memex-explorer Documentation

Release 0.4

Andy Terrel, Christine Doig, Ben Zaitlen, Karan Dodia, Brittain Ha

January 22, 2017

Contents

Memex Explorer is a web application that provides easy-to-use interfaces for gathering, analyzing, and graphing web crawl data.

For usage instructions, please refer to the User's Guide.

For more information about the project architecture, please refer to our Developer's Guide and API Guide.

Memex Explorer is built by Continuum Analytics, with grants and support from the NASA Jet Propulsion Laboratory, Kitware, and the NYU Polytechnic School of Engineering.

Contents:

CHAPTER 1

User's Guide to Memex Explorer

NOTE: Memex Explorer is still under active development, and this guide is constantly evolving as a result. For documentation requests, please file an issue and we will endeavor to address it as soon as possible.

1.1 Application Structure

The goal of Memex explorer is the bring together the functionality of several applications in a seamless way, in order to assist the user in searching the deep web for domain specific information. Memex Explorer has integration with several applications, providing a front-end to various crawlers and domain search tools.

Web Crawling With Memex Explorer you can create, run, and analyze Nutch and ACHE crawls. The crawl operation is heavily abstracted and simplified. Users provide a list of seed URLs to start the crawl, and in the case of ACHE's targeted crawling, a machine learning model to determine the relevancy of crawled pages.

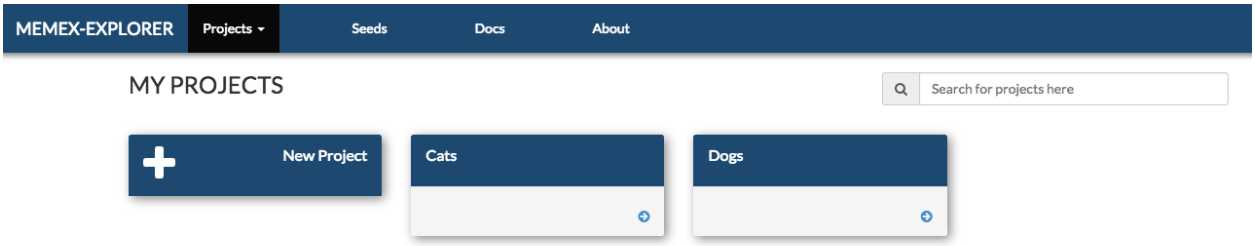
Dataset Analysis Memex Explorer allows you to upload a large number of files, which will be analyzed by Tika and placed into our Elasticsearch instance. Tika will exctact metadata from these documents, giving you a better overview of them.

Domain Discovery Tool Through the use of Domain Discovery Tool, the user can search for content in the web and build data models based on clustering algorithms. The user can search the web and highlight relevant and irrelevant pages, and DDT will produce data model files, which you can use with Ache crawls in Memex Explorer.

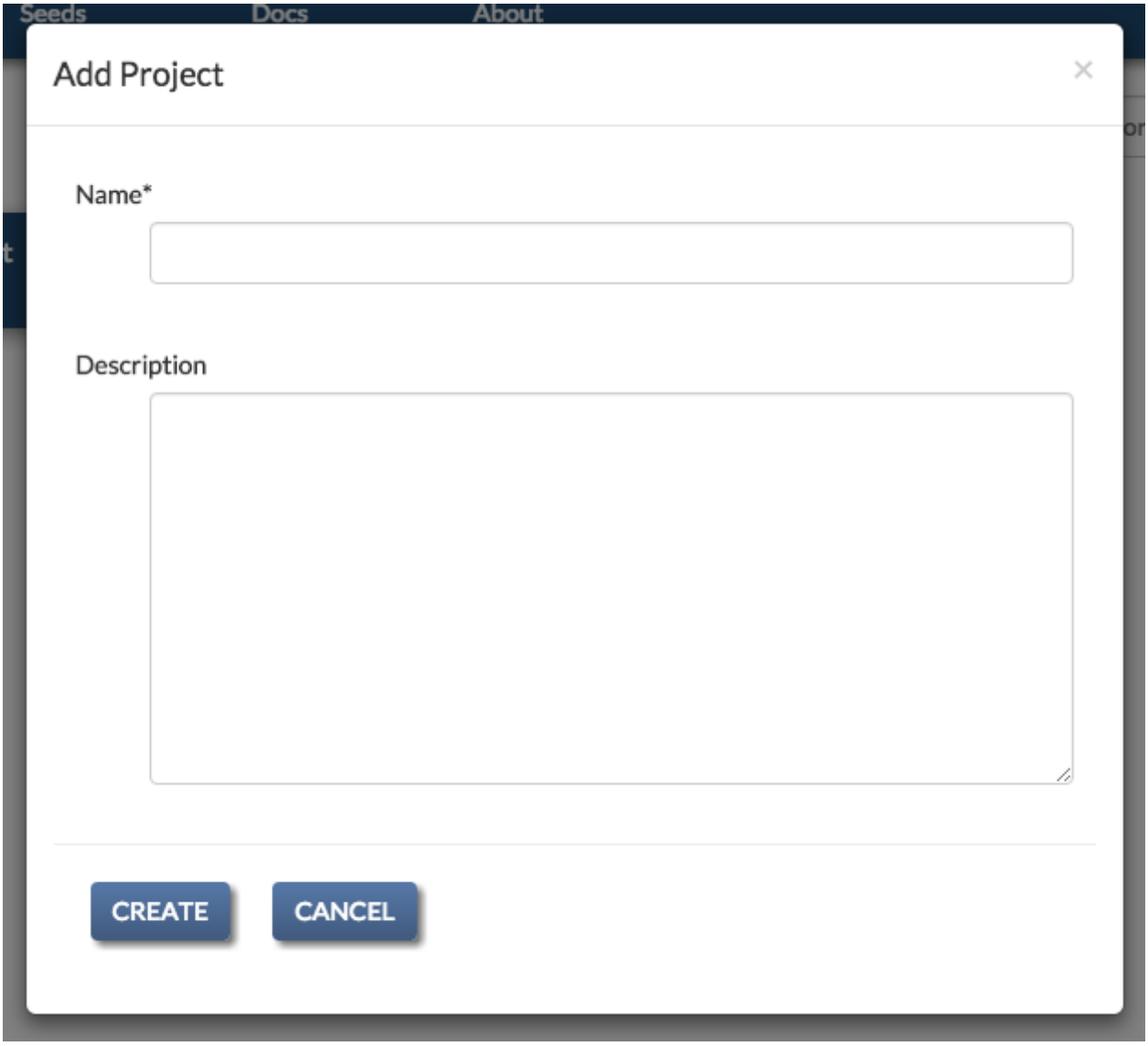
DataWake DataWake is a server and firefox plugin that tracks your search investigations. It keeps track of where you search, so that "trails" can be built out of the information that you gather. These trails can be converted to seeds lists in Memex Explorer, and can be used in both Nutch and Ache crawls.

1.2 Home Page

The landing page lists the currently registered projects. All the capabilities of Memex Explorer live under this project abstraction.

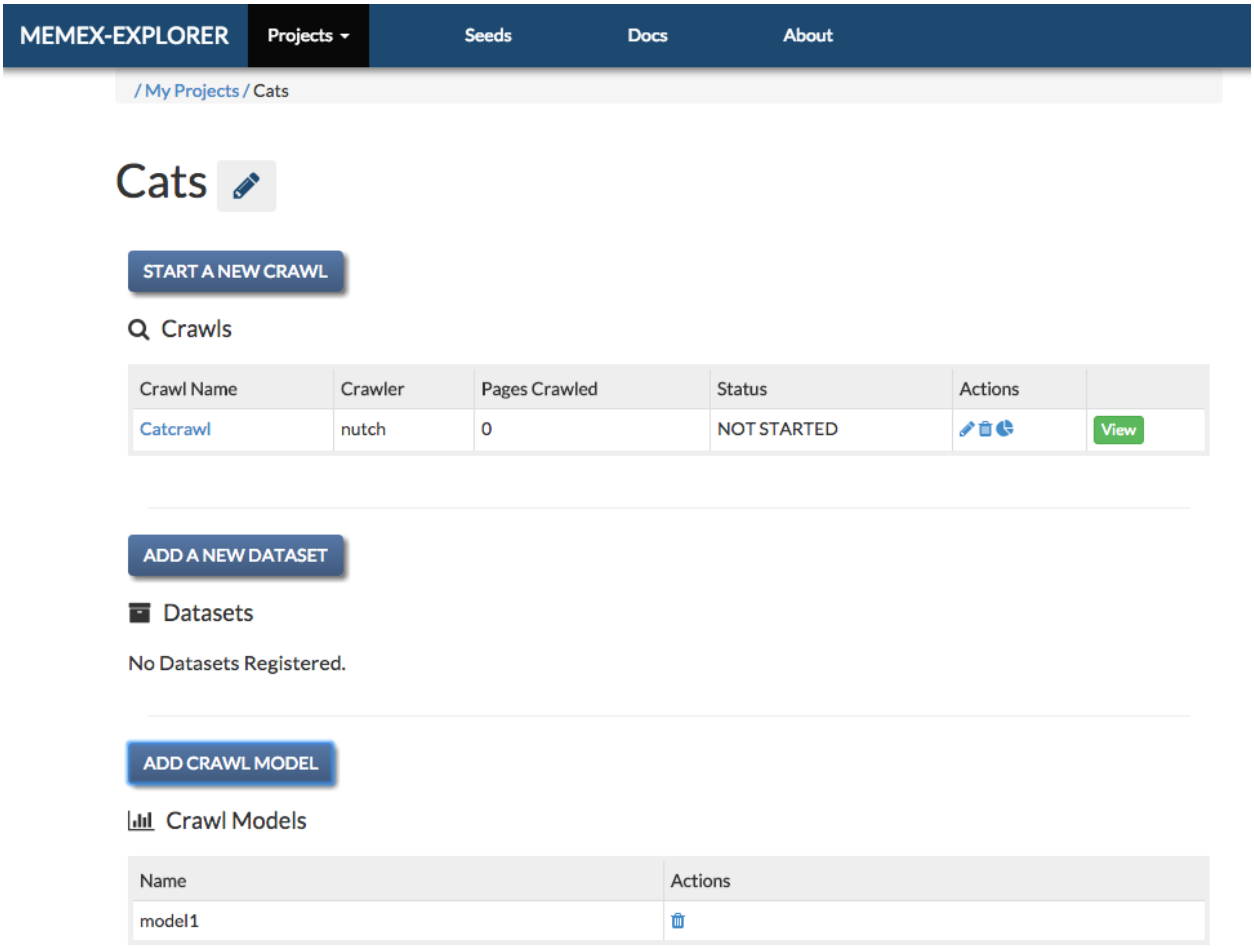


Creating a project just requires adding a name and an optional description.



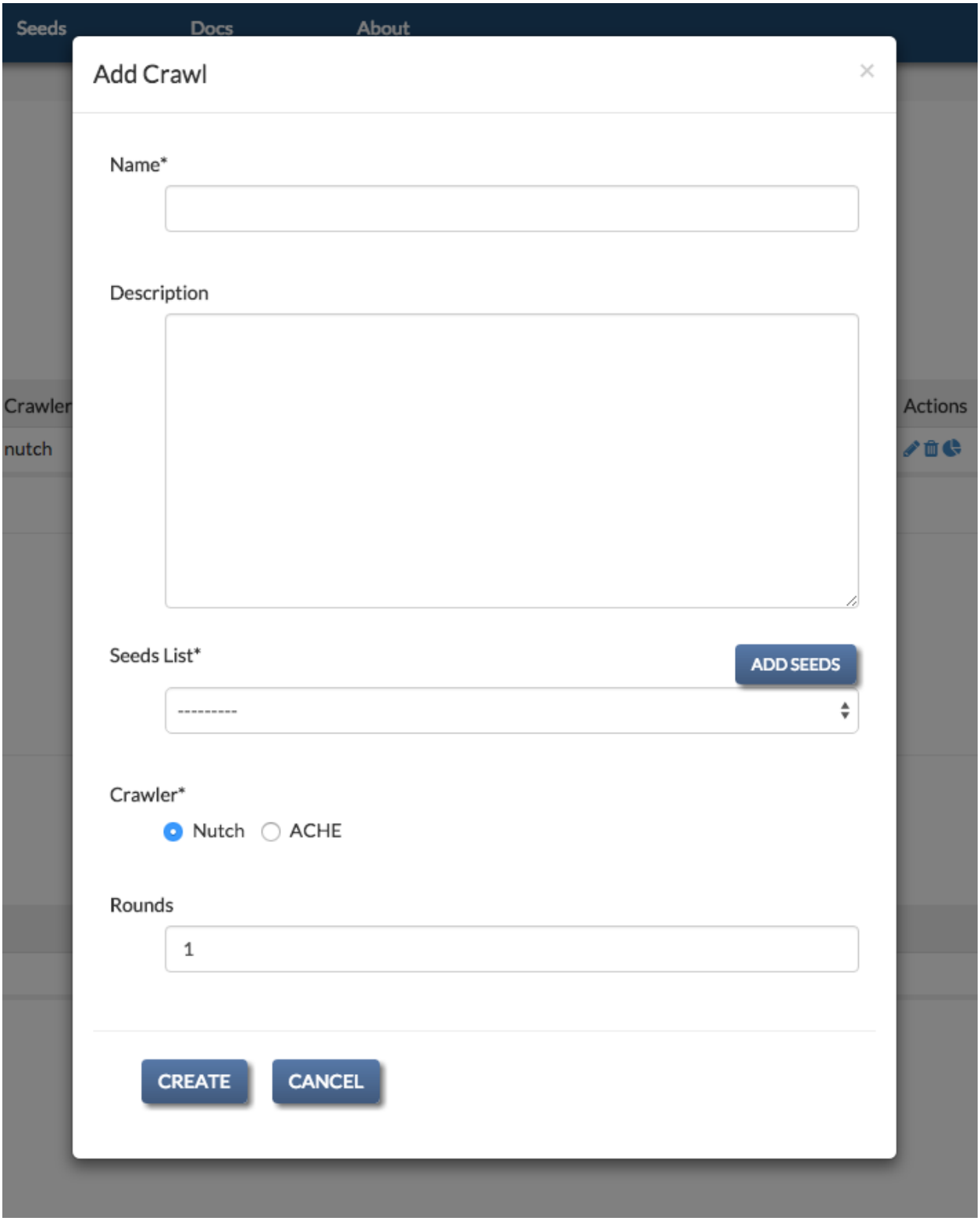
1.3 Project Page

The project page lists the currently available services in Memex Explorer. These services can all be access from the project page.



1.3.1 Registering a Crawl

To register a new crawl, click the "Add Crawl" button above the Crawls table. This will open a popup for adding crawls. If necessary, you can also create seeds list objects and crawl models from the same form.



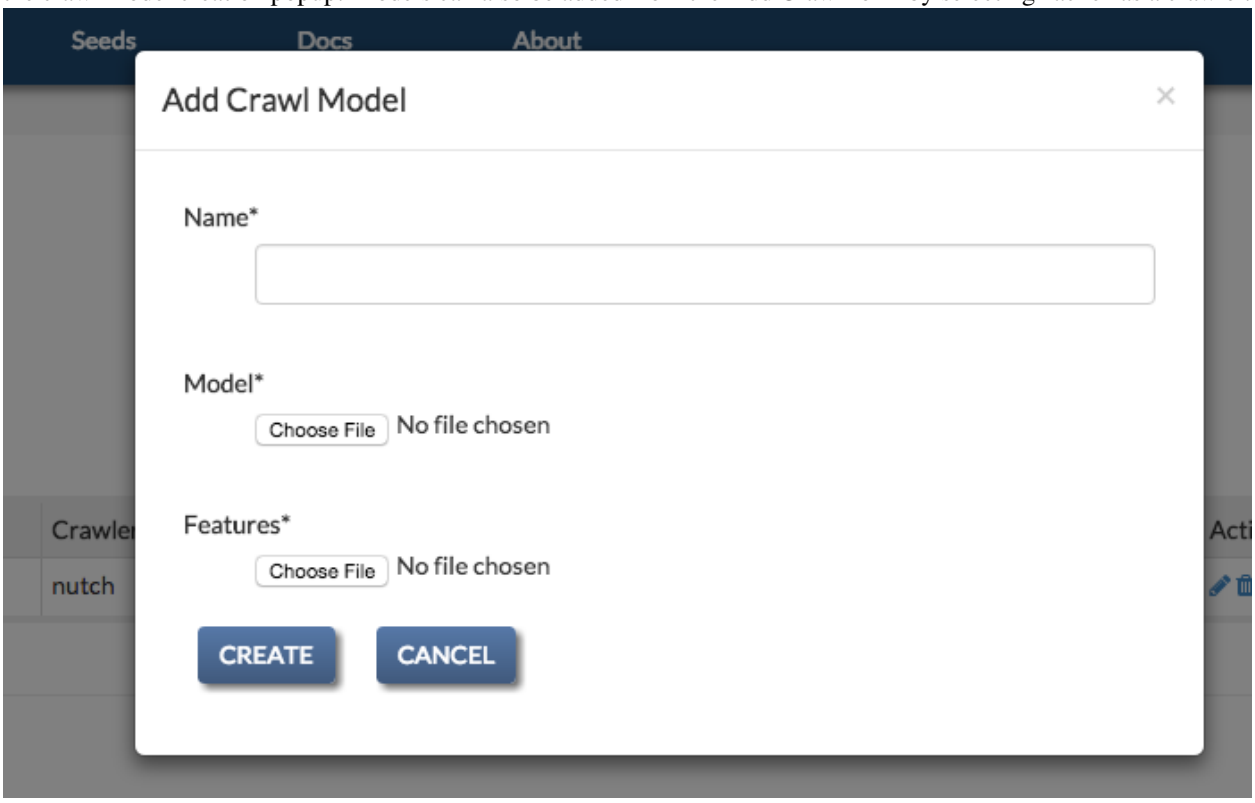
For both crawls, you have to supply a seeds list object, which contains the list of urls to be crawled. The seeds list object can be created from the Add Crawl form.

For ACHE crawls, you have to specify a crawl model for the crawl, which can also be added from the Add Crawl form.

1.3.2 Registering a Crawl Model

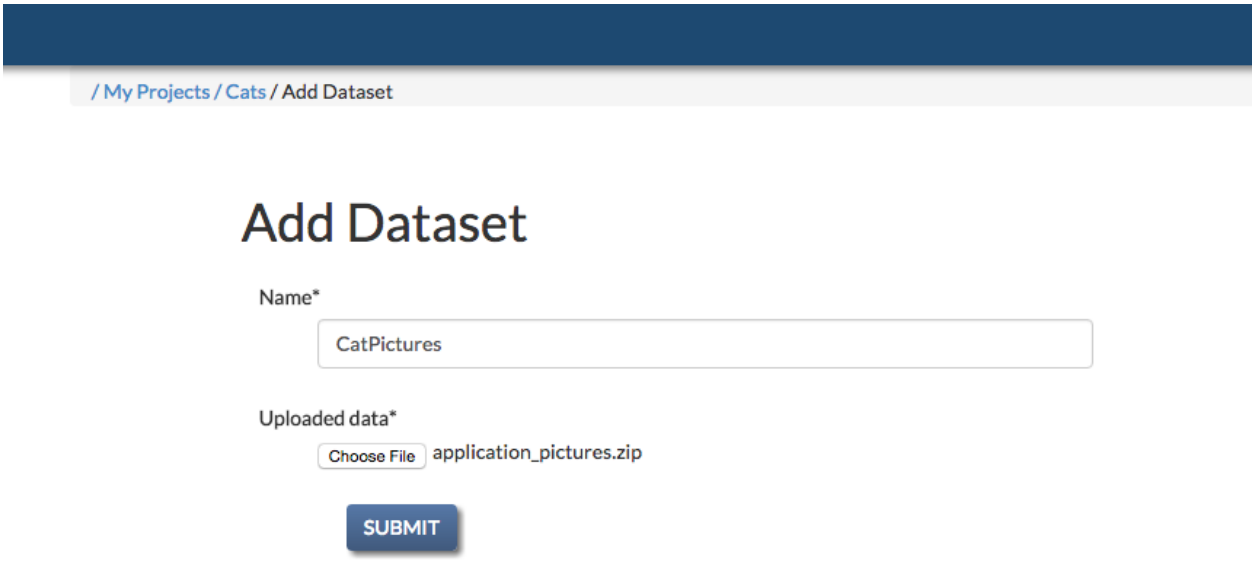
ACHE crawls require a *Crawl Model* to power the page classifier. The model consists of two elements: a "model" file and "features" file. These can be generated by following the instructions on the ACHE GitHub page.

To register a new crawl model, click on the "Add Crawl Model" button in the Crawl Models header. This will bring up the crawl model creation popup. Models can also be added from the Add Crawl form by selecting "ache" as a crawler.

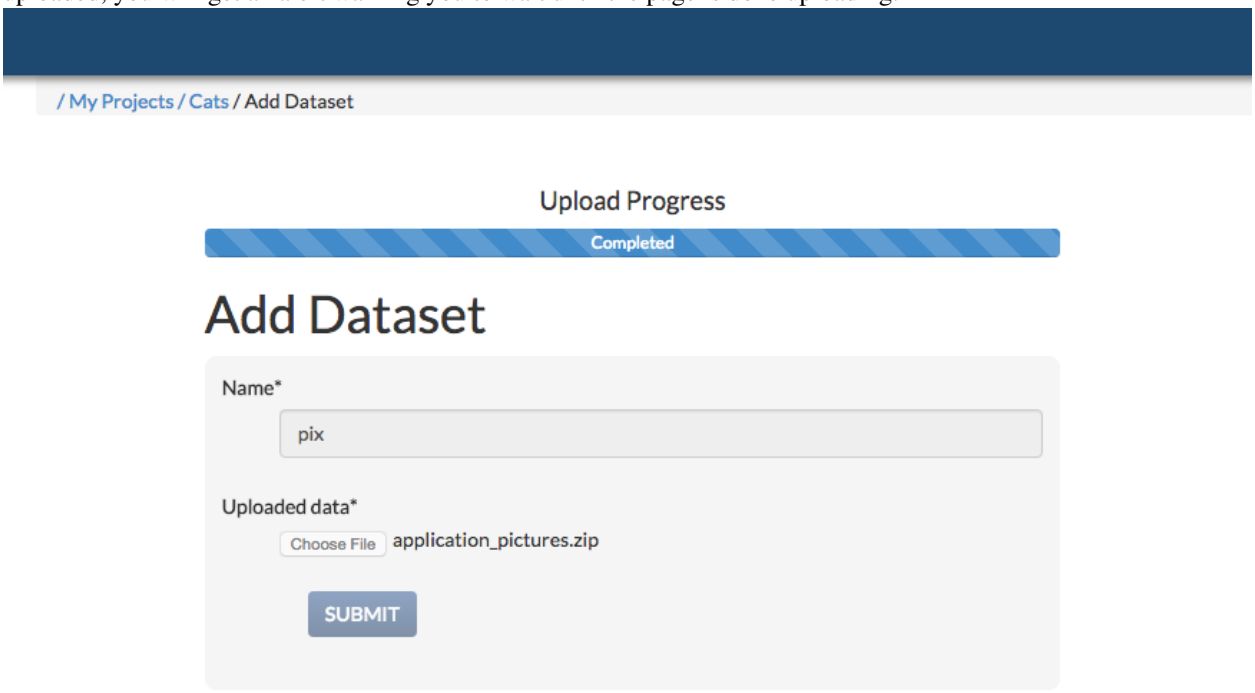


1.3.3 Uploading Files and Dataset Creation

With Memex Explorer you can create indices by uploading zipfiles of important documents. Memex Explorer will analyze these documents with Tika. You can then easily access the documents from the local Elasticsearch index, and incorporate them into other data analysis tools. You can create the dataset by clicking "Add Dataset" on the project page.

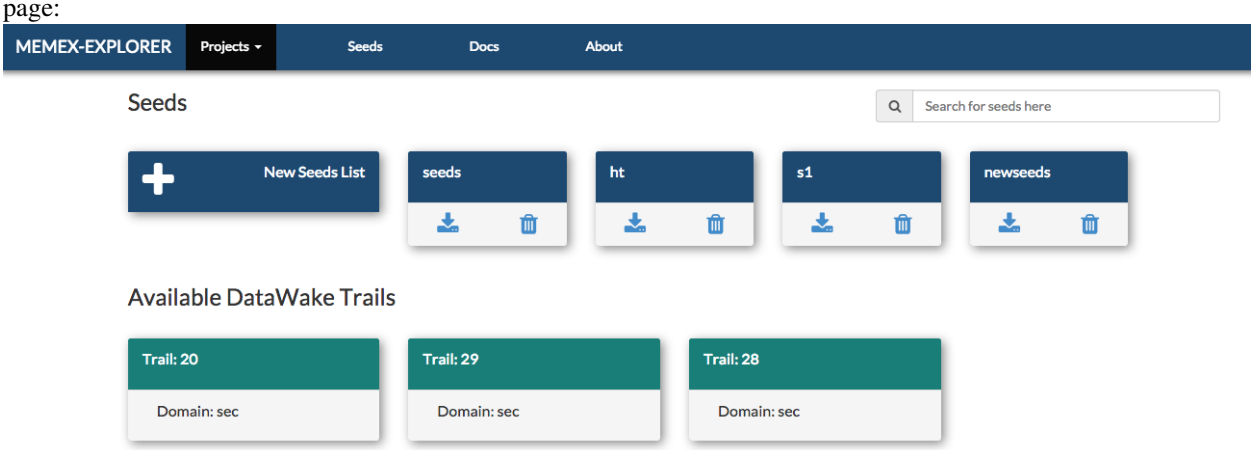


The add dataset page has a progress bar, and when your dataset has been successfully uploaded, you will get a success message and an alert to close the page. If you attempt to close the page before the files have been successfully uploaded, you will get an alert warning you to wait until the page is done uploading.



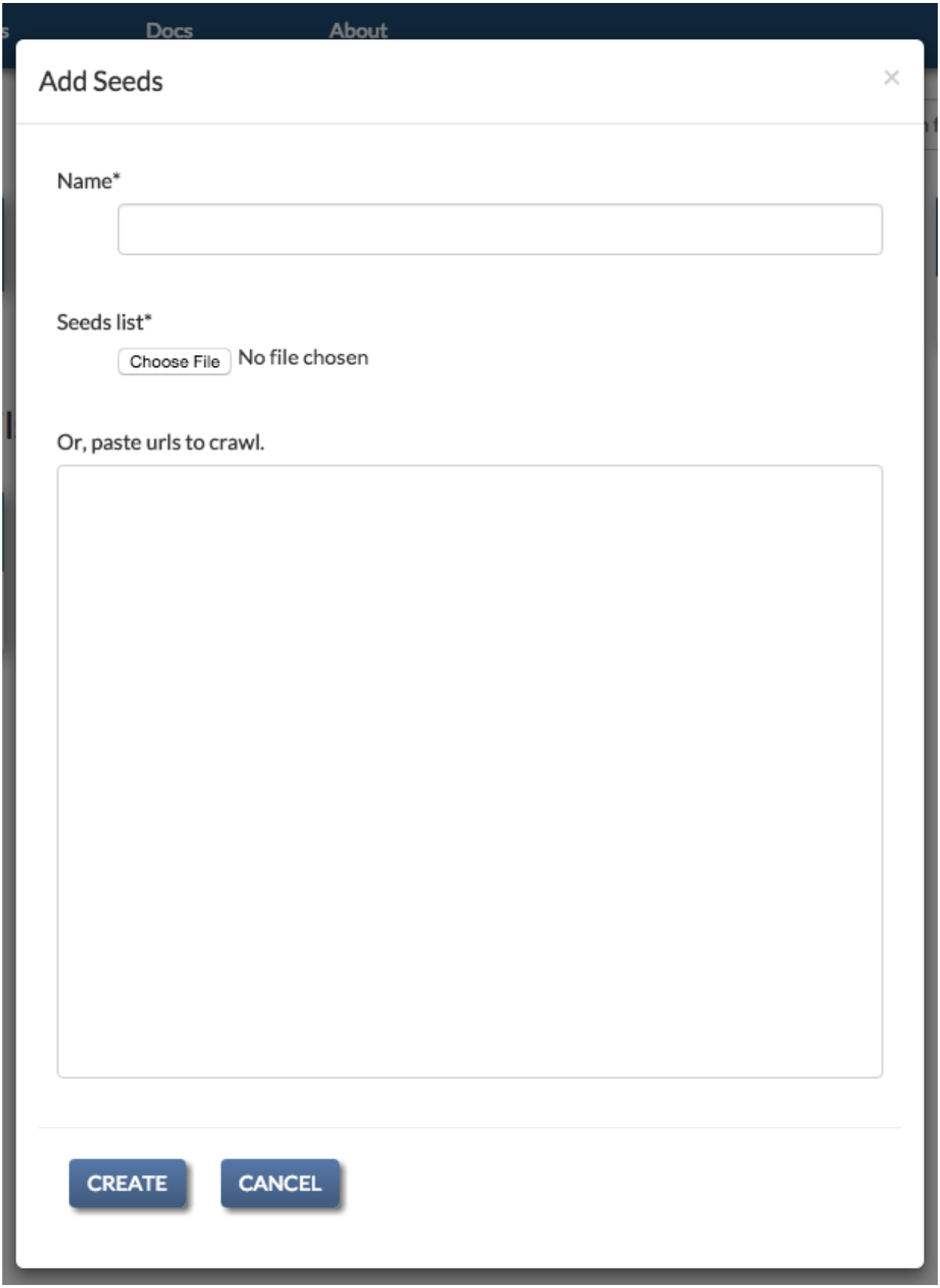
1.4 Seeds List Page

Seeds for crawls are independent of projects. They are created by clicking on the "Seeds" button on the navbar. From the seeds list page you can create seeds lists from files, text, or from datawake trails. You can also edit the seeds on a separate page. In addition, you can delete and download any of the seeds objects that you create. This is the seeds list page:



1.4.1 Registering a Seeds List

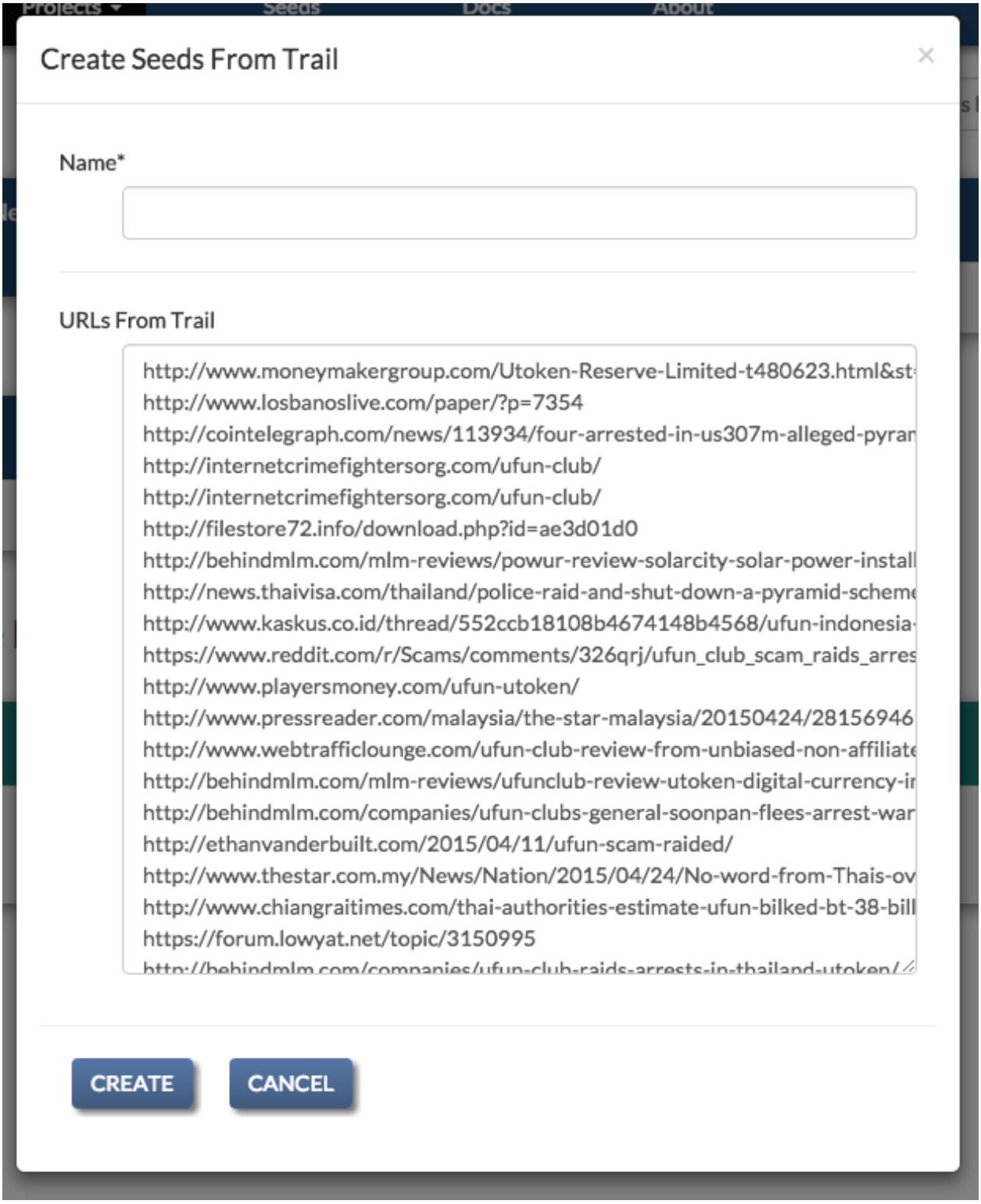
Each crawl requires a seeds list object. Ache requires the seed list in a textfile, whereas Nutch requires a seeds list injection. The seeds list object handles both of these requirements. It creates a file for Ache and contains fields for injecting seeds through the Nutch REST Api. All seeds objects can be added on the "Add Crawl" popup. This is the seeds list form.


Docs
About
Add Seeds
Name*
Seeds list*
Choose File
No file chosen
Or, paste urls to crawl.
CREATE
CANCEL

Seeds require a valid name, and either a file or URLs placed in the textarea below. If any of your seeds are invalid, you will get a form error, and all the invalid urls will be highlighted.

1.4.2 Creating a Seeds List from a DataWake Trail

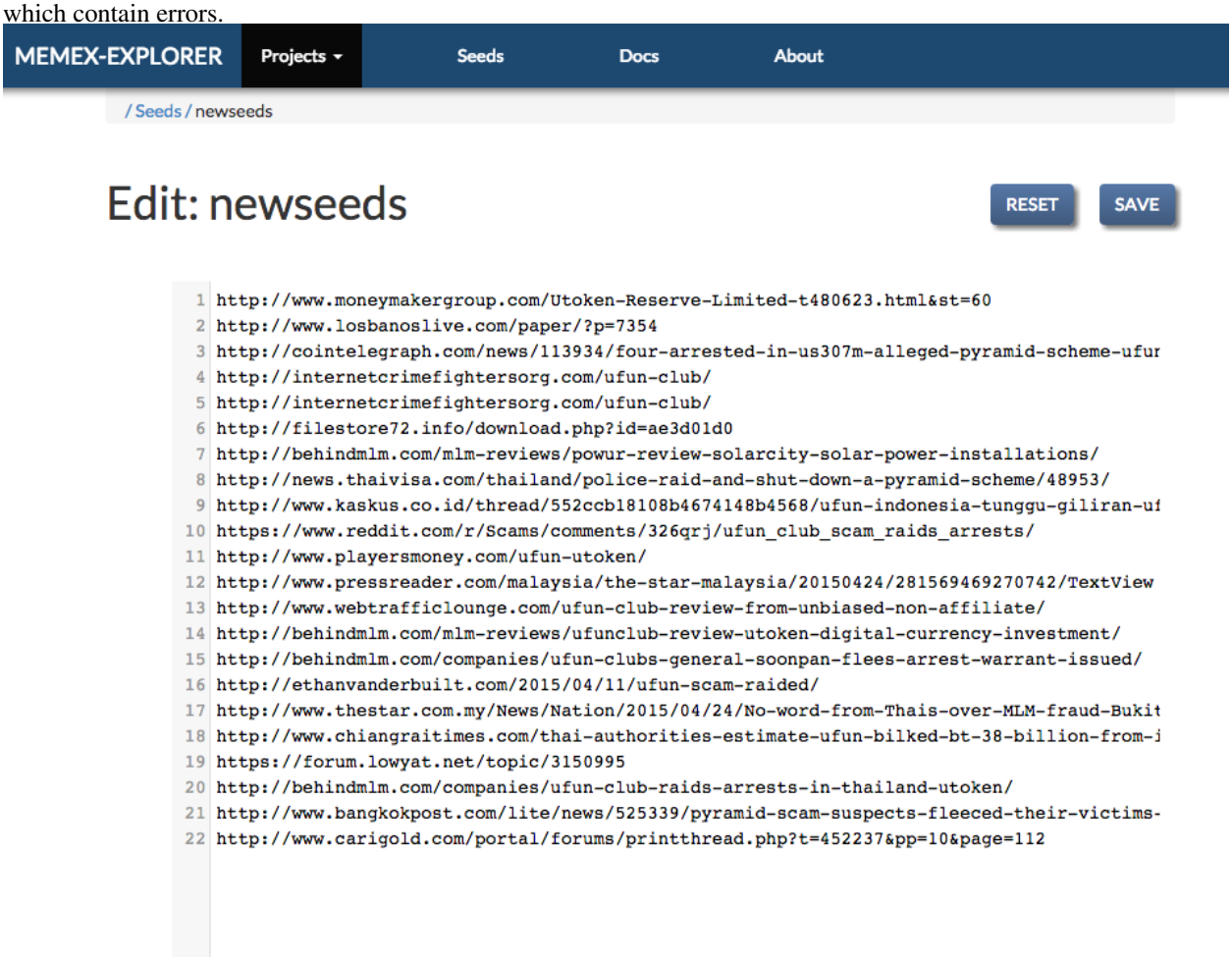
If you are using DataWake, and Memex Explorer has access to the index used by DataWake, you will be able to create seeds lists from DataWake trails. To create a seeds list, all that is required is a valid name. After you create the seeds list, you can edit it just like any other seeds list.



1.4.3 Editing a Seeds List

Once you have created your seeds list, you can edit through our built in editor. This editor allow you to change the content of your seeds list, by adding or removing seeds. It will also validate all of the URLs and display the ones

which contain errors.



CHAPTER 2

Memex Explorer Crawler Guide

memex-explorer uses two crawlers, *Ache* and *Nutch*.

2.1 Crawler Overview

Both crawlers have their own unique designs, and both use the data they collect in unique ways.

There is some commonality between the two, however. They both require a list of URLs to crawl, called a seeds list, and they both share similar interactivity with the *Crawler Control Buttons*.

This section will go over the common elements of the two crawlers.

2.1.1 Creating a Seeds List

The common point between the two crawlers is that they both use the same kind of seeds list for their crawling. The seeds list is comprised of a list of urls separated by line breaks. Both Nutch and Ache use them in different ways, and the result you get directly from the crawlers is different for each of them. Here is a sample seeds list:

```
http://www.reddit.com/r/aww
http://gizmodo.com/of-course-japan-has-an-island-where-cats-outnumber-peop-1695365964
http://en.wikipedia.org/wiki/Cat
http://www.catchannel.com/
http://mashable.com/category/cats/
http://www.huffingtonpost.com/news/cats/
http://www.lolcats.com/
```

Simply put, the seeds list should contain pages that are relevant to the topics you are searching. Both Nutch and Ache provide insight into the relevance of your seeds list, but in different ways.

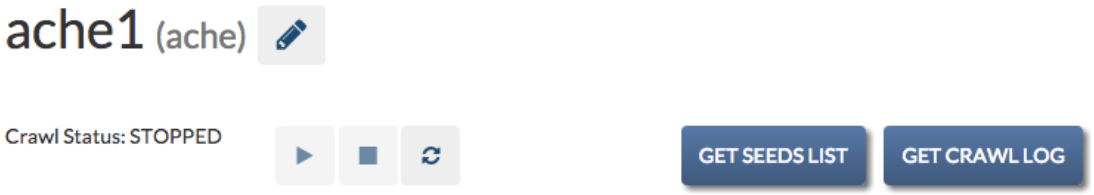
For the purposes of memex-explorer, the extension and name of your seeds list does not matter. It will be automatically renamed and stored according to the specifications of the crawler.

Seeds lists are created on the seeds page, and seeds lists can be created from the add crawl page.

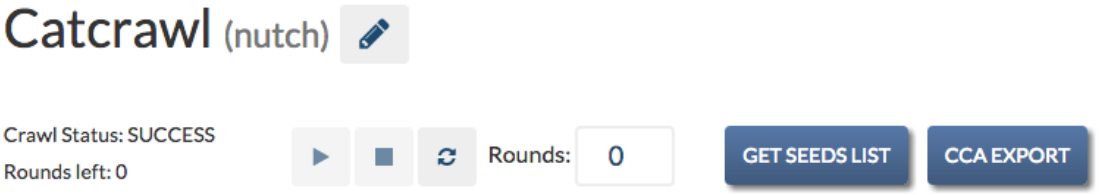
2.1.2 Crawler Control Buttons

Here we have an overview of the buttons available to each crawler for controlling the crawlers. The buttons behave differently depending on which one you are using.

These are the buttons available for Ache:



These are the buttons available for Nutch:



Options Button

Symbolized by the "pencil" icon. This allows you to change various settings on the crawl. See *Crawl Settings*.

Start Button

Symbolized by the "play" button. This will start the crawler for you, and will display the status as "starting" immediately after pressing it, and "running" after the crawl has been started.

Stop Button

Symbolized by the "stop" button. Stops the crawl.

In the case of Ache, the crawler stops immediately. In the case of Nutch, the crawler stops after it has finished the current process. However, the data on the current round of the crawl will be lost.

Restart Button

Symbolized by the "refresh" icon. Restarts the current crawl. This button is only available after the crawl has stopped.

With Ache, it will immediately start a brand new Ache crawl, deleting all of the previous crawl information. With Nutch, it will start a new crawler round, using the information gathered by the crawl in the previous round.

Get Crawl Log

This button will let you download the log of the current running crawl. This allows you to see the progress of the crawl and any errors that may be occurring during the crawl. This is only available for Ache crawls.

CCA Export

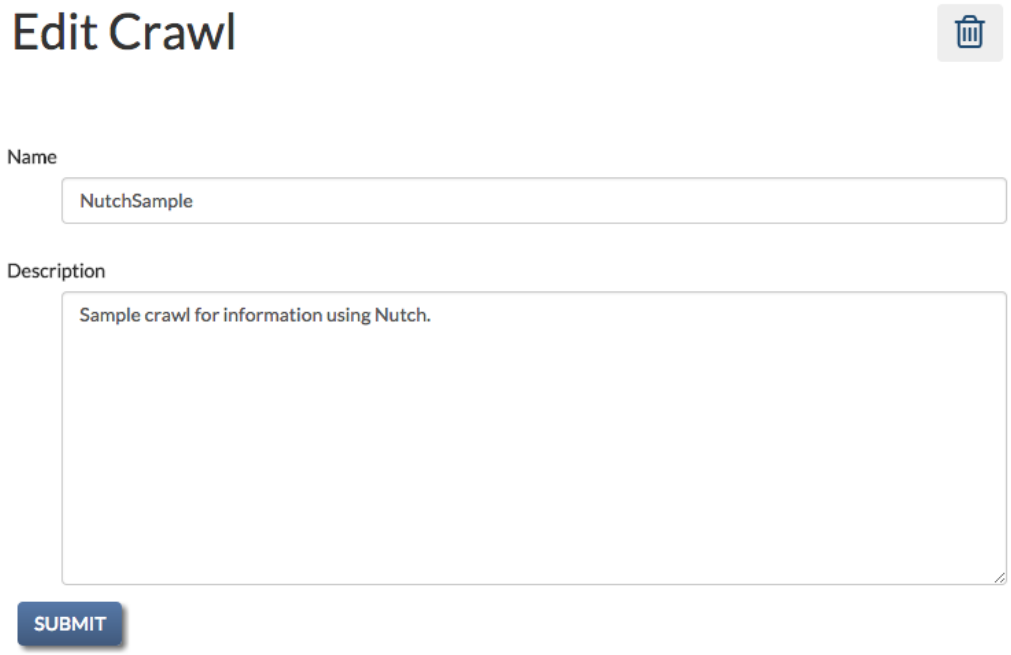
This button is Nutch only. It allows you to export your crawl data into the CCA format.

Rounds Input

Nutch only. This allows you to specify how many rounds you want the crawl to run. You can press the stop button at any time and it will stop when it is done with the current round.

2.1.3 Crawl Settings

The crawl settings page allows you to delete the crawl, as well as change the name or description of the crawl. It is accessed by clicking the "pencil" icon next to the name of the crawl.



Here you can change the name or description of the crawl. You can also delete the crawl.

2.2 Nutch

Nutch is developed by Apache, and has an interface with Elasticsearch. All Nutch crawls create Elasticsearch indices by default.

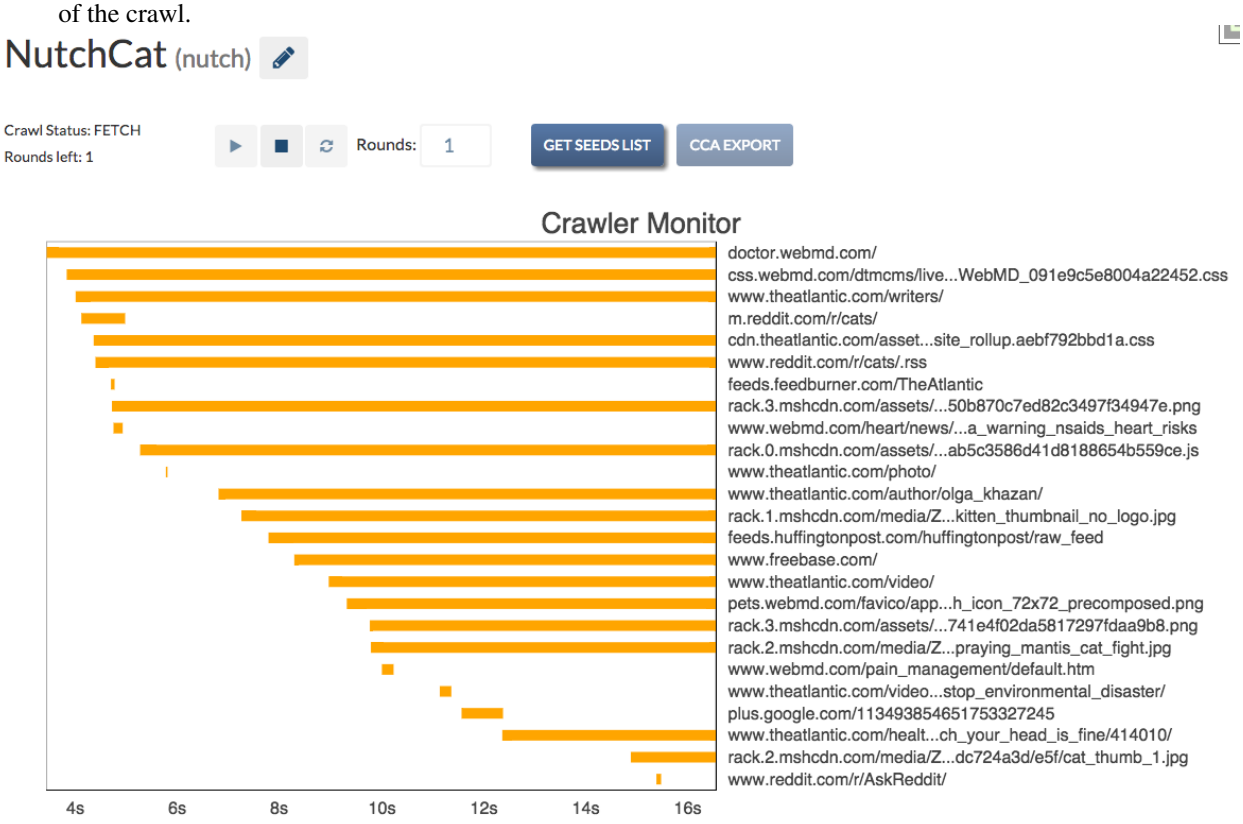
With Nutch, you can define how long you want to crawl by setting the number of rounds to crawl. You can keep track of the overall crawl time and the sites currently being crawled by looking at the Nutch crawl visualizations.

The number of pages left to crawl in a Nutch round increases significantly after each round. You might pass it a seeds list of 100 pages to crawl, and it can find over 1000 pages to crawl for the next round. Because of this, Nutch is a much easier crawler to get running.

Memex Explorer currently uses the Nutch REST API for running all crawls.

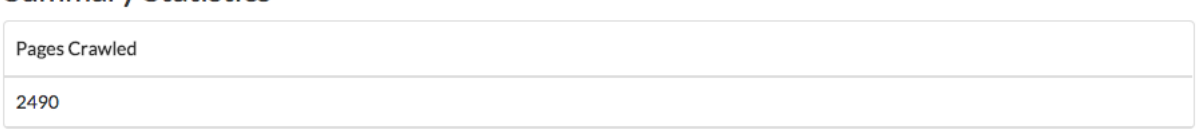
2.2.1 Nutch Dashboard

Memex explorer recently added features for monitoring the status of Nutch crawls. You can now get real-time information about which pages Nutch is currently crawling, and information about the duration of the crawl.



Statistics

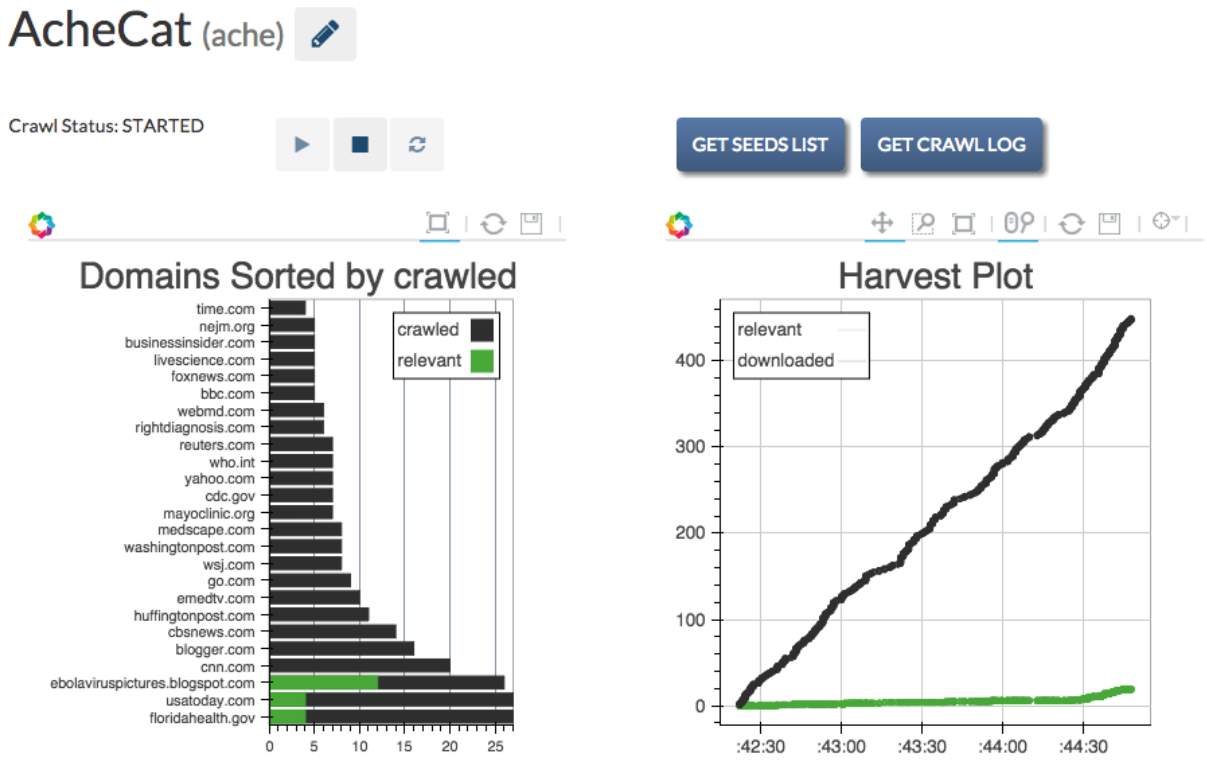
Nutch will tell you how many pages have been crawled after the current round has finished.



2.3 Ache

Ache is developed by NYU. Ache is different from Nutch because it requires a crawl model to be created before you can run a crawl (see *Building a Crawl Model*). Unlike Nutch, Ache can be stopped at any time. However, if you restart an Ache crawl, it will erase all the data from the previous crawl.

2.3.1 Ache Dashboard



Plots

Memex Explorer uses Bokeh for its plots. There are two plots available for analyzing Ache crawls, Domain Relevance and Harvest Rate.

The Domain Relevance plot sorts domains by the number of pages crawled, and adds information for relevancy of that domain to your crawl model. This plot helps you understand how well your model fits.

The Harvest Rate plot shows the overall performance of the crawl in terms how many pages were relevant out of the total pages crawled.

Statistics

Like Nutch, Ache also collects statistics for its crawls, and allows you to see the head of the seeds list.

Harvest rate reflects the relevance to the model of the pages crawled. In this case, 58% of the pages crawled were relevant according to the model.

Ache Specific Buttons

Ache has a "Download Relevant Pages" button, which will allow you download which pages Ache has found to be relevant to your seeds list and your crawl model.

Building a Crawl Model

Ache requires a crawl model to run. For information on how to build crawl models, see the Ache readme.

For more detailed information on Ache, head to the Ache Wiki.

CHAPTER 3

Developer's Guide to Memex Explorer

3.1 Setting up Memex Explorer

To set up your machine, you will need Anaconda or Miniconda installed. Miniconda is a minimal Anaconda installation that bootstraps conda and Python on any operating system. Install Anaconda or Miniconda from their respective sites.

Memex Explorer requires conda, either from Miniconda or Anaconda.

3.1.1 Application Setup

To set up a developer's environment, clone the repository, then run the app_setup.sh script:

```
$ git clone https://github.com/memex-explorer/memex-explorer.git
$ cd memex-explorer/source
$ ./app_setup.sh
```

You can then start the application from this directory:

```
$ source activate memex
$ supervisord
```

Memex Explorer will now be running locally at http://localhost:8000.

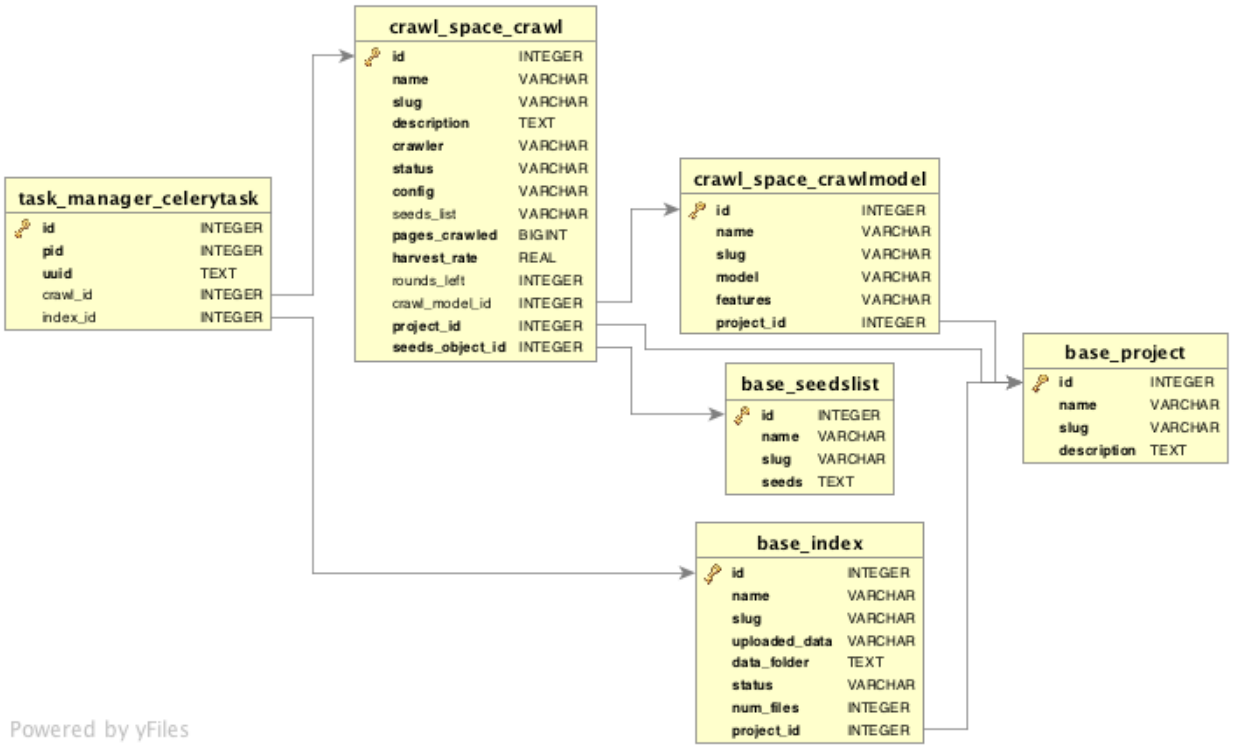
3.1.2 Tests

To run the tests, return to the root directory and run:

```
$ py.test
```

3.1.3 The Database Model

The current entity relation diagram:



Updating the Database

As of version 0.4.0, Memex Explorer will start tracking all database migrations. This means that you will be able to upgrade your database and preserve the data without any issues.

If you are using a version that is 0.3.0 or earlier, and you are unable to update your database without server errors, the best course of action is to delete the existing file at *source/db.sqlite3* and start over with a fresh database.

3.1.4 Enabling Non-Default Services

Nutch Visualizations

Nutch visualizations are not enabled by default. Nutch visualizations require RabbitMQ, and the method for installing RabbitMQ varies depending on the operating system. RabbitMQ can be installed via Homebrew on Mac, and apt-get on Debian systems. For more information on how to install RabbitMQ, read this page. Note: You may also need to change the below command to *sudo rabbitmq-server*, depending on how RabbitMQ is installed on your system and the permissions of the current user.

RabbitMQ and Bokeh-Server are necessary for creating the Nutch visualizations. The Nutch streaming visualization works by creating and subscribing to a queue of AMQP messages (hosted by RabbitMQ) being dispatched from Nutch as it runs the crawl. A background task reads the messages and updates the plot (hosted by Bokeh server).

To enable Bokeh visualizations for Nutch, change *autostart=false* to *autostart=true* for both of these directives in *source/supervisord.conf*, and then kill and restart supervisor.

```
[program:rabbitmq]
command=rabbitmq-server
```

```
priority=1
-autostart=false
+autostart=true

[program:bokeh-server]
command=bokeh-server --backend memory --port 5006
priority=1
-autostart=false
+autostart=true
```

Domain Discovery Tool (DDT)

Domain Discovery Tool can be installed as a conda package. Simply run *conda install ddt* to download the package for DDT.

Like with Nutch visualizations, to enable DDT, change the directive in *source/supervisord*.

```
[program:ddt]
command=ddt
priority=5
-autostart=false
+autostart=false
```

Temporal Anomaly Detection (TAD)

TAD does not currently have a conda package. Like the Nutch visualizations, it also has a RabbitMQ dependency. For instructions on installing TAD, visit the github repository.

Like DDT and Nutch Visualizations, you also have to change the supervisor directive.

```
[program:tad]
command=tad
priority=5
-autostart=false
+autostart=false
```

CHAPTER 4

Manual Testing Guide

By following this guide, you will be able to test all the significant elements of the application. All of the files required for testing are in the repository under "source/test_resources".

4.1 Testing Projects

4.1.1 Project Creation

1. When you start up the application, you should see a landing page with a button for adding a new project.
   (a) Click the new project button.
   (b) Provide a name and a description for the project on the next page, and press submit.
   (c) Verify that your new project shows up on the project page list.
   (d) Click on the new project and go to the project page. Verify that there are no crawls, models, or datasets yet.

4.1.2 Project Settings

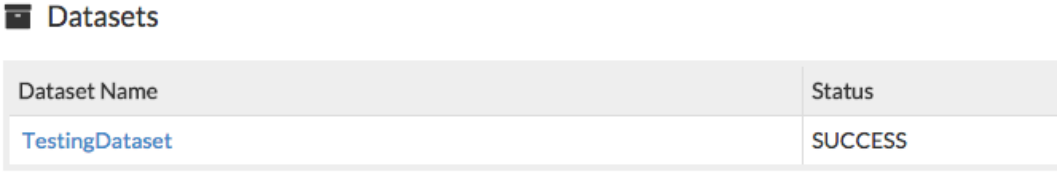
1. Click the "pencil" icon next to the name of the project on the project overview page.
   (a) Supply a different name and description for the project, and hit "submit".
   (b) Verify that the project was edited successfully by checking the success message at the top of the page.
2. Go back to the settings page.
   (a) Click on the "trashcan" icon. Verify that there is a popup asking you whether you want to delete the project.
   (b) Click on the trash icon and click yes.
   (c) Verify that you are taken to the landing page, and that there are no projects listed on the landing page.

4.2 Testing Indices

4.2.1 Index Creation

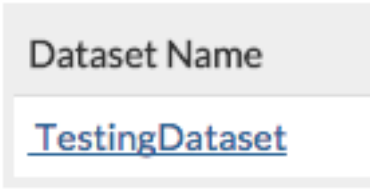
1. Create a new project.

2. Click on the "Add Index" button either in the sidebar or under the list of indices on the project page.
    (a) Add an index. Give the index a name and a zip file. There are two zipfiles in the repository to use, located at "source/resources/test_resources". Click submit.
    (b) Verify that the index was added successfully by checking for the success message at the top of the page.
    (c) Verify that the index was successfully created by checking the status next to the name of the index.



4.2.2 Index Settings

1. Click on the link to the index on the project overview page. This will take you to the index settings page.



    (a) Supply a new zipfile for the index creation. Use the zipfile that you did not use earlier – "sample2.zip" if you earlier used "sample.zip".
    (b) Verify that the index was updated successfully by checking the indices list.
    (c) Verify that the new files were added to the newly created index.
2. Return to the index settings page and click the "trashcan" icon. As before, confirm that the cancel button works, and then delete the index.
    (a) Confirm that the index was deleted successfully by looking at the list of indices on the project overview page.

4.3 Testing Seeds

At the navbar, click on the "Seeds" tab.

1. Create a Seeds List
    (a) Create a seeds list by providing a file.
    (b) Create another by pasting URLs into the textbox.
    (c) Paste in invalid URL into the textbox, and verify that it is highlighted red.
2. Edit a seeds list
    (a) Click on the icon for the seeds list to access the edit seeds page.

(b) Remove some URLs and click "Reset" to return to the original seeds list.

(c) Make one of the URLs invalid, and press "Save"

(d) Verify that the invalid URL is highlighted with red.

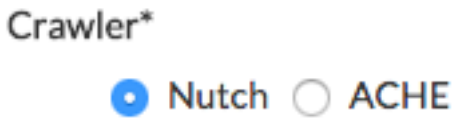
(e) Fix or remove the URL and click "Save"

4.4 Testing Crawls

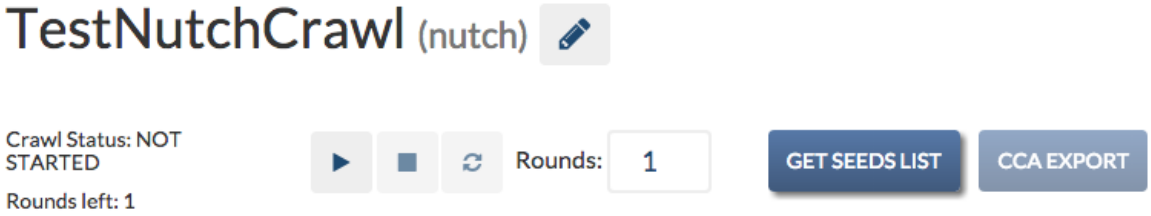
4.4.1 Testing Nutch Crawls

Included with the repository is a test seeds file. You can use this file to testing of nutch and ache crawls. The seeds file is located at "source/test_resources/test_crawl_data/cats.seeds".

1. From the project overview page, click the Add Crawl button on the list of crawls or in the sidebar dropdown.
2. At the add crawl page, supply a name and description.
   (a) Make sure that the "nutch" option is selected.



   (a) Select one of the previously created seed lists and create the crawl.
2. Verify that the crawl has been added successfully to the crawls list table.
3. Go to the crawl page by following the link in the crawls list table.
   (a) Verify that the crawl status and available buttons are the same as in this image.



   (a) The following buttons should be available: "Start Crawl", "Get Seeds List". All other buttons should be greyed-out.

   (b) The crawl status should be set to NOT STARTED with 0 rounds left to crawl.
2. Start a crawl and verify that the crawl completes successfully.
   (a) When you start the crawl, there should be two rounds left.

   (b) At the end of the first round, summary statistics should list total pages crawled as between 6 and 9.

   (c) After the first round is done, the status should show "SUCCESS" before going onto the next round.

   (d) On the start of the next round, the crawl status should change to "STARTED"

   (e) At the end of the second round, the rounds left should be zero.

   (f) The pages crawled should be between 300 and 400.

4.4.2 Test Crawl Settings

1. On the crawl page, click the "gears" icon to access the settings.
   (a) Change the name and description of the crawl, and submit.
   (b) Click the "trashcan" icon to delete the crawl.
   (c) Hit cancel on the popup first, and then delete the crawl.
   (d) Verify that you are brought to the project overview page.

CHAPTER 5

Glossary

Service Anything that provides an external functionality not included directly in Memex Explorer. Current examples include particular applications such as DDT, Tika, Kibana, and Elasticsearch.

Stack A particular set of Services in a working configuration. This term is not used frequently in the documentation.

Instance A version of Memex Explorer running on a given host as well as its associated stack and databases. An instance may have multiple projects.

Project An in-Memex Explorer data and application warehouse. Each project usually shares its application stack with other projects.

Domain Challenge A problem set like human trafficking, MRS, ebola.

Skin A particular UI (Text, CSS, etc...) on a particular webpage for a domain challenge

Celery A task manager implemented in Python which manages several tasks in Memex Explorer, including the crawlers.

Redis A key-value store database which used by Celery to keep information about task history and task queues.

Django A python web application framework. Django is the core of the Memex Explorer application.

Crawl Space Provides service for crawling the web using Nutch or Ache.

Task Manager Manages the application tasks, like running crawls. Task manager is not available from the Memex Explorer GUI interface.