# AdvancedAjaxInteraction

## AdvancedAjaxInteraction

This page provides commentary and thoughts on adapting Nutch not only to fetch AJAX/JavaScript driven dynamic HTML content, but also for interacting with that content (potentially a number of times) within a fetching scenario.



### Lets Begin with a Scenario

So lets say that as a Nutch crawl administrator your client has tasked you with the following ***"Get me domain specific material from a database such as NTIS"*** (NTIS; the National Technical Information Service, serves as the largest central resource for government-funded scientific, technical, engineering, and business related information available today.) What this really translates to is the following:

- use Nutch to log in to a database which requires HTTP POST authentication
- follow the redirect to the database landing query form
- submit a query to the form which will return a ranked list of search results for the given query
- interpret the JavaScript for each result in the ranked list
- use an HtmlParseFilter to obtain high level article/document content
- submit a GET request to invoke JavaScript which will return a PDF of the full textual content for this document
- return the full document (PDF) content and metadata along with the HTML parse filter data

### Crawling JavaScript/AJAX sites

In order to crawl webpages that rely on JavaScript/AJAX to dynamically load content you will want to use the Protocol-Selenium Plugin. This plugin will load the pages that you're crawling in Selenium so that JavaScript will be handled properly.

If you need to interact with the pages that you're crawling (E.g., JavaScript based pagination, clicking elements to dynamically load content) you will want to use the Protocol-InteractiveSelenium plugin. With this plugin you will create Handlers that interact with the pages in a defined way using the Selenium WebDriver interface. With this you'll be able to do any Selenium based interactions that you wish on a per-URL basis.

### Related Development Issues

- Nutch Selenium Plugin NUTCH-1933
  ** momer/nutch-selenium - This plugin allows you to fetch javascript pages using Selenium, while relying on the rest of the awesome Nutch stack! (ported to issue NUTCH-1933)
  ** momer/nutch-selenium-grid-plugin - This plugin allows you to fetch javascript pages using an existing Selenium Hub/Node set-up, while relying on the rest of the awesome Nutch stack!
- http://grid.selenium.googlecode.com/git-history/24150d2e97090b8b439bcc6a396911fb53200749/src/main/webapp/step_by_step_installation_instructions_for_osx.html - Installation instructions for Selenium Grid 2 on a Mac (needed for the momer/nutch-selenium-grid-plugin).
- http://grid.selenium.googlecode.com/git-history/00eae2a86d81c4ef8da355b0a8b916a9095a5cd9/src/main/webapp/download.html - latest version of Selenium Grid (Ver 1.0.8).

### Related Articles

- AJAX/JavaScript Enabled Parsing with Apache Nutch and Selenium

### FAQs

- How do I suppress Firefox from popping up during a Selenium crawl?
  1. Assign Firefox a particular space: Move Firefox to a dedicated space. Then, right-click on the Firefox icon in the Dock and go to Options > Assign To > This Desktop 2. Add the following key as <dict>'s children in /Applications/Firefox.app/Contents/Info.plist
     <key> LSBackgroundOnly </key>
     <string> True </string> 3. Quit Firefox. 4. Start crawling with Selenium. You will notice that Firefox will open silently in its own assigned space.