University of Florida PubMed Harvest

PubMed Harvester retrieves data from PubMed using web services, translates it into meaningful RDF which is then transferred into the VIVO model.

The Process

Extracting

In order to accomplish this part of the process, one will need to decide what data is to be extracted from PubMed.

The PubMed site (http://www.ncbi.nlm.nih.gov/pubmed) is very useful for making such decision and for constructing the search text for Harvester.

Example:

1. Go to http://www.ncbi.nlm.nih.gov/pubmed

2. Click on the "Advanced search" link

3. Search for publications that are linked to University of Florida and were published on June 01, 2011.

- Under the "Search Builder", select "Affiliation" and enter "University of Florida" into the text box. Then click on the "Add to Search Box" button.
- Select "Completion Date", and enter "2011/06/01" into both text boxes, meaning from 2011/06/01 to 2011/06/01. Then click on the "Add to Search Box" button.
- Copy the text from the "Search Box" and paste to a text editor for later use for Harvester.

(University of Florida[Affiliation]) AND "2011/06/01"[Completion Date] : "2011/06/01"[Completion Date]

- Click on the "Search" button.
- One should see results displayed on the PubMed site.

4. In this example, the above search text was used for Harvester. Hence, the default "termSearch" in the file /config/tasks/ufl.pubmedfetch.xml was replaced.

```
<?xml version="1.0" encoding="UTF-8"?>
<Task>
        <Param name="email">swilliams@ichp.ufl.edu</Param>
        <Param name="termSearch">(University of Florida[Affiliation]) AND "2011/06/01"[Completion Date] : "2011
/06/01"[Completion Date]</Param>
        <Param name="numRecords">ALL</Param>
        <Param name="batchSize">1000</Param>
</Task>
```

5. Edit the file /scripts/run-pubmed.sh

Un-comment out this line so that the script points to the task file ufl.pubmedfetch.xml for information about data extraction:

$PubmedFetch -X config/tasks/ufl.pubmedfetch.xml -o $H2RH -OdbUrl=$RAWRHDBURL

Transforming

The harvested data need to be mapped to the VIVO ontology. Since the initial harvest and the desired RDF/XML are both XML, mapping using XSL transformations seemed most appropriate. The details of those transformations had to be clear and distinct.

Translation

PubMed uses the Medline schema for storing citations. Medline is the Medical Literature Analysis and Retrival System Online (Medlars Online). This page shows details about what attributes from PubMed are transformed into what elements in VIVO's schema.

Scoring

Visit this page for general scoring methodology used by the Harvester. By default, PubMed Harvester uses two algorithms, EqualityTest and NormalizedLevenshteinDifference for scoring.

Matching

Visit this page for general matching methodology used by the Harvester.

Changing namespace

Get unmatched Authors into current namespace by modifying the file /scripts/run-pubmed.sh.

Uncomment this line to Execute ChangeNamespace to get unmatched Authors into current namespace:

$ChangeNamespace $CNFLAGS -u ${BASEURI}author/

Comment out this line:

$Qualify $MATCHEDINPUT -n ${BASEURI}author/ -c

Executing

Edit the file /config/models/vivo.xml, and modify "dbUrl", "dbUser", "dbPass", and "namespace" for your specific database settings and VIVO namespace.

Run /scripts/run-pubmed.sh to execute the process.

Lessons