PubmedFetch

Overview

This tool is used to ingest data from the National Library of Medicine's PubMed publication repository using their EUtils web service. PubMed allows queries to be designed to return very specific record sets based on a range of different attributes such as date added, date modified, number range, affiliation, etc. See their help for using the PubMed search tool and also the advanced search tool, which helps you build your query.

The publications that match the query given to the tool will be fetched as XML data and loaded into the output RecordHandler.

Setup

Command Line Arguments

PubmedFetch extends NIHFetch, so it uses the same arguments.

Short Option	Long Option	Parameter Value Map	Description	Required
b	batchSize	NUMBER	number of records to fetch per batch	true
m	email	EMAIL_ADDRESS	your contact email address	true
n	numRecords	NUMBER	maximum records to return – set to ALL in order to retrieve all records without limit	true
o	output	CONFIG_FILE	RecordHandler config file path	true
O	outputOverride	override the RH_PARAM of output recordhandler using VALUE	false	
t	termSearch	SEARCH_STRING	term to search against pubmed	true

Configuration File

As with all the Harvester command-line tools, you can provide all the arguments as parameters is a configuration file (~~X~~-config). Here is a sample configuration for PubmedFetch.

```
<?xml version="1.0" encoding="UTF-8"?>
<Task>
  <Param name="email">sample.email@mydomain.tld</Param>
  <Param name="termSearch">sample AND edu[ad]</Param>
  <!-- these are set inside the example scripts, so are not needed
  <Param name="output">config/recordhandlers/tfrh.xml</Param>
  <Param name="outputOverride">fileDir=harvested-data/examples/pubmed</Param>
  -->
  <Param name="numRecords">ALL</Param>
  <Param name="batchSize">1000</Param>
</Task>
```

Execution

After version 1.2 of the harvester, execution of pubmedfetch has changed. The following refers to 1.1.1 and below:

To execute the PubmedFetch tool from the commandline, there is a convenient environment config file that, when loaded in a bash shell, will allow you to execute PubmedFetch with a simple `$PubmedFetch [args]`. For information about that, see Environment Config File.

Or you can simply call (paths relative to base harvester folder):

```
java -cp bin/harvester-<version>.jar:bin/dependency/* -Dprocess-task=PubmedFetch org.vivoweb.harvester.fetch.
nih.PubmedFetch
```

Design

See Design of PubmedFetch and its javadoc page