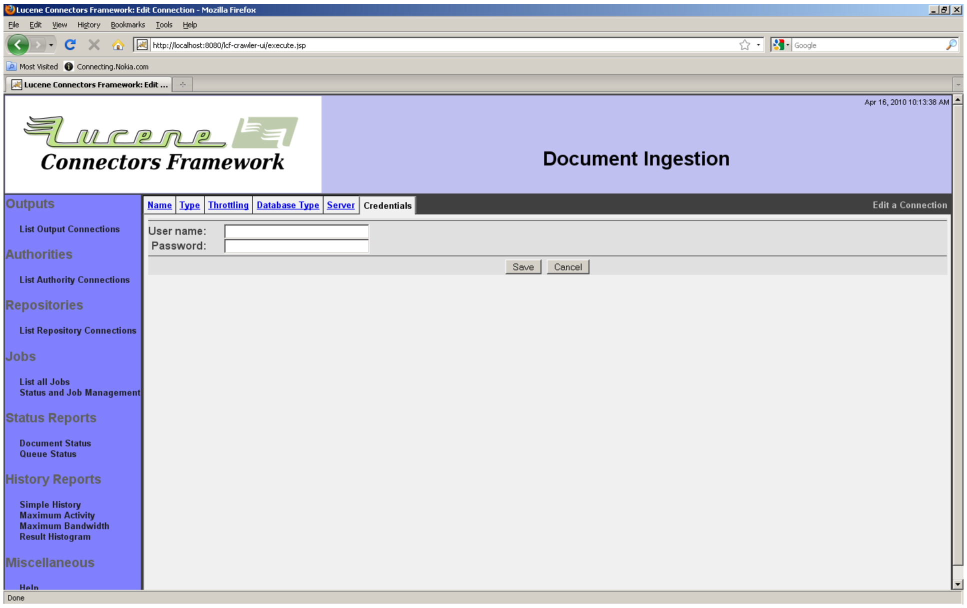


Database host and port: localhost
Database service name or instance/database: database

The server name and port must be provided in the "Database host and port" field. For example, for Oracle, the standard Oracle installation uses port 1521, so you would enter something like, "my-oracle-server:1521" for this field. Postgresql uses port 5432 by default, so "my-postgresql-server:5432" would be required. SQL Server's standard port is 1433, so use "my-sql-server:1433".

The service name or instance name field describes which instance and database to connect to. For Oracle or Postgresql, provide just the database name. For SQL Server, use "my-instance-name/my-database-name". For SQL Server using the default instance, use just the database name.

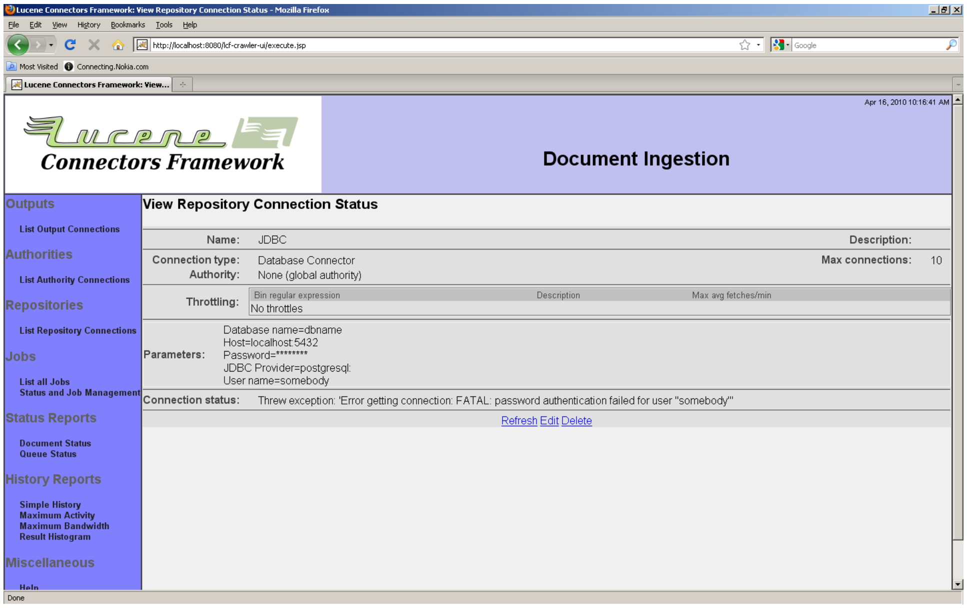
The "Credentials" tab is straightforward:



User name:
Password:

Enter the database user credentials.

After you click the "Save" button, you will see a connection summary screen, which might look something like this:


Lucene Connectors Framework: View Repository Connection Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited
Connecting.Nokia.com
Lucene Connectors Framework: View...
Apr 16, 2010 10:16:41 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Refresh
Edit
Delete
Done

View Repository Connection Status

Name: JDBC
Description:
Connection type: Database Connector
Max connections: 10
Authority: None (global authority)
Throttling:

Bin regular expression	Description	Max avg fetches/min
No throttles		

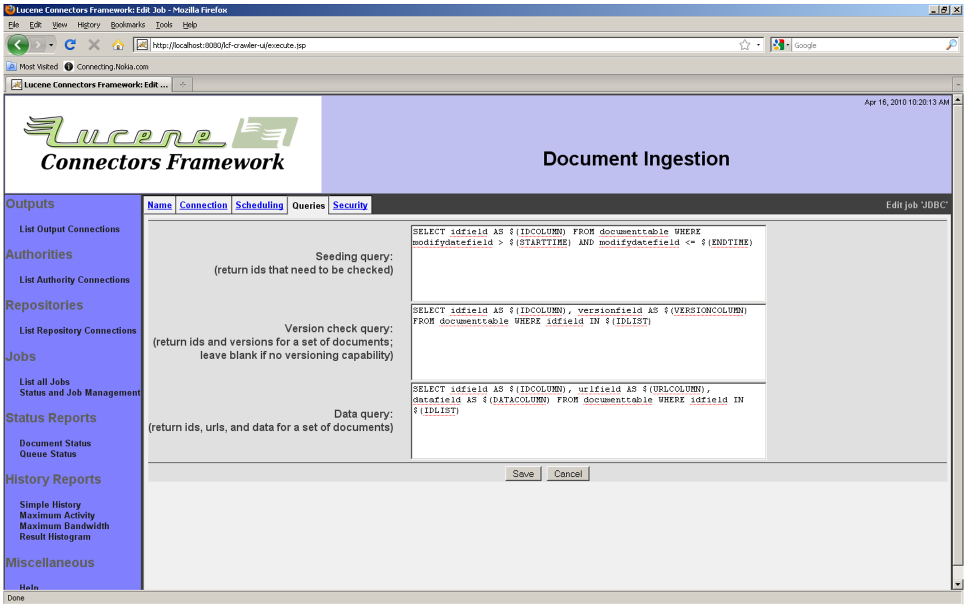
Parameters:
Database name=dbname
Host=localhost:5432
Password=********
JDBC Provider=postgresql:
User name=somebody

Connection status: Threw exception: 'Error getting connection: FATAL: password authentication failed for user "somebody"'

Note that in this example, the generic database connection is not properly authenticated, which is leading to an error status message instead of "Connection working".

When you configure a job to use a repository connection of the generic database type, several additional tabs are presented. These are, in order, "Queries", and "Security".

The "Queries" tab looks something like this:



Seeding query:
(return ids that need to be checked)

```
SELECT idfield AS $(IDCOLUMN) FROM documenttable WHERE
modifydatefield > $(STARTTIME) AND modifydatefield <= $(ENDTIME)
```

Version check query:
(return ids and versions for a set of documents;
leave blank if no versioning capability)

```
SELECT idfield AS $(IDCOLUMN), versionfield AS $(VERSIONCOLUMN)
FROM documenttable WHERE idfield IN $(IDLIST)
```

Data query:
(return ids, urls, and data for a set of documents)

```
SELECT idfield AS $(IDCOLUMN), urlfield AS $(URLCOLUMN),
datafield AS $(DATACOLUMN) FROM documenttable WHERE idfield IN
$(IDLIST)
```

You must supply at least two queries. (A third query is optional.) The purpose of these queries is to obtain the data needed for the database to be properly crawled. But in order for you to write these queries, you must make some decisions first. Basically, you need to figure out how best to map the constructs within your database to the requirements of the Framework.

- Obtain a list of document identifiers corresponding to changes and additions that occurred within a specified time window (see below)
- Given a set of document identifiers, find the corresponding version strings (see below)
- Given a set of document identifiers and version strings, find information about the document, consisting of the document's data, access URL, and metadata

The Framework uses a unique document identifier to describe every document within the confines of a defined repository connection. This document identifier is used as a primary key to locate information about the document. When you set up a generic-database-type job, the database you are connecting to must have a similar concept. If you pick the wrong thing for a document identifier, at the very least you could find that the crawler runs very slowly.

Obtaining the list of document identifiers that represents the changes that occurred over the given time frame must return at least all such changes. It is acceptable (although not ideal) for the returned list to be bigger than that.

If you want your database connection to function in an incremental manner, you must also come up with the format of a "version string". This string is used by the Framework to determine if a document has changed. It must change whenever anything that might affect the document's indexing changes. (It is not a problem if it changes for other reasons, as long as it fulfills that principle criteria.)

The queries you provide get substituted before they are used by the connection. The example queries, which are present when the queries tab is first opened for a new job, show many of these substitutions in roughly the manner in which they are intended to be used. For example, "$(IDCOLUMN)" will substitute a column name expected by the connection to contain the document identifier into the query. The list of substitution strings are as follows:

String name	Meaning/use
IDCOLUMN	The name of an expected resultset column containing a document identifier
VERSIONCOLUMN	The name of an expected resultset column containing a version string
URLCOLUMN	The name of an expected resultset column containing a URL
DATACOLUMN	The name of an expected resultset column containing document data
STARTTIME	A query string value containing a start time in milliseconds since epoch
ENDTIME	A query string value containing an end time in milliseconds since epoch
IDLIST	A query string value containing a parenthesized list of document identifier values

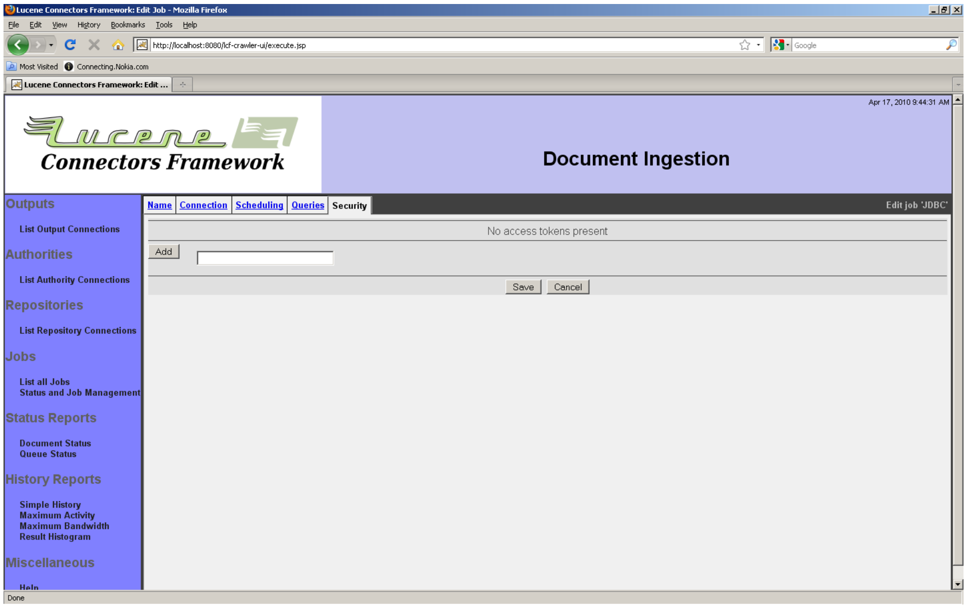
Use caution when constructing queries that include time-based components. "$(STARTTIME)" and "$(ENDTIME)" provide times in milliseconds since epoch. If the modified date field is not in this unit, the seeding query may not select the desired document identifiers. You should convert "$(STARTTIME)" and "$(ENDTIME)" to the appropriate timestamp unit for your system within your query.

The following table gives several sample query fragments that can be used to convert the helper strings "$(STARTTIME)" and "$(ENDTIME)" into other date and time types. The first column names the SQL database type that the following query phrase corresponds to, the second column names the output data type for the query phrase, and the third gives the query phrase itself using "$(STARTTIME)" as an example time in milliseconds since epoch. These query phrases are intended as guidlines for creating an appropriate query phrase in each language. Each query phrase is designed to work with the most current version of the database software available at the time of publishing for this document. If your modified date field is not of the type given in the second column, the query phrase may not provide an appropriate output for date comparisons.

Database Type	Date Type	Sample Query Phrase
Oracle	date	`TO_DATE ( '1970/01/01:00:00:00', 'yyyy/mm/dd:hh:mi:ss') + ROUND ($(STARTTIME)/86400000)`
Oracle	timestamp	`TO_TIMESTAMP('1970-01-01 00:00:00') + interval '$(STARTTIME)/1000' second`
Postgres SQL	timestamp	`date '1970-01-01' + interval '$(STARTTIME) milliseconds'`
MS SQL Server ($>$6.5)	datetime	`DATEADD(ms, $(STARTTIME), '19700101')`
Sybase (10+)	datetime	`DATEADD(ms, $(STARTTIME), '19700101')`

When you create a job based on a general database connection, the job's queries are initially populated with examples. These examples should give you a good idea of what columns your queries should return - in most cases, the only columns you need to return are the ones that appear in the example queries. However, for the file data query, you may also return columns that are not specified in the example. When you do this, the extra return column values will be passed to the index as metadata for the document. The metadata name used will be the corresponding column name of the resultset.

The "Security" tab simply allows you to add specific access tokens to all documents indexed with a general database job. In order for you to know what tokens to add, you must decide with what authority connection these documents will be secured, and understand the form of the access tokens used by that authority connection type. This is what the "Security" tab looks like:



No access tokens present

Enter a desired access token, and click the "Add" button. You may enter multiple access tokens.

4.7 IBM FileNet P8 Repository Connection

More here later

4.8 EMC Documentum Repository Connection

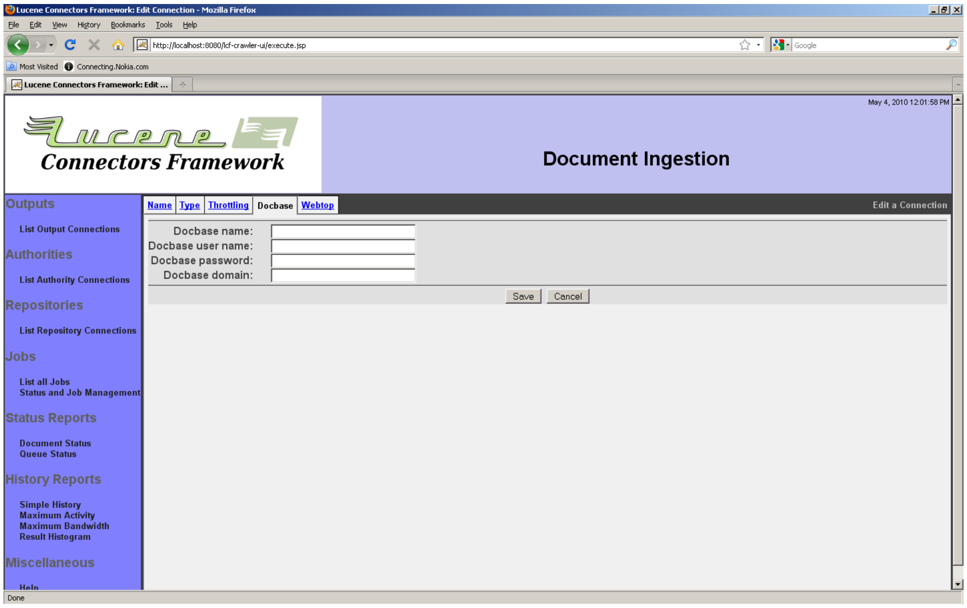
The EMC Documentum connection type allows you index content from a Documentum Content Server instance. A single connection allows you to reach all documents contained on a single Content Server instance. Multiple connections are therefore required to reach documents from multiple Content Server instances.

For each Content Server instance, the Documentum connection type allows you to index any Documentum content that is of type dm_document, or is derived from dm_document. Compound documents are handled as well, but only by mean of the component documents that make them up. No other Documentum construct can be indexed at this time.

Documents described by Documentum connections are typically secured by a Documentum authority. If you have not yet created a Documentum authority, but would like your documents to be secured, please follow the direction in the section titled "EMC Documentum Authority Connection".

A Documentum connection has the following special tabs: "Docbase", and "Webtop". The "Docbase" tab allows you to select a Content Server to connect to, and also to provide appropriate credentials. The "Webtop" tab describes the location of a Webtop server that will be used to display the documents from that Content Server, after they have been indexed.

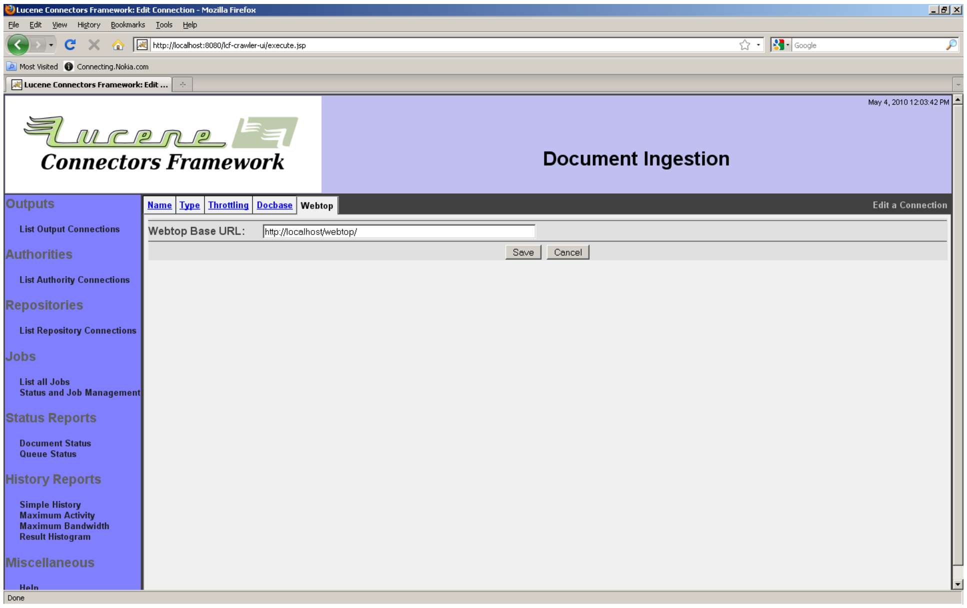
The "Docbase" tab looks like this:



Docbase name:
Docbase user name:
Docbase password:
Docbase domain:

Enter the Content Server Docbase instance name, and provide your credentials. You may leave the "Domain" field blank, if the Content Server instance does not have AD integration enabled.

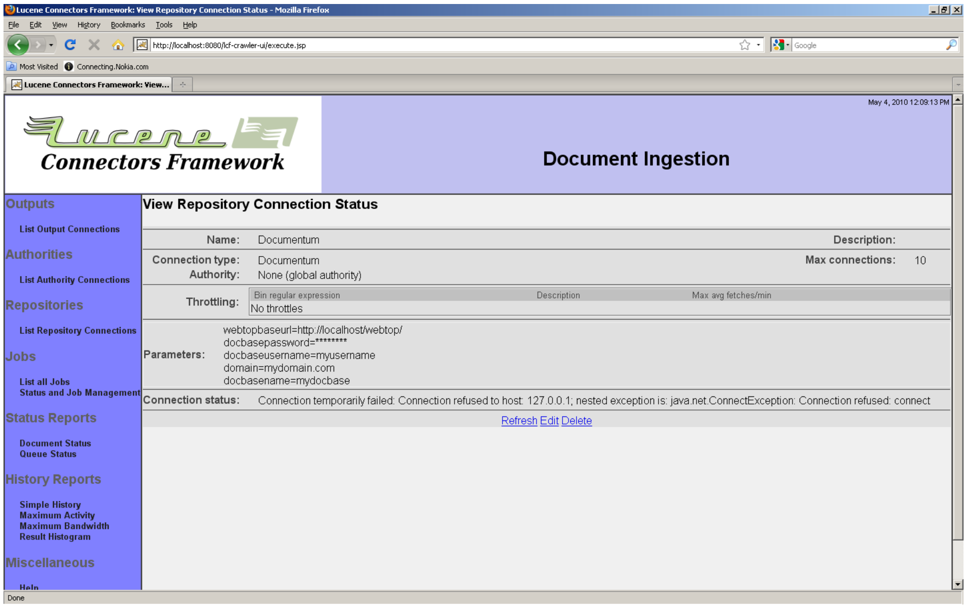
The "Webtop tab looks like this:


Lucene Connectors Framework: Edit Connection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/execute.jsp
Most Visited Connecting.Nokia.com
Lucene Connectors Framework: Edit ...
May 4, 2010 12:03:42 PM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Type
Throttling
Docbase
Webtop
Edit a Connection
Save
Cancel
Done

Webtop Base URL: http://localhost/webtop/

Enter the components of the base URL of the Webtop instance you want to use for serving the documents. Remember that this information will only be used to construct a URL to the document to allow user inspection; it will not be used for any crawling activities.

When you are done, click the "Save" button. When you do, a connection summary and status screen will be presented:



View Repository Connection Status

Name: Documentum
Description:
Connection type: Documentum
Max connections: 10
Authority: None (global authority)

Throttling:

Bin regular expression	Description	Max avg fetches/min
No throttles		

Parameters:

```
webtopbaseurl=http://localhost/webtop/
docbasepassword=********
docbaseusername=myusername
domain=mydomain.com
docbasename=mydocbase
```

Connection status: Connection temporarily failed: Connection refused to host: 127.0.0.1; nested exception is: java.net.ConnectException: Connection refused: connect

Pay careful attention to the status, and be prepared to correct any problems that are displayed.

A job created to use a Documentum connection has the following additional tabs associated with it: "Paths", "Document Types", "Content Types", "Security", and "Path Metadata".

The "Paths" tab allows you to construct the paths within Documentum that you want to scan for content. If no paths are selected, all content will be considered eligible.

The "Document Types" tab allows you to select what document types you want to index. Only document types that are derived from dm_document, which are flagged by the system administrator as being "indexable", will be presented for your selection. On this tab also, for each document type you index, you may choose included specific metadata for documents of that type, or you can check the "All metadata" checkbox to include all metadata associated with documents of that type.

The "Content Types" tab allows you to select which Documentum mime-types are to be included in the document set. Check the types you want to include, and uncheck the types you want to exclude.

The "Security" tab allows you to disable or enable Documentum security for the documents described by this job. You can turn off native Documentum security by clicking the "Disable" radio button. If you do this, you may also enter your own access tokens, which will be applied to all documents described by the job. The form of the access tokens you enter will depend on the governing authority connection type. Click the "Add" button to add each access token.

The "Path Metadata" tab allows you to send each document's path information as metadata to the index. To enable this feature, enter the name of the metadata attribute you want this information to be sent into the "Path attribute name" field. Then, add the rules you want to the list of rules. Each rule has a match expression, which is a regular expression where parentheses ("(" and ")") mark sections you are interested in. These sections are called "groups" in regular expression parlance. The replace string consists of constant text plus substitutions of the groups from the match, perhaps modified. For example, "$(1)" refers to the first group within the match, while "$(1l)" refers to the first match group mapped to lower case. Similarly, "$(1u)" refers to the same characters, but mapped to upper case.

For example, suppose you had a rule which had ".*/(.*)/(.*)/.*" as a match expression, and "$(1) $(2)" as the replace string. If presented with the path `Project/Folder_1/Folder_2/Filename`, it would output the string `Folder_1 Folder_2`.

If more than one rule is present, the rules are all executed in sequence. That is, the output of the first rule is modified by the second rule, etc.

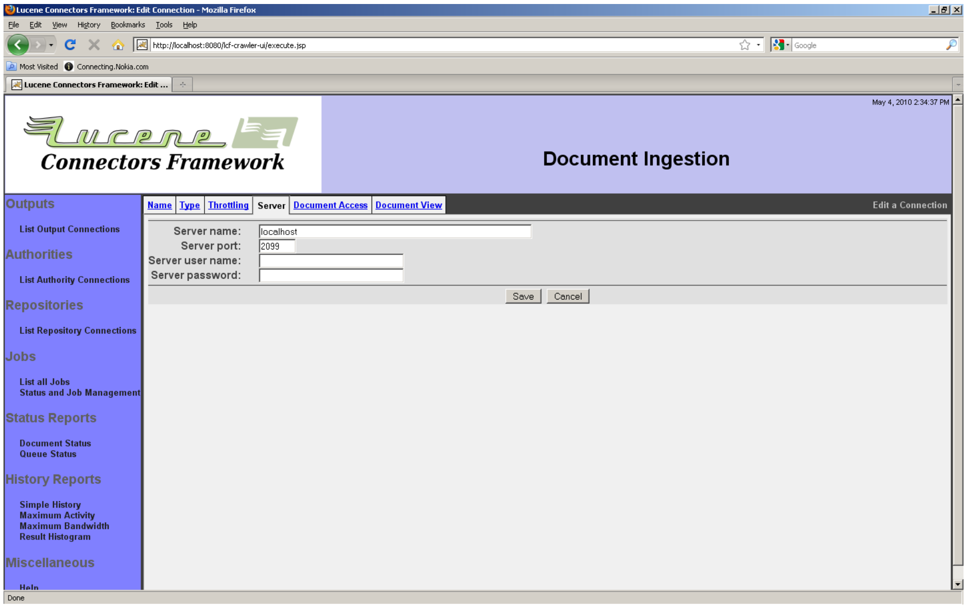
4.9 OpenText LiveLink Repository Connection

The OpenText LiveLink connection type allows you to index content from LiveLink repositories. LiveLink has a rich variety of different document types and metadata, which include basic documents, as well as compound documents, folders, workspaces, and projects. A LiveLink connection is able to discover documents contained within all of these constructs.

Documents described by LiveLink connections are typically secured by a LiveLink authority. If you have not yet created a LiveLink authority, but would like your documents to be secured, please follow the direction in the section titled "OpenText LiveLink Authority Connection".

A LiveLink connection has the following special tabs: "Server", "Document Access", and "Document View". The "Server" tab allows you to select a LiveLink server to connect to, and also to provide appropriate credentials. The "Document Access" tab describes the location of the LiveLink web interface, relative to the server, that will be used to fetch document content from LiveLink. The "Document View" tab affects how URLs to the fetched documents are constructed, for viewing results of searches.

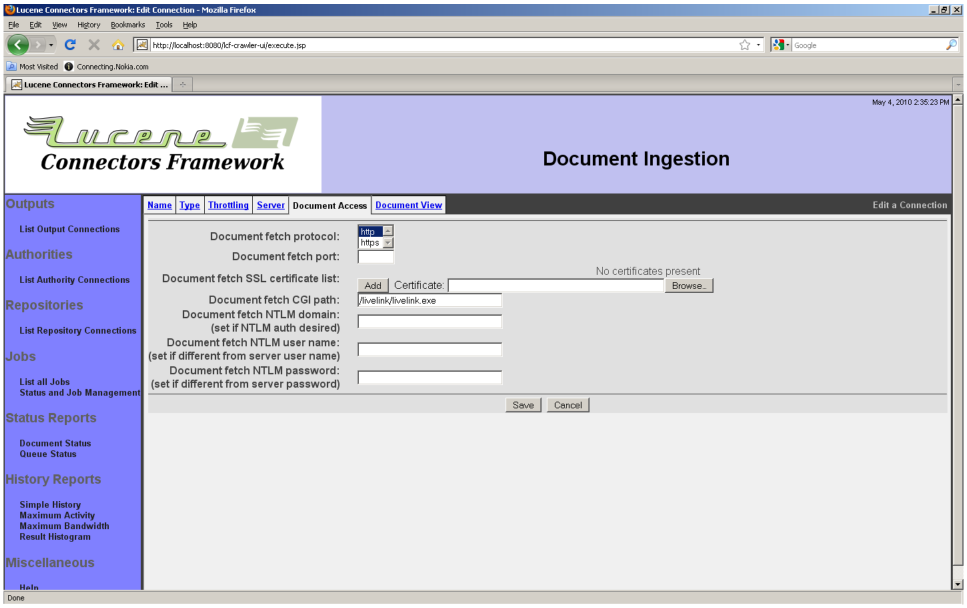
The "Server" tab looks like this:



Server name: localhost
Server port: 2099
Server user name:
Server password:

Enter the LiveLink server name, port, and credentials.

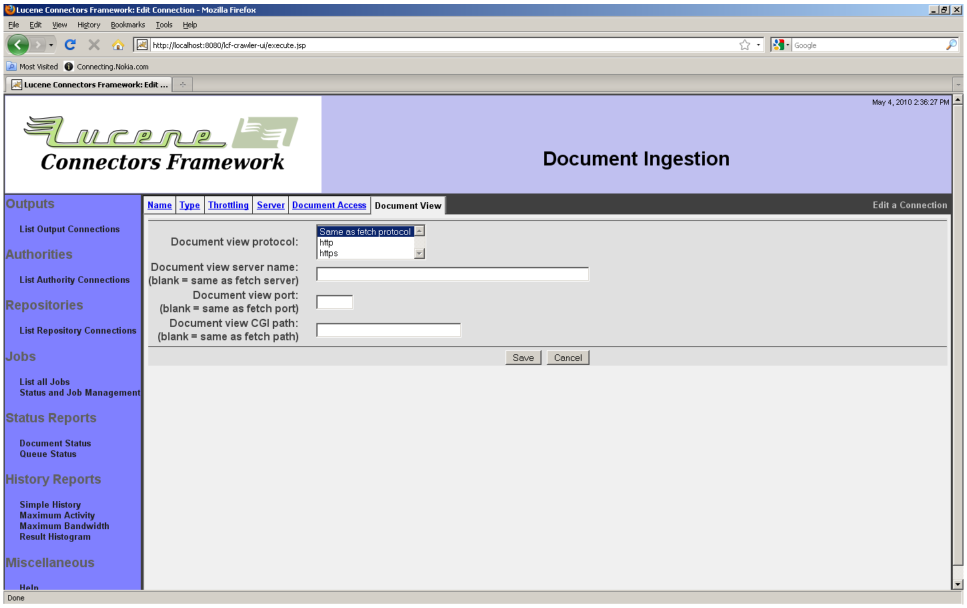
The "Document Access" tab looks like this:



Document fetch protocol: http / https
Document fetch port:
Document fetch SSL certificate list: No certificates present
Certificate:
Document fetch CGI path: /livelink/livelink.exe
Document fetch NTLM domain:
(set if NTLM auth desired)
Document fetch NTLM user name:
(set if different from server user name)
Document fetch NTLM password:
(set if different from server password)

The server name is presumed to be the same as is on the "Server" tab. Select the desired protocol. If your LiveLink server is using a non-standard HTTP port for the specified protocol, enter the port number. If your LiveLink server is using NTLM authentication, enter an AD user name, password, and domain. If your LiveLink is using HTTPS, browse locally for the appropriate certificate authority certificate, and click "Add" to upload that certificate to the connection's trust store. (You may also use the server's certificate, but that is less resilient because the server's certificate may be changed periodically.)

The "Document View" tab looks like this:



Lucene Connectors Framework: Edit Connection - Mozilla Firefox

http://localhost:8080/lcf-crawler-ui/execute.jsp

May 4, 2010 2:36:27 PM

Document Ingestion

Document view protocol: Same as fetch protocol / http / https

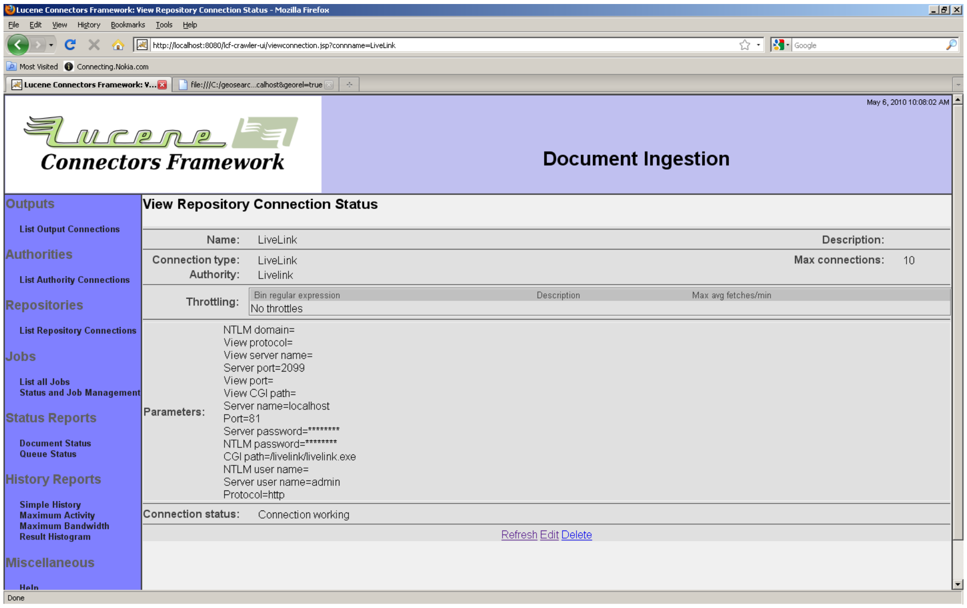
Document view server name: (blank = same as fetch server)

Document view port: (blank = same as fetch port)

Document view CGI path: (blank = same as fetch path)

If you want each document's view URL to be the same as its access URL, you can leave this tab unchanged. If you want to direct users to a different CGI path when they view search results, you can specify that here.

When you are done, click the "Save" button. You will see a summary screen that looks something like this:


Lucene Connectors Framework: View Repository Connection Status - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/lcf-crawler-ui/viewconnection.jsp?connname=LiveLink
Google
Most Visited
Connecting.Nokia.com
Lucene Connectors Framework: V...
file:///C:/geosearc...calhost&georel=true
May 6, 2010 10:08:02 AM
Lucene Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Done

View Repository Connection Status

Name: LiveLink
Description:

Connection type: LiveLink
Authority: Livelink
Max connections: 10

Throttling:

Bin regular expression	Description	Max avg fetches/min
No throttles		

Parameters:

```
NTLM domain=
View protocol=
View server name=
Server port=2099
View port=
View CGI path=
Server name=localhost
Port=81
Server password=********
NTLM password=********
CGI path=/livelink/livelink.exe
NTLM user name=
Server user name=admin
Protocol=http
```

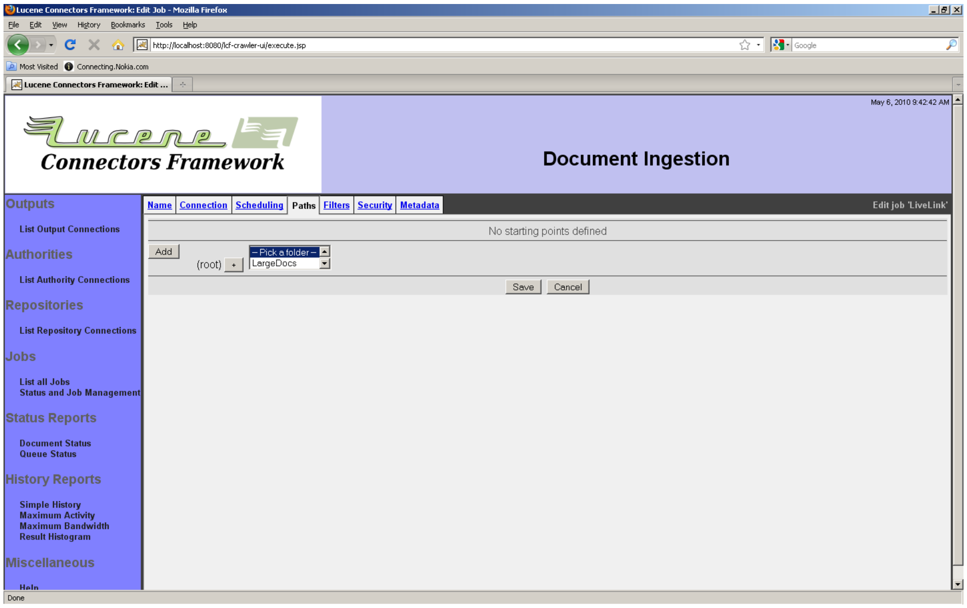
Connection status: Connection working

Refresh Edit Delete

Make note of and correct any reported connection errors. In this example, the connection has been correctly set up, so the connection status is "Connection working".

A job created to use a LiveLink connection has the following additional tabs associated with it: "Paths", "Filters", "Security", and "Metadata".

The "Paths" tab allows you to manage a list of LiveLink paths that act as starting points for indexing content:



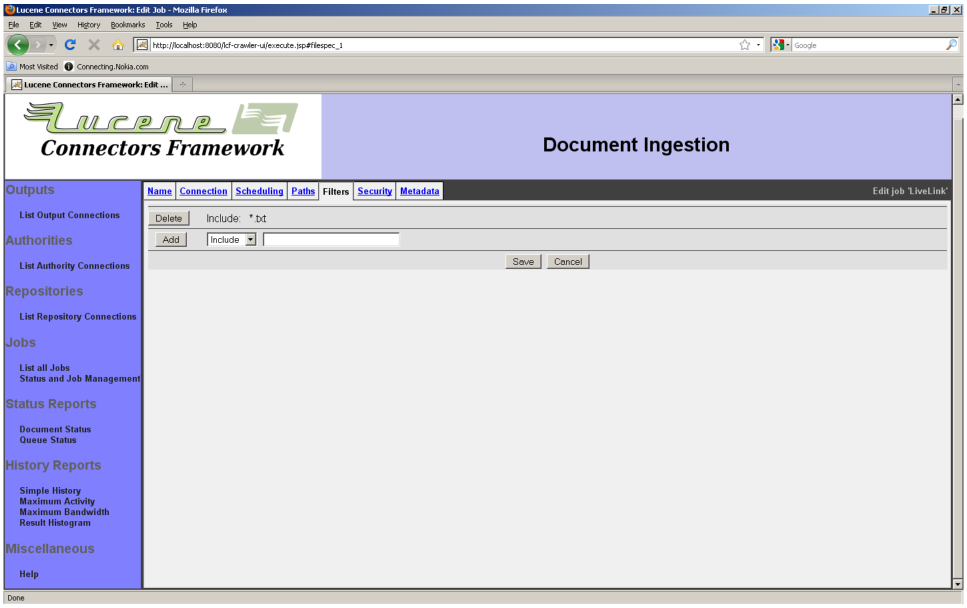
Edit job 'LiveLink'

No starting points defined

(root) -- Pick a folder -- LargeDocs

Build each path by selecting from the available dropdown, and clicking the "+" button. When your path is complete, click the "Add" button to add the path to the list of starting points.

The "Filters" tab controls the criteria the LiveLink job will use to include or exclude content. The filters are basically a list of rules. Each rule has a document match field, and a matching action ("Include" or "Exclude"). When a LiveLink connection encounters a document, it evaluates the rules from top to bottom. If the rule matches, then it will be included or excluded from the job's document set depending on what you have selected for the matching action. A rule's match field specifies a character match, where "*" will match any number of characters, and "?" will match any single character.


Lucene Connectors Framework: Edit Job - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Most Visited Connecting.Nokia.com
Lucene Connectors Framework: Edit ...
Lucene
Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name Connection Scheduling Paths Filters Security Metadata
Delete
Add
Save Cancel
Done

http://localhost:8080/lcf-crawler-ui/execute.jsp#filespec_1

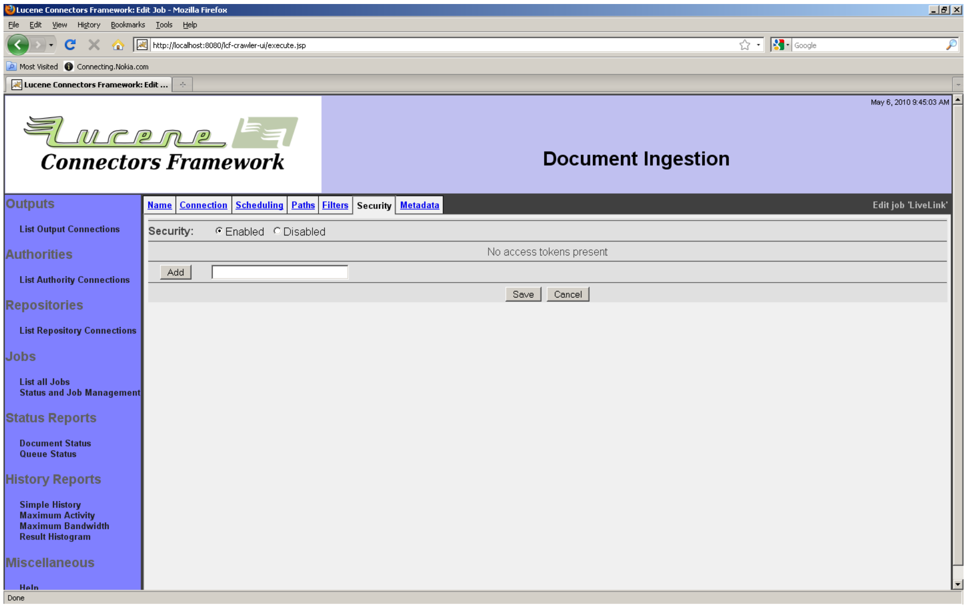
Edit job 'LiveLink'

Include: *.txt

Include

Enter the match field value, select the match action, and click the "Add" button to add to the list of filters.

The "Security" tab allows you to disable (or enable) LiveLink security for the documents associated with this job:


Lucene Connectors Framework: Edit Job - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Most Visited Connecting.Nokia.com
Lucene Connectors Framework: Edit ...
Lucene
Connectors Framework
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name Connection Scheduling Paths Filters Security Metadata
Add
Save Cancel
Done

http://localhost:8080/lcf-crawler-ui/execute.jsp

May 6, 2010 9:45:03 AM

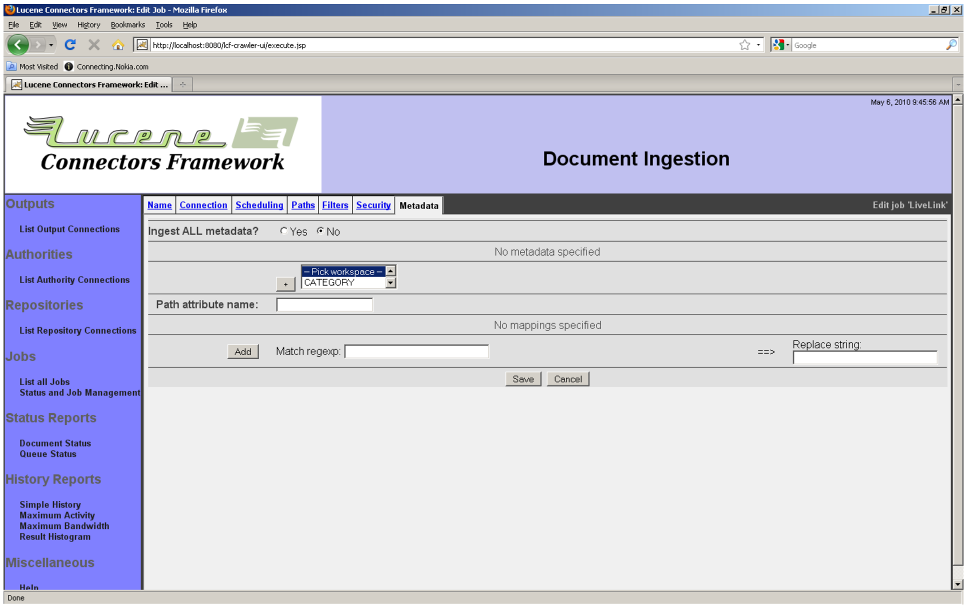
Edit job 'LiveLink'

Security: ● Enabled ○ Disabled

No access tokens present

If you disable security, you can add your own access tokens to all jobs in the document set as they are indexed. The format of the access tokens you would enter depends on the governing authority associated with the job's repository connection. Enter a token and click the "Add" button to add it to the list.

The "Metadata" tab allows you to select what specific metadata values from LiveLink you want to pass to the index:



Ingest ALL metadata? ○ Yes ● No

No metadata specified

– Pick workspace –
CATEGORY

Path attribute name:

No mappings specified

Match regexp: ==> Replace string:

If you want to pass all available LiveLink metadata to the index, then click the "All metadata" radio button. Otherwise, you need to build LiveLink metadata paths and add them to the metadata list. Select the next metadata path segment, and click the appropriate "+" button to add it to the path. You may add folder information, or a metadata category, at any point.

Once you have drilled down to a metadata category, you can select the metadata attributes to include, or check the "All attributes in this category" checkbox. When you are done, click the "Add" button to add the metadata attributes that you want to include in the index.

You can also use the "Metadata" tab to have the connection send path data along with each document, as a piece of document metadata. To enable this feature, enter the name of the metadata attribute you want this information to be sent into the "Path attribute name" field. Then, add the rules you want to the list of rules. Each rule has a match expression, which is a regular expression where parentheses ("(" and ")") mark sections you are interested in. These sections are called "groups" in regular expression parlance. The replace string consists of constant text plus substitutions of the groups from the match, perhaps modified. For example, "$(1)" refers to the first group within the match, while "$(1l)" refers to the first match group mapped to lower case. Similarly, "$(1u)" refers to the same characters, but mapped to upper case.

For example, suppose you had a rule which had ".*/(.*)/(.*)/.*" as a match expression, and "$(1) $(2)" as the replace string. If presented with the path `Project/Folder_1/Folder_2/Filename`, it would output the string `Folder_1 Folder_2`.

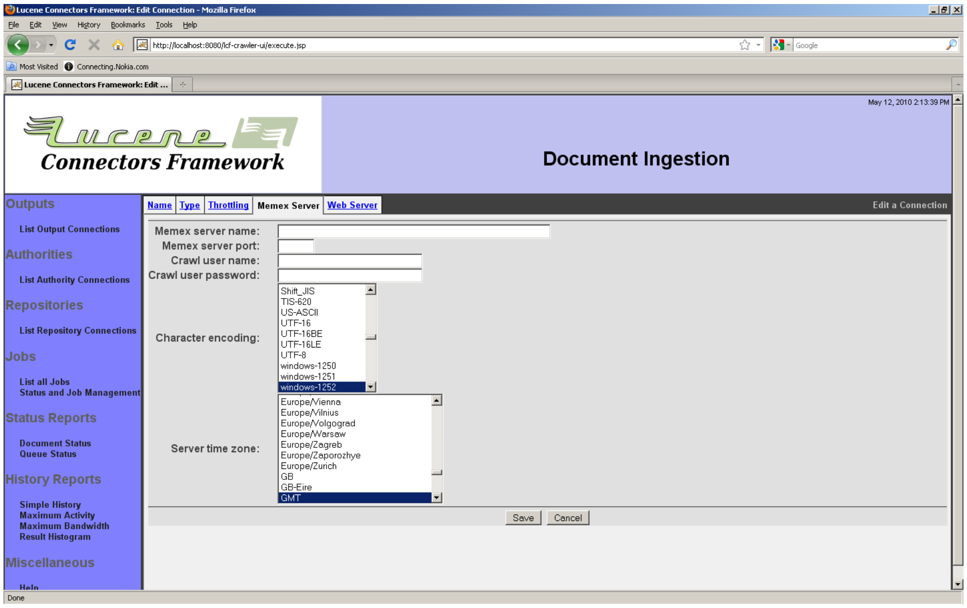
If more than one rule is present, the rules are all executed in sequence. That is, the output of the first rule is modified by the second rule, etc.

4.10 Memex Patriarch Repository Connection

A Memex Patriarch connection allows you to index documents from a Memex server.

Documents described by Memex connections are typically secured by a Memex authority. If you have not yet created a Memex authority, but would like your documents to be secured, please follow the direction in the section titled "Memex Patriarch Authority Connection".

A Memex connection has the following special tabs on the repository connection editing screen: the "Memex Server" tab, and the "Web Server" tab. The "Memex Server" tab looks like this:



Memex server name:

Memex server port:

Crawl user name:

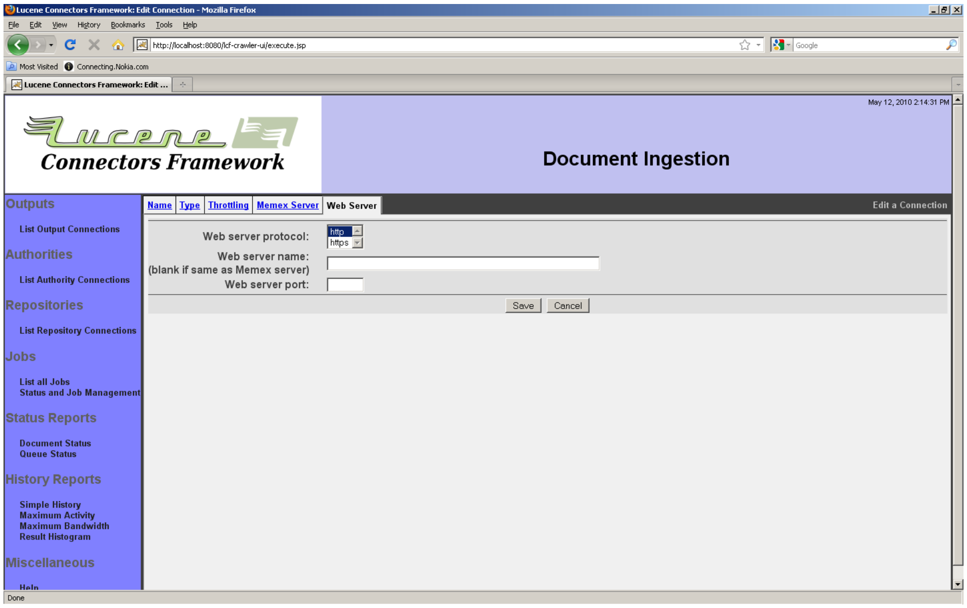
Crawl user password:

Character encoding: Shift_JIS, TIS-620, US-ASCII, UTF-16, UTF-16BE, UTF-16LE, UTF-8, windows-1250, windows-1251, windows-1252 (selected)

Server time zone: Europe/Vienna, Europe/Vilnius, Europe/Volgograd, Europe/Warsaw, Europe/Zagreb, Europe/Zaporozhye, Europe/Zurich, GB, GB-Eire, GMT (selected)

You must supply the name of your Memex server, and the connection port, along with the Memex credentials for a user that has sufficient permissions to retrieve Memex documents. You must also select the Memex server's character encoding, and timezone. If you do not know the encoding or timezone, consult your Memex system administrator.

The "Web Server" tab looks like this:



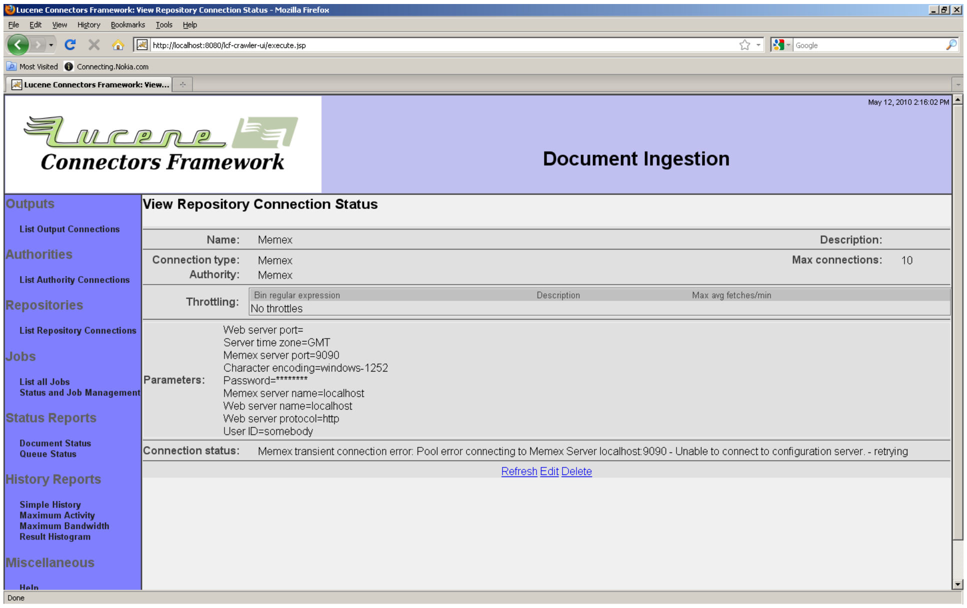
Web server protocol: http / https

Web server name:
(blank if same as Memex server)

Web server port:

Here you must provide information that allows a Memex connection to construct a unique URL for each of your Memex documents. Select a protocol, and fill in the server name and port.

When you are done, click the "Save" button. You should see a status page, something like this:



View Repository Connection Status

Name: Memex
Description:
Connection type: Memex
Max connections: 10
Authority: Memex
Throttling:

Bin regular expression	Description	Max avg fetches/min
No throttles		

Parameters:

```
Web server port=
Server time zone=GMT
Memex server port=9090
Character encoding=windows-1252
Password=********
Memex server name=localhost
Web server name=localhost
Web server protocol=http
User ID=somebody
```

Connection status: Memex transient connection error: Pool error connecting to Memex Server localhost:9090 - Unable to connect to configuration server. - retrying

Jobs based on Memex connections have the following special tabs: "Record Criteria", "Entities", and "Security".

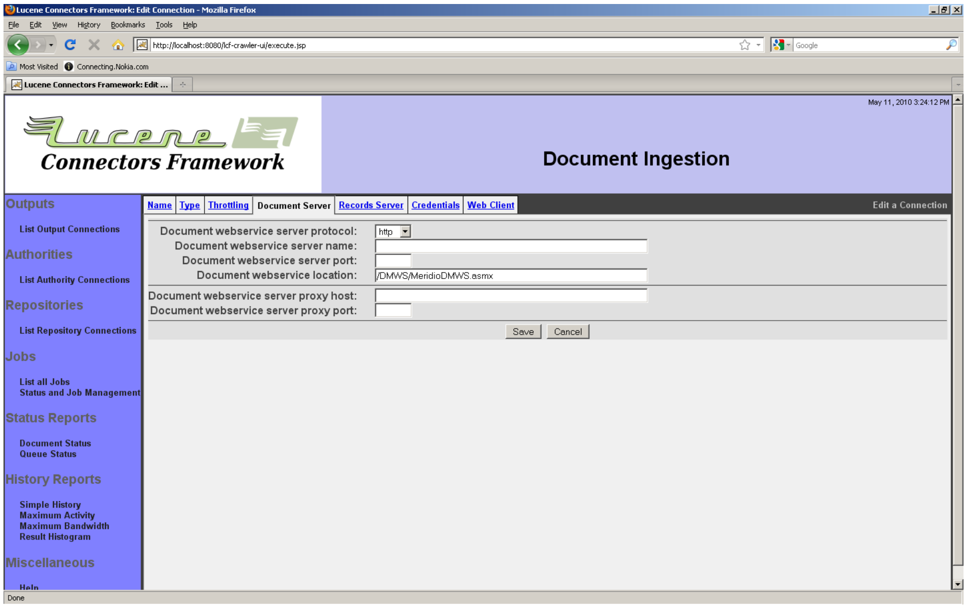
More here later

4.11 Autonomy Meridio Repository Connection

An Autonomy Meridio connection allows you to index documents from a set of Meridio servers. Meridio's architecture allows you to separate services on multiple machines - e.g. the document services can run on one machine, and the records services can run on another. A Meridio connection type correspondingly is configured to describe each Meridio service independently.

Documents described by Meridio connections are typically secured by a Meridio authority. If you have not yet created a Meridio authority, but would like your documents to be secured, please follow the direction in the section titled "Autonomy Meridio Authority Connection".

A Meridio connection has the following special tabs on the repository connection editing screen: the "Document Server" tab, the "Records Server" tab, the "Web Client" tab, and the "Credentials" tab. The "Document Server" tab looks like this:

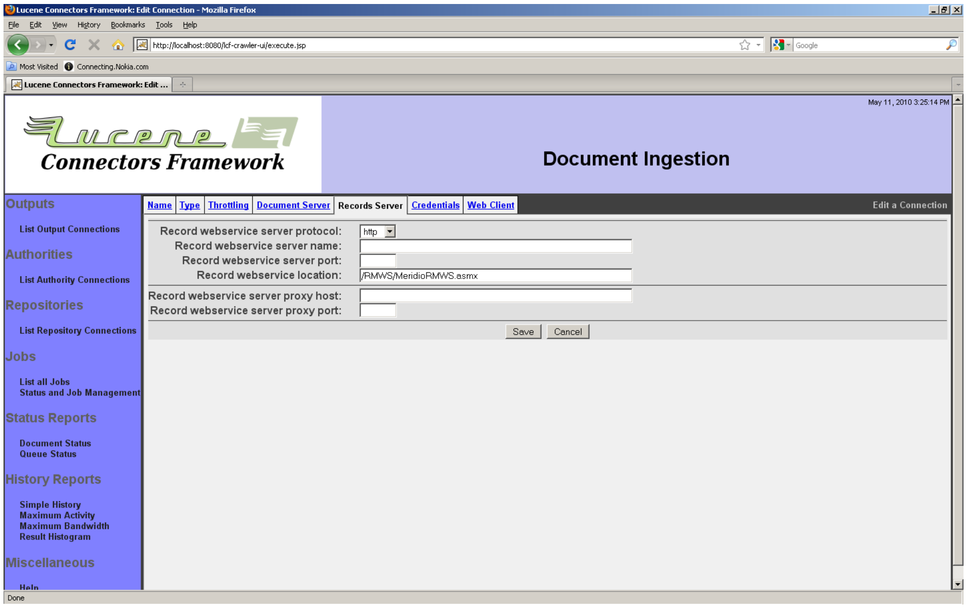


Document webservice server protocol:	http
Document webservice server name:	
Document webservice server port:	
Document webservice location:	/DMWS/MeridioDMWS.asmx
Document webservice server proxy host:	
Document webservice server proxy port:	

Select the correct protocol, and enter the correct server name, port, and location to reference the Meridio document server services. If a proxy is involved, enter the proxy host and port. Authenticated proxies are not supported by this connection type at this time.

Note that, in the Meridio system, while it is possible that different services run on different servers, this is not typically the case. The connection type, on the other hand, makes no assumptions, and permits the most general configuration.

The "Records Server" tab looks like this:

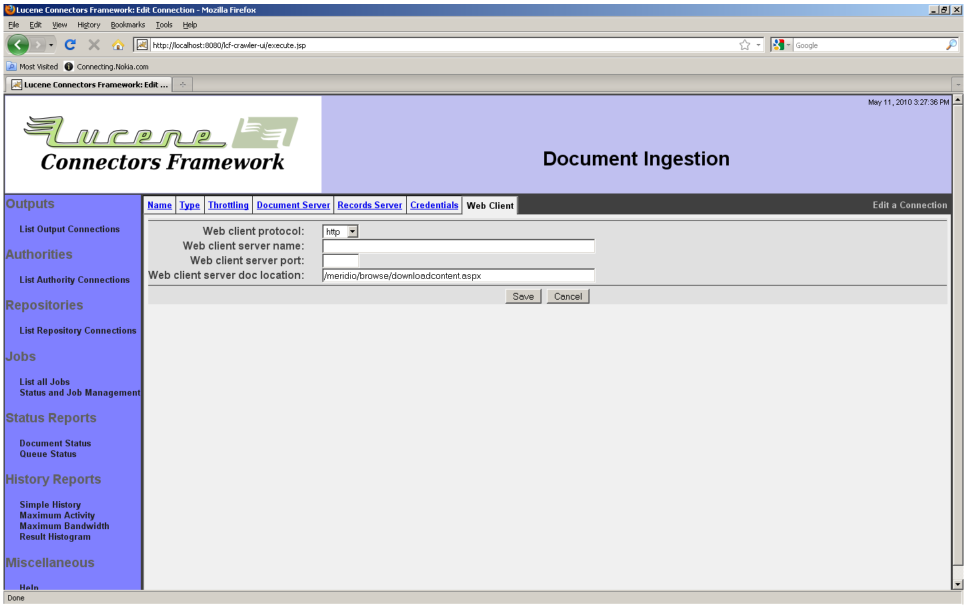


Record webservice server protocol: http
Record webservice server name:
Record webservice server port:
Record webservice location: /RMWS/MeridioRMWS.asmx
Record webservice server proxy host:
Record webservice server proxy port:

Select the correct protocol, and enter the correct server name, port, and location to reference the Meridio records server services. If a proxy is involved, enter the proxy host and port. Authenticated proxies are not supported by this connection type at this time.

Note that, in the Meridio system, while it is possible that different services run on different servers, this is not typically the case. The connection type, on the other hand, makes no assumptions, and permits the most general configuration.

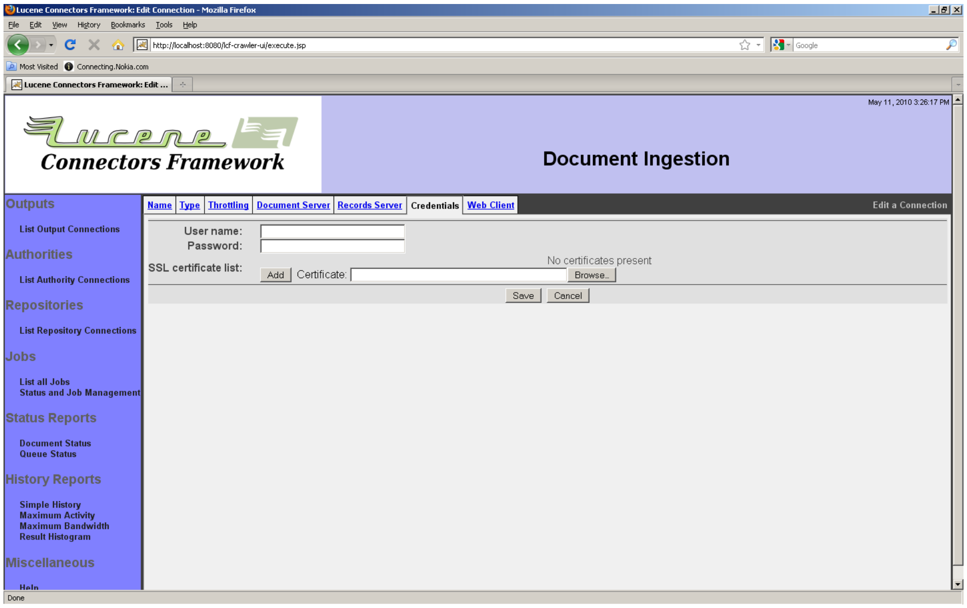
The "Web Client" tab looks like this:



Web client protocol: http
Web client server name:
Web client server port:
Web client server doc location: /meridio/browse/downloadcontent.aspx

The purpose of the Meridio Connection web client tab is to allow the connection to build a useful URL for each document it indexes. Select the correct protocol, and enter the correct server name, port, and location to reference the Meridio web client service. No proxy information is required, as no documents will be fetched from this service.

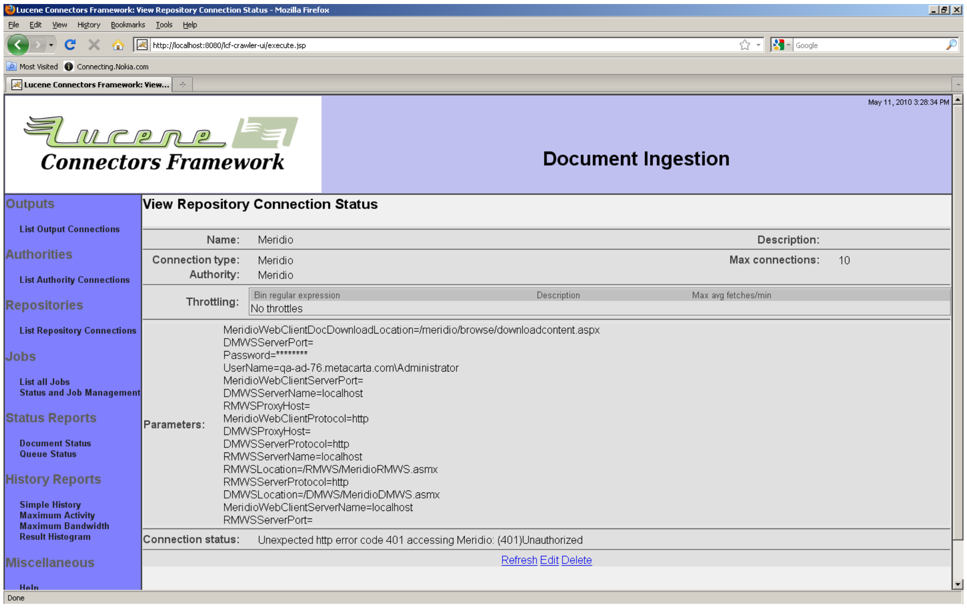
The "Credentials" tab looks like this:



User name:
Password:
SSL certificate list: No certificates present
Certificate:

Enter the Meridio server credentials needed to access the Meridio system.

When you are done, click the "Save" button to save the connection. You will see a summary screen, looking something like this:



View Repository Connection Status

Name: Meridio
Description:

Connection type: Meridio
Max connections: 10
Authority: Meridio

Throttling:

Bin regular expression	Description	Max avg fetches/min
No throttles		

Parameters:

```
MeridioWebClientDocDownloadLocation=/meridio/browse/downloadcontent.aspx
DMWSServerPort=
Password=********
UserName=qa-ad-76.metacarta.com\Administrator
MeridioWebClientServerPort=
DMWSServerName=localhost
RMWSProxyHost=
MeridioWebClientProtocol=http
DMWSProxyHost=
DMWSServerProtocol=http
RMWSServerName=localhost
RMWSLocation=/RMWS/MeridioRMWS.asmx
RMWSServerProtocol=http
DMWSLocation=/DMWS/MeridioDMWS.asmx
MeridioWebClientServerName=localhost
RMWSServerPort=
```

Connection status: Unexpected http error code 401 accessing Meridio: (401)Unauthorized

Note that in this example, the Meridio connection is not actually correctly configured, which is leading to an error status message instead of "Connection working".

Since Meridio uses Windows IIS for authentication, there are many ways in which the configuration of either IIS or the Windows domain under which Meridio runs can affect the correct functioning of the Meridio connection. It is beyond the scope of this manual to describe the kinds of analysis and debugging techniques that might be required to diagnose connection and authentication problems. If you have trouble, you will almost certainly need to involve your Meridio IT personnel. Debugging tools may include (but are not limited to):

- Windows security event logs
- ManifoldCF logs (see below)
- Packet captures (using a tool such as WireShark)

If you need specific ManifoldCF logging information, contact your system integrator.

Jobs based on Meridio connections have the following special tabs: "Search Paths", "Content Types", "Categories", "Data Types", "Security", and "Metadata".

More here later

4.12 Microsoft SharePoint Repository Connection

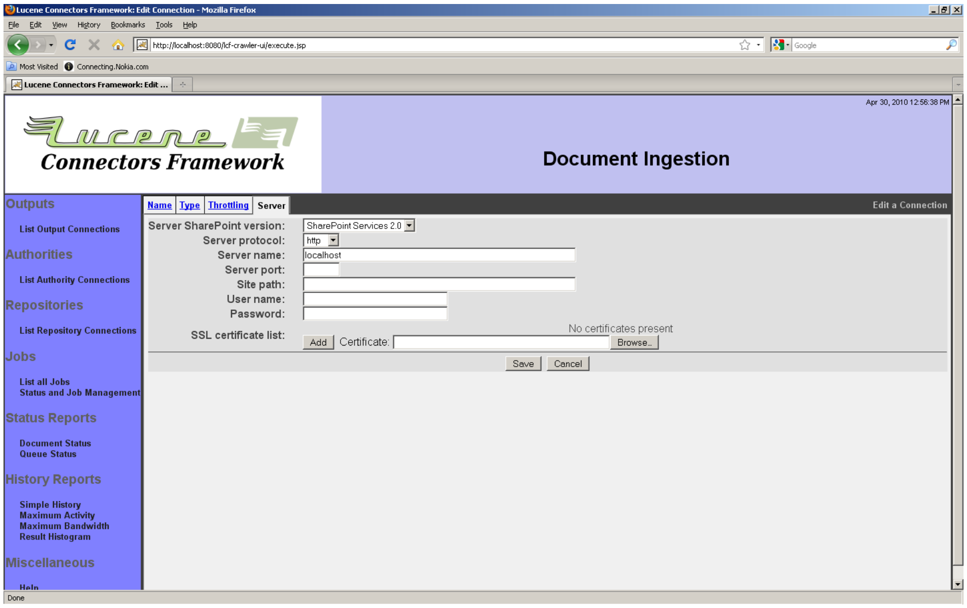
The Microsoft SharePoint connection type allows you to index documents from a Microsoft SharePoint site. Bear in mind that a single SharePoint installation actually represents a set of sites. Some sites in SharePoint are directly related to others (e.g. they are subsites), while some sites operate relatively independently of one another.

The SharePoint connection type is designed so that one SharePoint repository connection can access all SharePoint sites from a specific root site though its explicit subsites. It is the case that it is desirable in some very large SharePoint installations to access all SharePoint sites using a single connection. But the ManifoldCF SharePoint connection type today does not support that model as of yet. If this functionality is important for you, contact your system integrator.

SharePoint uses a web URL model for addressing sites, subsites, libraries, and files. The best way to figure out how to set up a SharePoint connection type is therefore to start with your web browser, and visit the root of the site you wish to crawl. Then, record the URL you see in your browser.

Documents described by SharePoint connections are typically secured by an Active Directory authority. If you have not yet created your Active Directory authority, but would like your documents to be secured, please follow the direction in the section titled "Active Directory Authority Connection".

A SharePoint connection has one special tab on the repository connection editing screen: the "Server" tab, which looks like this:



Server SharePoint version: SharePoint Services 2.0
Server protocol: http
Server name: localhost
Server port:
Site path:
User name:
Password:
SSL certificate list: No certificates present
Certificate:

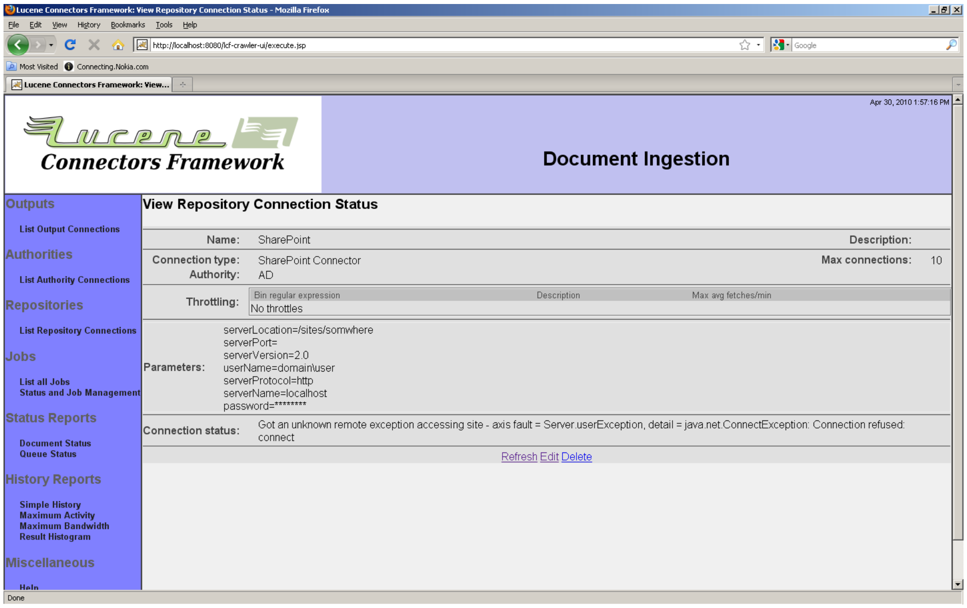
Select your SharePoint server version from the pulldown. If you do not select the correct server version, your documents may either be indexed with insufficient security protection, or you may not be able to index any documents. Check with your SharePoint system administrator if you are not sure what to select.

Select the server protocol, and enter the server name and port, based on what you recorded from the URL for your SharePoint site. For the "Site path" field, type in the portion of the root site URL that includes everything after the server and port, except for the final "aspx" file. For example, if the SharePoint URL is "http://myserver:81/sites/somewhere/index.asp", the site path would be "/sites/somewhere".

The SharePoint credentials are, of course, what you used to log into your root site. The SharePoint connection type always requires the user name to be in the form "domain\user".

If your SharePoint server is using SSL, you will need to supply enough certificates for the connection's trust store so that the SharePoint server's SSL server certificate can be validated. This typically consists of either the server certificate, or the certificate from the authority that signed the server certificate. Browse to the local file containing the certificate, and click the "Add" button.

After you click the "Save" button, you will see a connection summary screen, which might look something like this:



View Repository Connection Status

Name: SharePoint
Description:
Connection type: SharePoint Connector
Max connections: 10
Authority: AD

Throttling:

Bin regular expression	Description	Max avg fetches/min
No throttles		

Parameters:

```
serverLocation=/sites/somwhere
serverPort=
serverVersion=2.0
userName=domain\user
serverProtocol=http
serverName=localhost
password=********
```

Connection status: Got an unknown remote exception accessing site - axis fault = Server.userException, detail = java.net.ConnectException: Connection refused: connect

Note that in this example, the SharePoint connection is not actually referencing a SharePoint instance, which is leading to an error status message instead of "Connection working".

Since SharePoint uses Windows IIS for authentication, there are many ways in which the configuration of either IIS or the Windows domain under which SharePoint runs can affect the correct functioning of the SharePoint connection. It is beyond the scope of this manual to describe the kinds of analysis and debugging techniques that might be required to diagnose connection and authentication problems. If you have trouble, you will almost certainly need to involve your SharePoint IT personnel. Debugging tools may include (but are not limited to):

- Windows security event logs
- ManifoldCF logs (see below)
- Packet captures (using a tool such as WireShark)

If you need specific ManifoldCF logging information, contact your system integrator.

When you configure a job to use a repository connection of the generic database type, several additional tabs are presented. These are, in order, "Paths", "Security", and "Metadata".

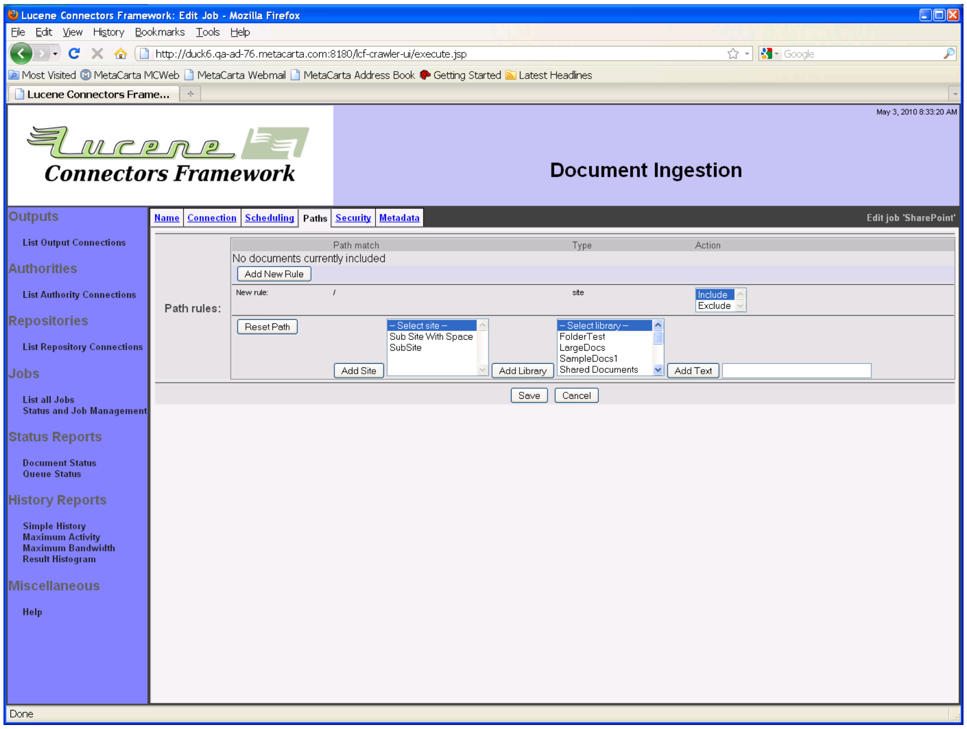
The "Paths" tab allows you to build a list of rules describing the SharePoint content that you want to include in your job. When the

SharePoint connection type encounters a subsite, library, or file, it looks through this list of rules to determine whether to include the subsite, library, or file. The first matching rule will determine what will be done.

Each rule consists of a path, a rule type, and an action. The actions are "Include" and "Exclude". The rule type tells the connection what kind of SharePoint entity it is allowed to exactly match. For example, a "File" rule will only exactly match SharePoint paths that represent files - it cannot exactly match sites or libraries. The path itself is just a sequence of characters, where the "*" character has the special meaning of being able to match any number of any kind of characters, and the "?" character matches exactly one character of any kind.

The rule matcher extends strict, exact matching by introducing a concept of implicit inclusion rules. If your rule action is "Include", and you specify (say) a "File" rule, the matcher presumes implicit inclusion rules for the corresponding site and library. So, if you create an "Include File" rule that matches (for example) "/MySite/MyLibrary/MyFile", there is an implied "Site Include" rule for "/MySite", and an implied "Library Include" rule for "/MySite/MyLibrary". Similarly, if you create a "Library Include" rule, there is an implied "Site Include" rule that corresponds to it. Note that these shortcuts only applies to "Include" rules - there are no corresponding implied "Exclude" rules.

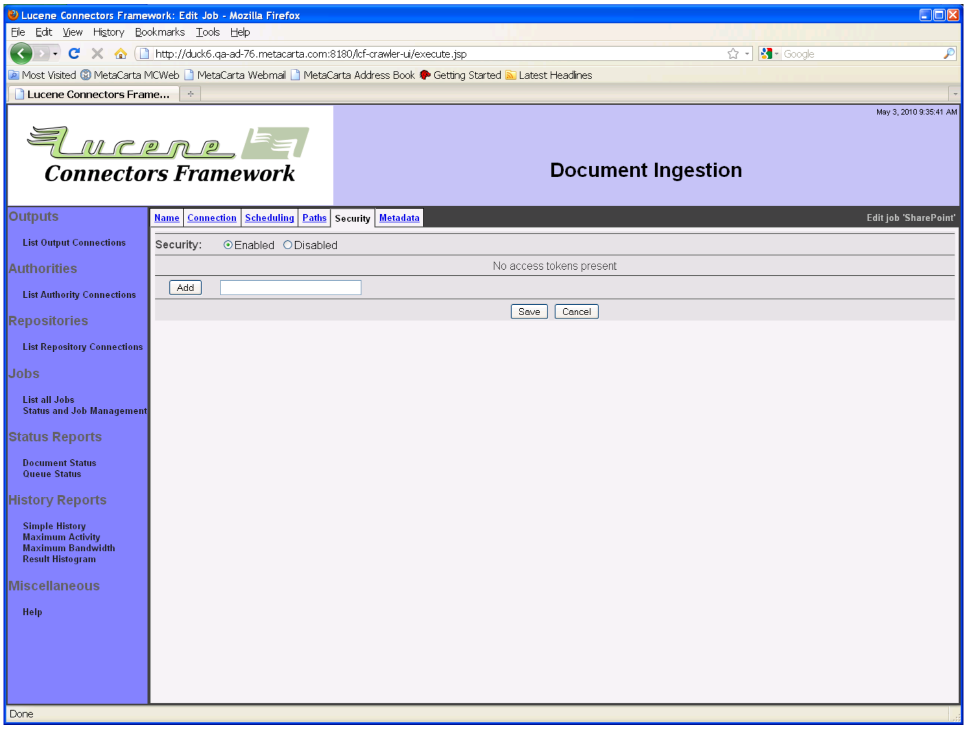
The "Paths" tab allows you to build these rules one at a time, and add them either to the bottom of the list, or insert them into the list of rules at any point. Either way, you construct the rule you want to append or insert by first constructing the path, from left to right, using your choice of text and context-dependent pulldowns with existing server path information listed. This is what the tab may look like for you. Bear in mind that if you are using a connection that does not display the status, "Connection working", you may not see the selections you should in these pulldowns:



No documents currently included

To build a rule, first build the rule's matching path. Make an appropriate selection or enter desired text, then click either the "Add Site", "Add Library", or "Add Text" button, depending on your choice. Repeat this process until the path is what you want it to be. At this point, if the SharePoint connection does not know what kind of entity your path describes, you will need to select the SharePoint entity type that you want the rule to match also. Select whether this is an include or exclude rule. Then, click the "Add New Rule" button, to add your newly-constructed rule at the end of the list.

The "Security" tab allows you to specify whether SharePoint's security model should be applied to this set of documents, or not. You also have the option of applying some specified set of access tokens to the documents described by the job. The tab looks like this:


Lucene Connectors Framework: Edit Job - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://duck6.qa-ad-76.metacarta.com:8180/lcf-crawler-ui/execute.jsp
Most Visited MetaCarta MCWeb MetaCarta Webmail MetaCarta Address Book Getting Started Latest Headlines
Lucene Connectors Frame...
Lucene Connectors Framework
May 3, 2010 9:35:41 AM
Document Ingestion
Outputs
List Output Connections
Authorities
List Authority Connections
Repositories
List Repository Connections
Jobs
List all Jobs
Status and Job Management
Status Reports
Document Status
Queue Status
History Reports
Simple History
Maximum Activity
Maximum Bandwidth
Result Histogram
Miscellaneous
Help
Name
Connection
Scheduling
Paths
Security
Metadata
Add
Save
Cancel
Done

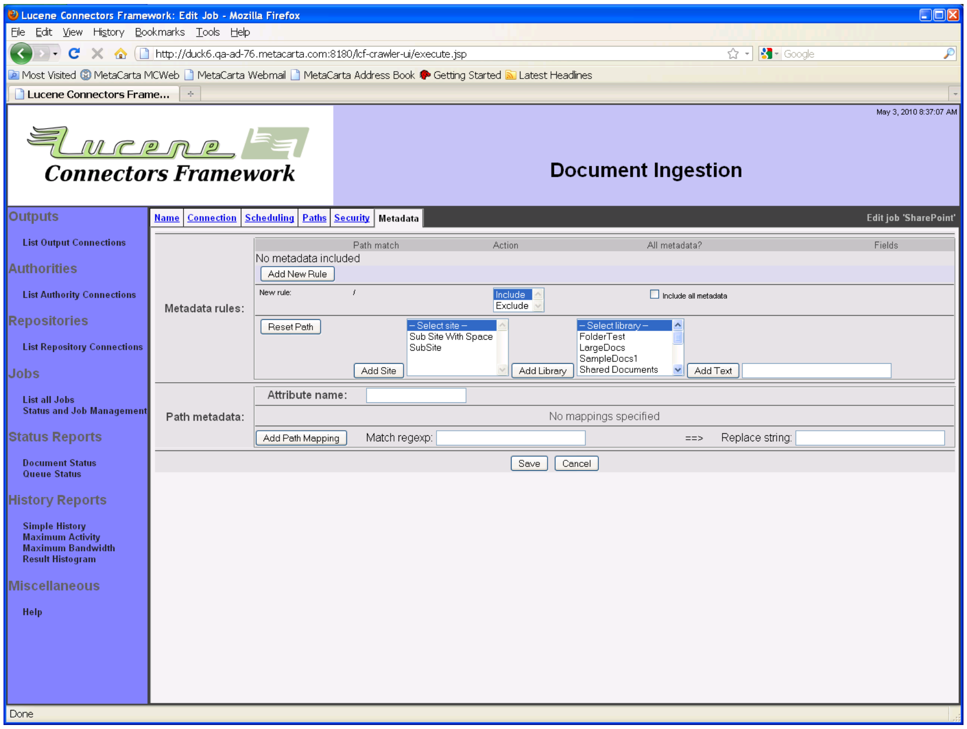
Edit job 'SharePoint'

Security: ● Enabled ○ Disabled

No access tokens present

Select whether SharePoint security is on or off using the radio buttons provided. If security is off, you may add access tokens in the text box and click the "Add" button. The access tokens must be in the proper form expected by the authority that governs your SharePoint connection for this feature to be useful.

The "Metadata" tab allows you to specify what metadata will be included for each document. The tab is similar to the "Paths" tab, which you may want to review above:



The main difference is that instead of rules that include or exclude individual sites, libraries, or documents, the rules describe inclusion and exclusion of document metadata. Since metadata is associated with files, all of the metadata rules are applied only to file paths, and there are no such things as "site" or "library" metadata path rules.

If an exclusion rule matches a file's path, it means that no metadata from that file will be included at all. There is no way to individually exclude a single field using an exclusion rule.

To build a rule, first build the rule's matching path. Make an appropriate selection or enter desired text, then click either the "Add Site", "Add Library", or "Add Text" button, depending on your choice. Repeat this process until the path is what you want it to be. Select whether this is an include or exclude rule. Either check the box for "Include all metadata", or select the metadata you want to include from the pulldown. (The choices of metadata fields you are presented with are determined by which SharePoint library is selected. If your rule path does not uniquely specify a library, you cannot select individual fields to include. You can only select "All metadata".) Then, click the "Add New Rule" button, to put your newly-constructed rule at the end of the list.

You can also use the "Metadata" tab to have the connection send path data along with each document, as a piece of document metadata. To enable this feature, enter the name of the metadata attribute you want this information to be sent into the "Attribute name" field. Then, add the rules you want to the list of rules. Each rule has a match expression, which is a regular expression where parentheses ("(" and ")") mark sections you are interested in. These sections are called "groups" in regular expression parlance. The replace string consists of constant text plus substitutions of the groups from the match, perhaps modified. For example, "$(1)" refers to the first group within the match, while "$(1l)" refers to the first match group mapped to lower case. Similarly, "$(1u)" refers to the same characters, but mapped to upper case.

For example, suppose you had a rule which had ".*/(.*)/(.*)/.*" as a match expression, and "$(1) $(2)" as the replace string. If presented with the path `Project/Folder_1/Folder_2/Filename`, it would output the string `Folder_1 Folder_2`.

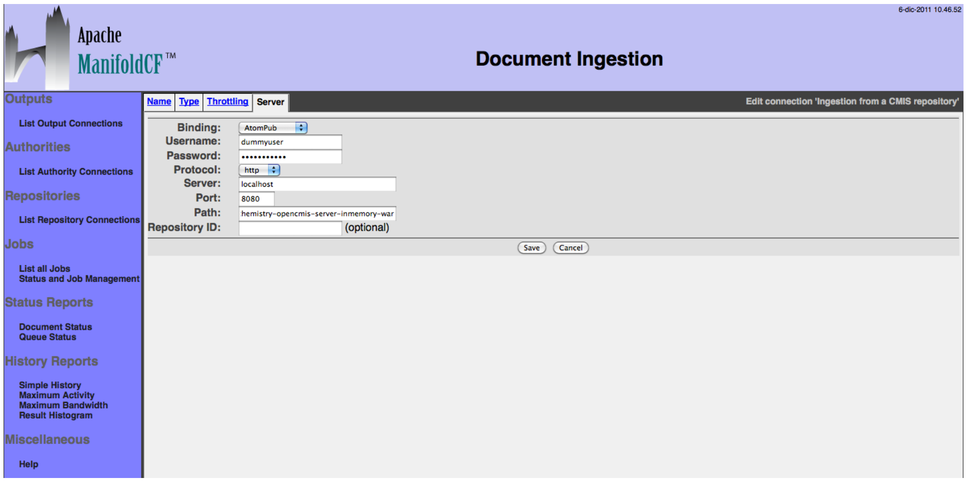
If more than one rule is present, the rules are all executed in sequence. That is, the output of the first rule is modified by the second rule, etc.

4.13 CMIS Repository Connection

The CMIS Repository Connection type allows you to index content from any CMIS-compliant repository.

By default each CMIS Connection manages a single CMIS repository, this means that if you have multiple CMIS repositories exposed by a single endpoint, you need to create a specific connection for each CMIS repository.

A CMIS connection has the following configuration parameters on the repository connection editing screen:



Edit connection 'Ingestion from a CMIS repository'

Binding: AtomPub
Username: dummyuser
Password: •••••••••••
Protocol: http
Server: localhost
Port: 8080
Path: hemistry-opencmis-server-inmemory-war
Repository ID: (optional)

Select the correct CMIS binding protocol (AtomPub or Web Services) and enter the correct username, password and the endpoint to reference the CMIS document server services.

The endpoint consists of the HTTP protocol, hostname, port and the context path of the CMIS service exposed by the CMIS server:

```
http://HOSTNAME:PORT/CMIS_CONTEXT_PATH
```

Optionally you can provide the repository ID to select one of the exposed CMIS repository, if this parameter is null the CMIS Connector will consider the first CMIS repository exposed by the CMIS server.

Note that, in a CMIS system, a specific binding protocol has its own context path, this means that the endpoints are different:

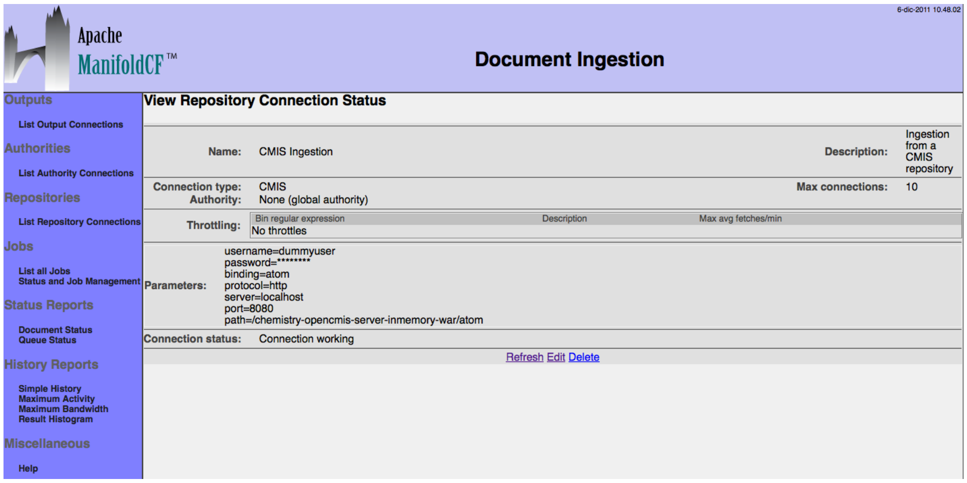
for example the endpoint of the AtomPub binding exposed by the actual version of the InMemory Server provided by the OpenCMIS framework is the following:

```
http://localhost:8080/chemistry-opencmis-server-inmemory-war-0.5.0-SNAPSHOT/atom
```

The Web Services binding is exposed using a different endpoint:

```
http://localhost:8080/chemistry-opencmis-server-inmemory-war-0.5.0-SNAPSHOT/services/RepositoryService
```

After you click the "Save" button, you will see a connection summary screen, which might look something like this:



View Repository Connection Status

Name: CMIS Ingestion
Description: Ingestion from a CMIS repository
Connection type: CMIS
Authority: None (global authority)
Max connections: 10

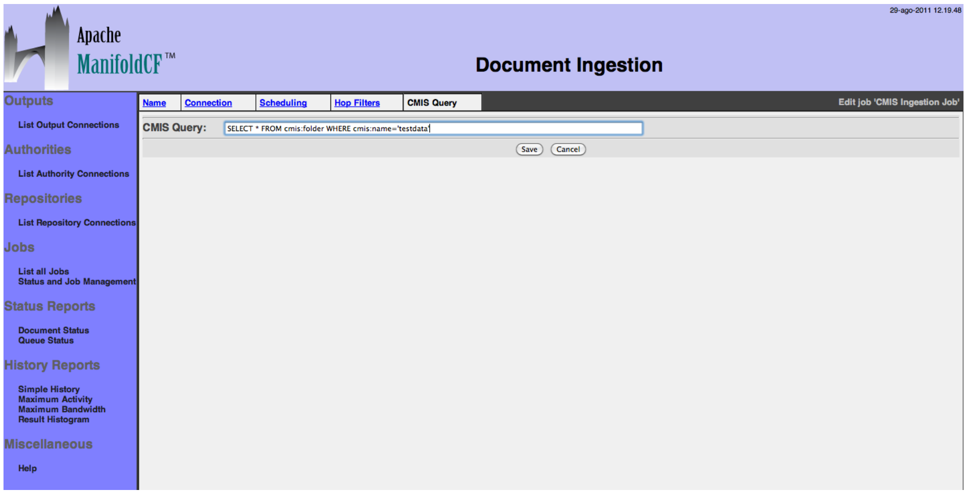
Throttling:	Bin regular expression	Description	Max avg fetches/min
	No throttles		

Parameters:

```
username=dummyuser
password=********
binding=atom
protocol=http
server=localhost
port=8080
path=/chemistry-opencmis-server-inmemory-war/atom
```

Connection status: Connection working

When you configure a job to use the CMIS repository connection an additional tab is presented. This is the "CMIS Query" tab:

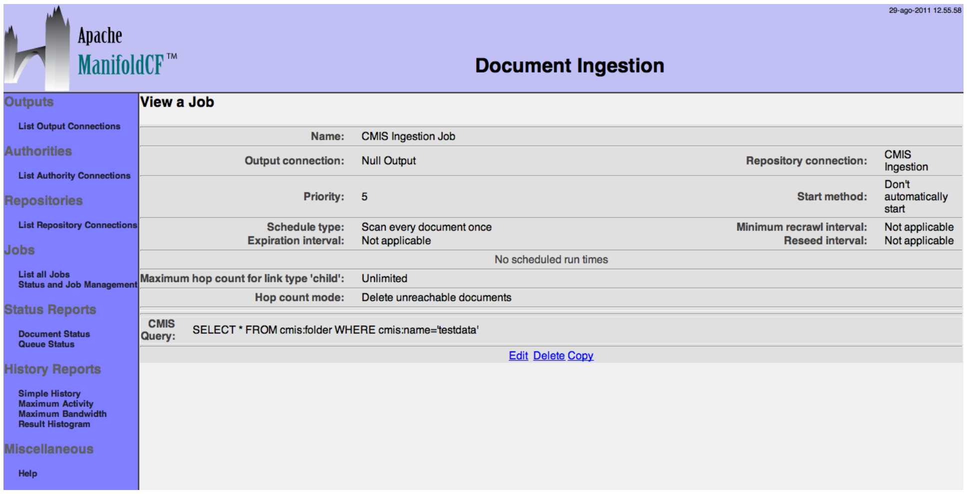


CMIS Query: SELECT * FROM cmis:folder WHERE cmis:name='testdata'

The CMIS Query tab allows you to specify the query based on the CMIS Query Language to get all the result documents that need to be ingested.

Note that the CMIS Connector during the ingestion process, for each result, if it will find a folder node (that must have cmis:folder as the baseType), it will ingest all the children of the folder node; otherwise it will directly ingest the document (that must have cmis:document as the baseType).

When you are done, and you click the "Save" button, you will see a summary page looking something like this:



View a Job

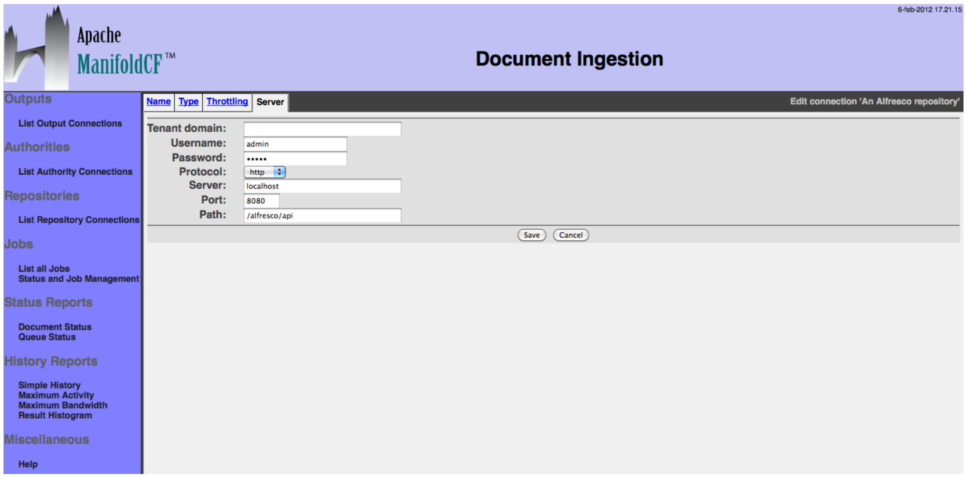
Name:	CMIS Ingestion Job		
Output connection:	Null Output	Repository connection:	CMIS Ingestion
Priority:	5	Start method:	Don't automatically start
Schedule type: Expiration interval:	Scan every document once Not applicable	Minimum recrawl interval: Reseed interval:	Not applicable Not applicable
No scheduled run times			
Maximum hop count for link type 'child':	Unlimited		
Hop count mode:	Delete unreachable documents		
CMIS Query:	SELECT * FROM cmis:folder WHERE cmis:name='testdata'		

4.14 Alfresco Repository Connection

The Alfresco Repository Connection type allows you to index content from an Alfresco repository.

This connector is compatible with any Alfresco version (2.x, 3.x and 4.x).

An Alfresco connection has the following configuration parameters on the repository connection editing screen:



Edit connection 'An Alfresco repository'

Tenant domain:	
Username:	admin
Password:	•••••
Protocol:	http
Server:	localhost
Port:	8080
Path:	/alfresco/api

Enter the correct username, password and the endpoint to reference the Alfresco document server services.

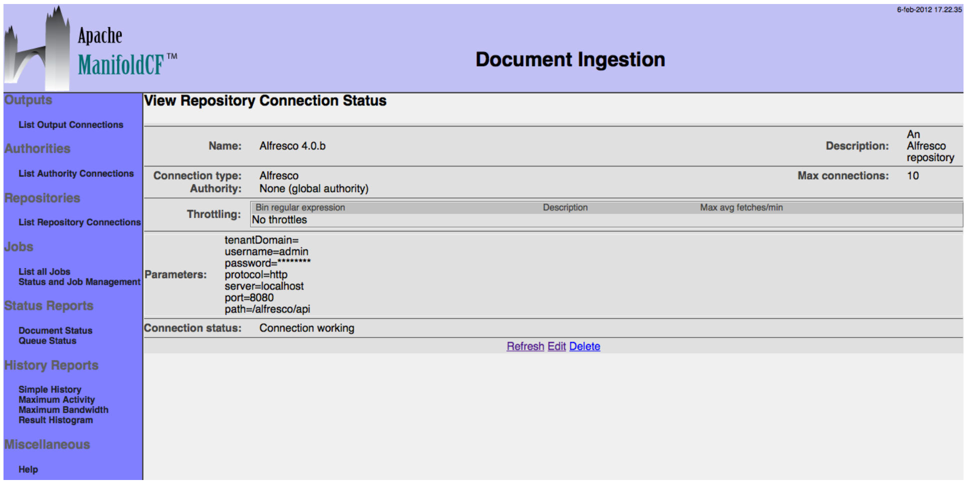
If you have a Multi-Tenancy environment configured in the repository, be sure to also set the tenant domain parameter.

The endpoint consists of the HTTP protocol, hostname, port and the context path of the Alfresco Web Services API exposed by the Alfresco server.

By default, if you don't have changed the context path of the Alfresco webapp, you should have an endpoint address similar to the following:

```
http://HOSTNAME:PORT/alfresco/api
```

After you click the "Save" button, you will see a connection summary screen, which might look something like this:



View Repository Connection Status

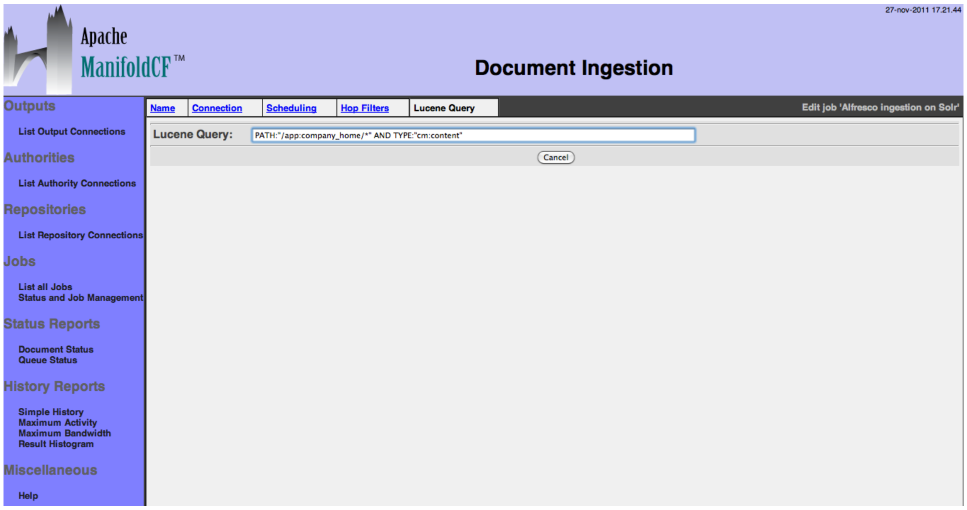
Name: Alfresco 4.0.b
Description: An Alfresco repository
Connection type: Alfresco
Max connections: 10
Authority: None (global authority)
Throttling:

Bin regular expression	Description	Max avg fetches/min
No throttles		

Parameters:
tenantDomain=
username=admin
password=********
protocol=http
server=localhost
port=8080
path=/alfresco/api

Connection status: Connection working

When you configure a job to use the Alfresco repository connection an additional tab is presented. This is the "Lucene Query" tab:



Edit job 'Alfresco ingestion on Solr'

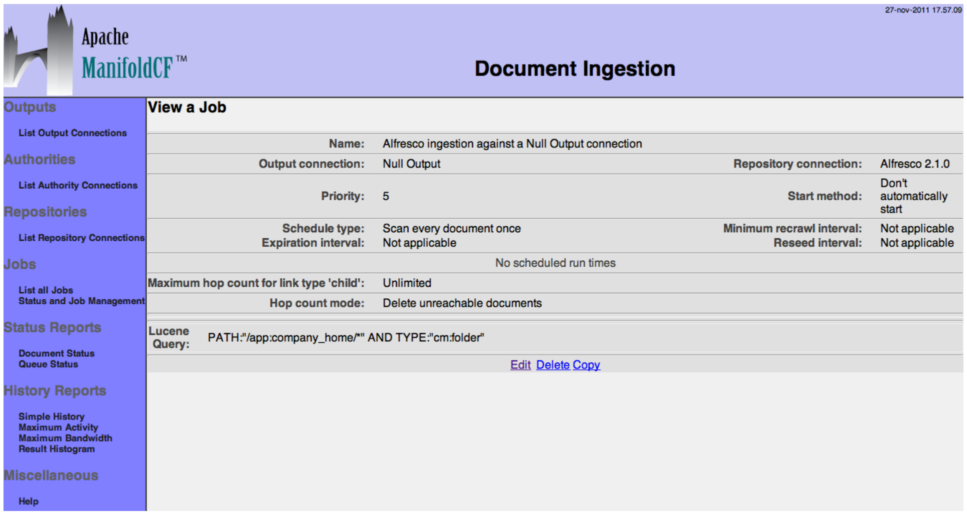
Lucene Query: PATH:"/app:company_home/*" AND TYPE:"cm:content"

The Lucene Query tab allows you to specify the query based on the Lucene Query Language to get all the result documents that need to be ingested.

Please note that if the Lucene Query is null, the connector will ingest all the contents in the Alfresco repository under the Company Home.

Please also note that the Alfresco Connector during the ingestion process, for each result, if it will find a folder node (that must have any child association defined for the node type), it will ingest all the children of the folder node; otherwise it will directly ingest the document (that must have any d:content as one of the properties of the node).

When you are done, and you click the "Save" button, you will see a summary page looking something like this:



View a Job

Name:	Alfresco ingestion against a Null Output connection		
Output connection:	Null Output	Repository connection:	Alfresco 2.1.0
Priority:	5	Start method:	Don't automatically start
Schedule type: Expiration interval:	Scan every document once Not applicable	Minimum recrawl interval: Reseed interval:	Not applicable Not applicable
No scheduled run times			
Maximum hop count for link type 'child':	Unlimited		
Hop count mode:	Delete unreachable documents		
Lucene Query:	PATH:"/app:company_home/*" AND TYPE:"cm:folder"		