Ex Libris – Rosetta Harvest

Harvest utility:

Options:

-a	login with auto user
-c	specify the collection code to use
-f	harvest from this date (yyyy-MM-dd)
-g	generate report to list items for a CONTENTdm collection not in Rosetta
-h	show help
-k	retention schedule
-l	sources for harvesting
-m	specify material-flow
-o	origin-field <arg> CDM ONLY! Specify which CONTENTdm field contains the file name
-p	specify producer
-r	specify rights
-s	start harvesting from this item index (not item ID)
-t	start harvest through this item index (not item ID)
-u	harvest from -f specified date until this date (yyyy-MM-dd)
-w	harvest metadata buy do not download the object(s)
-x	harvest metadata and objects but do not submit sip to Rosetta

Harvest report

- harvest –g cdm GEA status

The report is saved in /operational_shared/reports

Harvest collection in increments:

- -s (start index) and -t (through index) options with listing command
  - harvest -l -s 1 -t 1000 cdm GEA
  - harvest -l -s 1001 -t 2000 cdm GEA
- from date through date
  - harvest -f 2013-10-01 -u 2013-11-01

Origin field

- o identi cdm GeorgeBeard

The way it's setup is that it first looks in the "fullers" field. If "fullers" has a text node, the harvester uses that as the master file name (or in the case of the example below "MSS_P_3_0047a.jpg"). If "fullers" has no text node value, then the harvester uses the text node value from the "identi" field.

```
<fullrs name="Full resolution">Volume3\MSS_P_3_0047a.jpg</fullrs>
<identi dc="identifier" dcterms="identifier" name="Identifier">MSS_P_3_0047.jpg</identi>
<identi>PH500_fd57_item-1a.jpg; PH500_fd57_item-1a.jpg</identi>
<find>35931162992005_PH500_fd57_item-1a.jpg</find>
<dmrecord>535</dmrecord>
<fila>187321514112005_451_MssP24_B4_F4.jpg</fila>
```

Retention:

The "-k" flag for setting the retention code.
harvest -k 27441 cdm GeorgBeard 0

Sources to harvest from:

- cdm – CONTENTdm
- csv – spreadsheet

- ojs – Open Journal System
- ia – Internet Archive
-

The CDM harvester. Once a master file has been copied into a SIP it is moved from /operational_shared/cdm/objects/ to /operational_shared/cdm/objects/deposited

The logging feature in Rosetta: Each time you run the harvester a log file will be generated in /operational_shared/logs/ in the format yyyyMMddTHHmmss.log (ex. 20130917T095937.log).

CDM item xml is cleansed of invalid xml characters before creating the mets

The harvester, if an item exists in "/operational_shared/cdm/objects/" will handle this exception such that if the file already exists, it will be renamed "FILE_NAME.tif.[timestamp]" similar to the mets deposited folder function.

Also Renamed the "deposited" folder for master files to "harvested" since that more correctly describes where the associated SIP is at in the process when the master file is moved.

Harvesting Internet Archive:

auprintempsmelod00goun_scandata.xml
auprintempsmelod00goun_orig_jp2.tar
auprintempsmelod00goun_meta.xml
auprintempsmelod00goun_marc.xml
auprintempsmelod00goun_jp2.zip
auprintempsmelod00goun_files.xml
auprintempsmelod00goun_dc.xml
auprintempsmelod00goun.pdf
auprintempsmelod00goun.gif