IP Example Script

Your first Harvest

Much of the work has been already done for your harvest but a few changes need to be done to the configuration so the harvest knows where your CTSA IP data is and records you wish to ingest.

- Change directory to example-scripts/bash-scripts/full-harvest-examples/example-ip
- Edit the raw-records.config.xml file
    - Set the fileDir parameter to folder containing the CTSA IP xml data file
- Edit the vivo.model.xml file
    - Set the dbURL, dbUser and dbPass
    - For more information on these parameters and their use, please see Harvester vivo configuration file
- Edit the run-ip.sh file and set the HARVESTER_INSTALL_DIR= to be the directory you unpacked the harvester in
- Run bash run-ip.sh
- Restart tomcat and apache2. You may also need to force the index to rebuild to see the new data. The index can be rebuilt by issuing the following URL in a browser:http://your.vivo.address/vivo/SearchIndex. This will require site admin permission, and prompt you to login if your not already.

The first run

Three folders will be created

- logs
- data
- previous-harvest

The logs folder contains the log from the run, the data folder contains the data from each run, and the previous-harvest folder contains the old harvest data for use during the update process at the end of the script. While you're testing, I would recommend treating each run as the first run (so no update logic will occur). You can do this by removing the previous-harvest folder before running again.

Inside the data folder, you will find the raw records utilized during the ingest. To see what rdf statements went into VIVO, you can view the vivo-additions.rdf.xml file. Conversely, to view what the harvester removed (because of updated data), you can view the vivo-subtractions.rdf.xml file. This file will be blank on your first run, since you have no previous harvest to compare the incoming data against.

Followup Runs and Queries

It is recommended to execute a single harvest and then run remove-last-ip-harvest.sh to remove the rdf each time you run a test harvest until your satisfied with the results. If then want to change the query to harvest more data in, make a duplicate copy of the example-ip folder and run the script from there (be sure you remove the previous-harvest folder before your first run).

Overview of run-ip.sh script

- Location of log file. If there is an issue with a harvest, this file proves invaluable in finding a solution to the problem.

```
echo "Full Logging in $HARVEST_NAME.$DATE.log"
if [ ! -d logs ]; then
  mkdir logs
fi
cd logs
touch $HARVEST_NAME.$DATE.log
ln -sf $HARVEST_NAME.$DATE.log $HARVEST_NAME.latest.log
cd ..
```

- Clear old data. To maintain data integrity, the previous information is removed.

```
rm -rf data
```

- Translate: The following will translate the input data into valid RDF.

```
 harvester-xsltranslator -X xsltranslator.config.xml
```

- To perform the additions and subtractions, first uncomment out code with the Find Addition and Find Subtractions

```
harvester-diff -X diff-additions.config.xml
harvester-diff -X diff-subtractions.config.xml
```

and then uncomment the Apply Subtractions to Previous model & Apply Additions to Previous model.

```
harvester-transfer -o previous-harvest.model.xml -r data/vivo-subtractions.rdf.xml -m
harvester-transfer -o previous-harvest.model.xml -r data/vivo-additions.rdf.xml
```

This will apply changes to the previous harvest. In order to apply changes to VIVO model, you will need to uncomment the the following :

```
harvester-transfer -o vivo.model.xml -r data/vivo-subtractions.rdf.xml -m
harvester-transfer -o vivo.model.xml -r data/vivo-additions.rdf.xml
```