Pubmed Example Script 1.2

Filename: example-scripts/full-harvest-examples/example-pubmed/run-pubmed.sh

This script fetches data from the Pubmed database and imports it into VIVO.

Steps

PubmedFetch. This step queries the Pubmed database, which returns an XML document. The query is specified in *pubmedfetch.config.xml*.

XSLTranslator. This is the step which transforms the Pubmed data into VIVO data. The Translate java program is called, using the XSLT mapping in *pubmed-to-vivo.datamap.xsl*. The RDF/XML files are placed in an output directory, defined in *xsltranslator.conf.xml*.

Transfer. In this step, the Transfer java program is called, which takes the VIVO RDF-XML output by the Translate program and stores it into an H2 model, which is used by Score, Match, Qualify, and ChangeNamespace.

Score. In this step, the Score program is called, which compares the input data with data already in VIVO based on certain parameters and marks up the data with numbers indicating how closely matched a given new record is with a given VIVO record.

Match. The Match program then looks over each of the scores, comparing them to a specified threshold value. If the scores are greater than the threshold value, Match will merge the data by linking the new data with the old data that it matched. Score and Match are called several times to match on different types of data.

Qualify. Clear out triples that are not supposed to be VIVO data but were created for scoring purposes.

ChangeNamespace. The file output by the XSLT file is internally linked in its own way. Each node has an identifier which consists of a namespace and an ID unique to just this file. ChangeNamespace will both change the namespace to the correct VIVO namespace, and then replace all the file-unique IDs with VIVO-unique IDs, creating new such IDs as necessary.

Diff. The new data is compared with data harvested in previous runs of the Pubmed harvester. Those triples that are present now but were not in the previous harvest are slated to be added to VIVO, and triples that were present in the previous harvest but not in the current one are slated to be removed from VIVO. The output of this step is required for the actual data import step, so if you want to prevent this calculation, delete or move the previous harvest directory *data/prevHarv*. Then every new triple from the harvest will be added to VIVO.

Transfer. The results of the Diff are used to add new data to VIVO and subtract removed data from VIVO. These changes are also applied to the previous harvest model in order to update it for the next harvest.