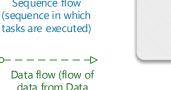






data from Data



storage to tasks)

















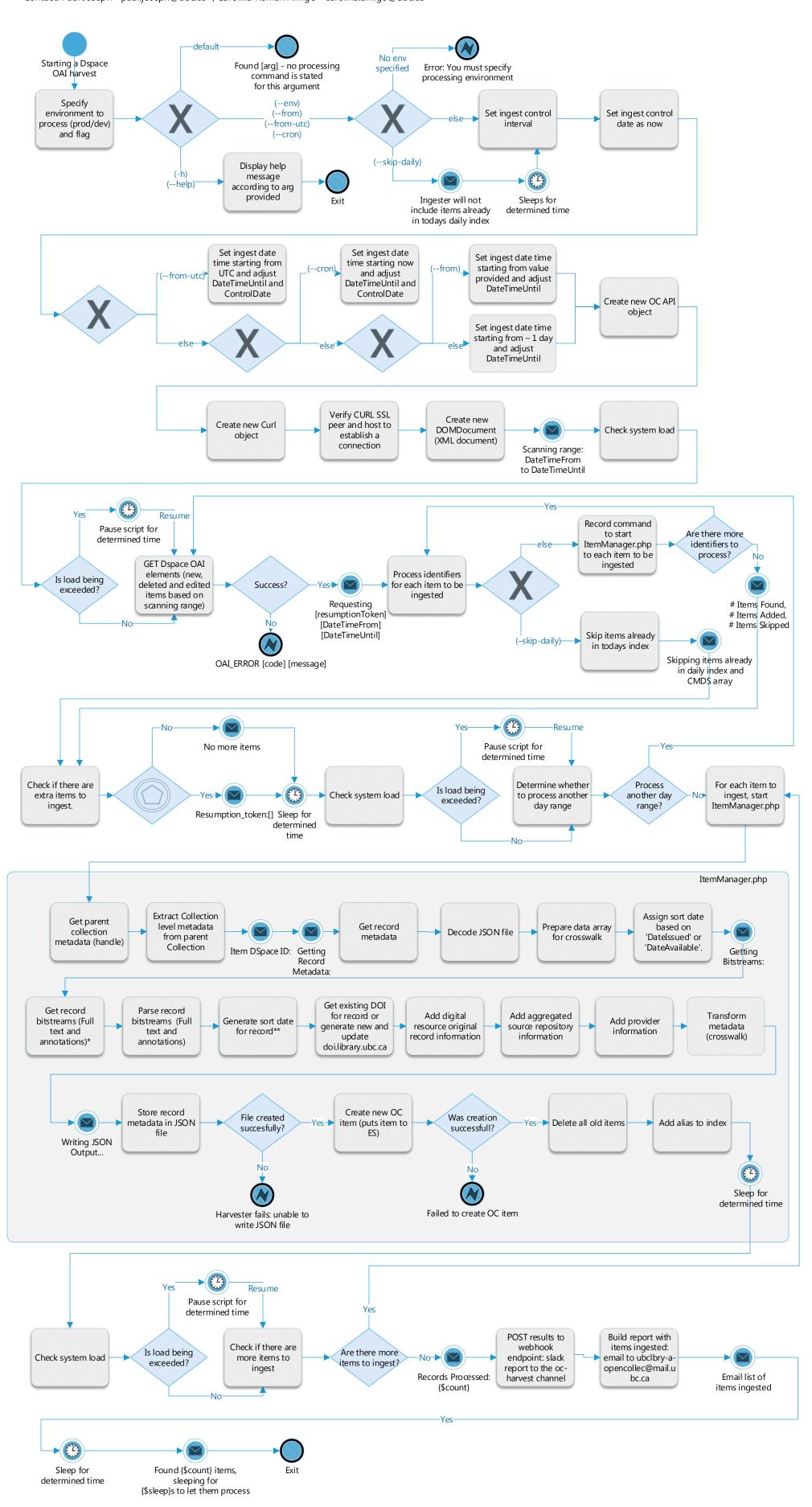


Parallel gateway (concurrent paths) Exclusive gateway (mutually

# **Harvesting Process Dspace – October 16<sup>th</sup> 2017**

Business Process Model and Notation (BPMN) diagram of the following code: https://git.library.ubc.ca/open-collections/harvester/blob/master/bin/dspace/process-collection-oai https://git.library.ubc.ca/open-collections/harvester/blob/master/src/DSpace/ItemManager.php

Contact: Paul Joseph <paul.joseph@ubc.ca>, Carolina Roman Amigo <carolina.amigo@ubc.ca>



\*NOTE: according to the latest cIRcle changes, embargoed bitstreams are not available for unauthorized users. OC account is not authorized.

\*\* In this format order: 'YYYY BC' or 'converts e.g. 16th century to 1500 AD' or 'coerce into YYYY-MM?-DD? AD and correct the incomplete dates'

# Legend



#### **Harvesting Process CONTENTdm - October 16<sup>th</sup> 2017** Business Process Model and Notation (BPMN) diagram of the following code: https://git.library.ubc.ca/open-collections/harvester/blob/master/bin/contentdm/process-collection https://git.library.ubc.ca/open-collections/harvester/blob/master/src/CONTENTdm/ItemManager.php Contact: Paul Joseph <paul.joseph@ubc.ca>, Carolina Roman Amigo <carolina.amigo@ubc.ca> No env specified Starting a Error: You must specify CONTENT dm harvest processing environment Ingesting Items within Create new OC API Determine max Specific Collection Specified object load Display help Specify message environment to (-h) according to arg process (prod/dev) -help) Reading all CONTENTdm provided and flag Collections (EXCEPT arphotos) -(--debug)-DEBUG\_MODE Verify CURL SSL Create new Is load being Create new Curl peer to establish a CONTENTdm API Verify system load exceeded? object connection object Pause script for determined time No Control Date equal Control date equal Create new ingest to date time control date to -2 provided Adding all items added after [control date] -else Look for JSON mapping file on OC GitHub repo For each collection -skip-daily)→( 🍑 get id for a specific Ingester will not Collection ID: [id] Sleeps for collection\* determined time include items already in todays daily index Skip collection File not found -else-Could not find a mapping file, you need to create one for the nick using the admin tool File found--(--debug)-Look for the items Check if item was list file /sync/ Is load being Decode JSON file ever ingested (is in else Verify system load {\$id}.json for a exceeded? any index) specific collection Pause script for determined time No -No items list found Skipping {\$id} - no items list found Skip items that are already in todays daily index SKIPPING\_IS\_ALREADY\_IN\_ TODAYS\_DAILY\_INDEX: {\$hdl} Skip item No In todays daily index Is date equal or Compare record Is in (no new item) Get record "date modified" more recent any index information (no -upsert) info with control than control 200 - EXISTING\_ITEM date. date? -Not in indexelse 404 - NEW\_ITEM\_NEVER\_INGESTED Ingest item (Submit item to elastic search) System deades skip Add item true 404 - SYSTEM\_DECIDES\_UPSERT DEBUG: did not submit item to elasticsearch For each item to (--upsert)ingest, start ItemManager.php 200/404 - USER\_DECIDES\_UPSERT Processing Item Handle: {\$id}.{\$hdl} | \$cmd ItemManager.php Resume Pause script for determined time Was it obtained Is load being Write daily log to a Was it obtained Check system load Get item nick\*\* Get item cisoptr\*\* exceeded? successfully? successfully? file Could not find a nick Could not find a cisoptr Get UBC internals Get record Find mapping File was located Create new OC API Request URL (repo, handle, metadata from (crosswalk) files successfully? object provenance nick) URL Item CONTENTdm ID: Getting Record {\$nickDotCisoptr} Metadata: Create file using admin tool parseItemMetadata parseItemMetadata Assign value "en" function starts function ends Yesto "language" Assign sort date based on 'Date' Prepare data array Find mapping File was located Is "language" -Yes Decode JSON file 'DateIssued' "ubc.date.sort" successfully? for crosswalk empty? 'DateAvailable'. Create file using admin tool No-Retrieve DOI for Is isShownAt Was it found? Look for cisoptr (DOI) empty? each record UPDATE\_DOI\_SYSTEM:{} Sleep for DOI from CONTENTdm: {\$doi} determined time Create new DOI Lookup DOI API object Cisoptr not found in CONTENTdm No DOI from CONTENTdm, need to get from System Create new DOI Update record Return existing API object with existing DOI DOI Sleep for determined time Add DOI Was DOI information to Update record Found DOI? Retum DOI Get new DOI IsShownAt obtained? with new DOI property could not get a doi Update record Return existing with existing DOI DOI Crosswalk function starts Add digital Add aggregated Add provider Get metadata resource original source repository application profile information record information information **METADATA** No value in field **ENTRIES START** Not empty metadata entry OR full text The property {} is not in the Metadata Application Profile Not empty Split crosswalked Return array with Perform crosswalk metadata for each value crosswalked values values METADATA **ENTRIES END** Crosswalk function ends Not empty metadata The property {} is not in the Metadata Application Profile Parse system internals (Digital Get image Call search Parse image Request URL information from identifier, internal Parse sort date Check system load transformationinformation URL handle, transcript, Getting image info etc.) ({\$nick}, {\$ciso}) Is it a Check for a Is load being Decode JSON file compoundexceeded? compound object object? Will find children: Processing Pause script for determined time Compound Object No -No-Not compound object, nothing more to do Get compound Is load being Request URL object information Check system load Resume exceeded? from URL Pause script for Getting Page: determined time -No Get each page Find mapping File was located Request URL metadata from Decode JSON file (crosswalk) files successfully? URL parseItemMetadata function starts Create file using admin tool parseItemMetadata Crosswalk Assign value "en" function ends Yesfunction starts to "language" Assign sort date based on 'Date' Prepare data array Is "language" Decode JSON file "ubc.date.sort" 'DateIssued' for crosswalk empty? 'DateCreated' or empty? 'DateAvailable'. METADATA **ENTRIES START** No-Crosswalk Get metadata function ends application profile No value in field Perform crosswalk Return array with Not empty metadata entry OR full text for each value crosswalked values METADATA **ENTRIES END** else The property {} is not in the Metadata Application Profile Parse system Get image internals (Digital Are there more Parse image Request URL child records to information from identifier, internal Parse sort date Yes information URL process? handle, transcript, Getting image info etc.) ({\$nick}, {\$ciso}) No All child records processed Resume Pause script for determined time JSON file Is load being Get full text Verify SortDate Check system load exceeded? exists? Inserting Into Writing JSON DailyIndex: Output... -Yes -No Find existing Create JSON file File writing Create new OC Was creation instances of the with metadata successfull? item and schedule item successfull? contents Remove JSON file for deletion sleep for determined time Filename:{} No Harvester fails: unable to Write log write JSON file Failed to create Send error message \$instance != \$dailyIndex DELETED\_OLD\_ENTRY (\_index) Remove dailyLog Remove JSON file (\_id) file MANUALLY\_DELETE CACHE\_BUSTING:{} IGNORING {\$instance['\_index']} Add alias to index Sleep for

Yes List duplicates using JSON encoding

\*\*Nick: nickname assigned to an item as an unique identifier. Cisoptr: numerical ID for CONTENTdm. A combination of nick+cisoptr retrieves one item.

Processor Found {\$skipped}

items that were already

in the daily index

Count records

found

(cisoptr)

Look for duplicates

(count > 1)

goodbye

Processor Found

{\$cisoCount}

records

Finished Looping

through Collection: {\$id}

Processor Found

{\$cisoCount} duplicate records

Exit

determined time

Is load being

exceeded?

\*Path example: /../../vendor/ubc-library/harvester-contentdm-crosswalks/{\$\_n}.json

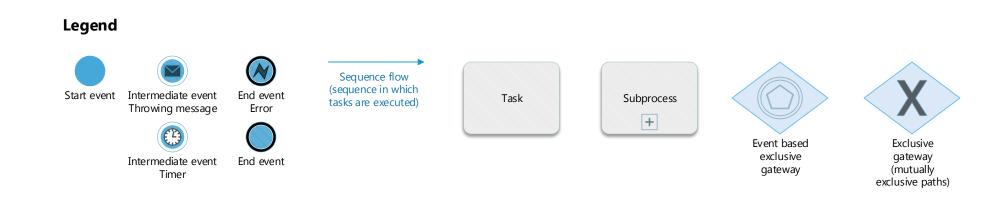
Pause script for

determined time

No-

Verify system load

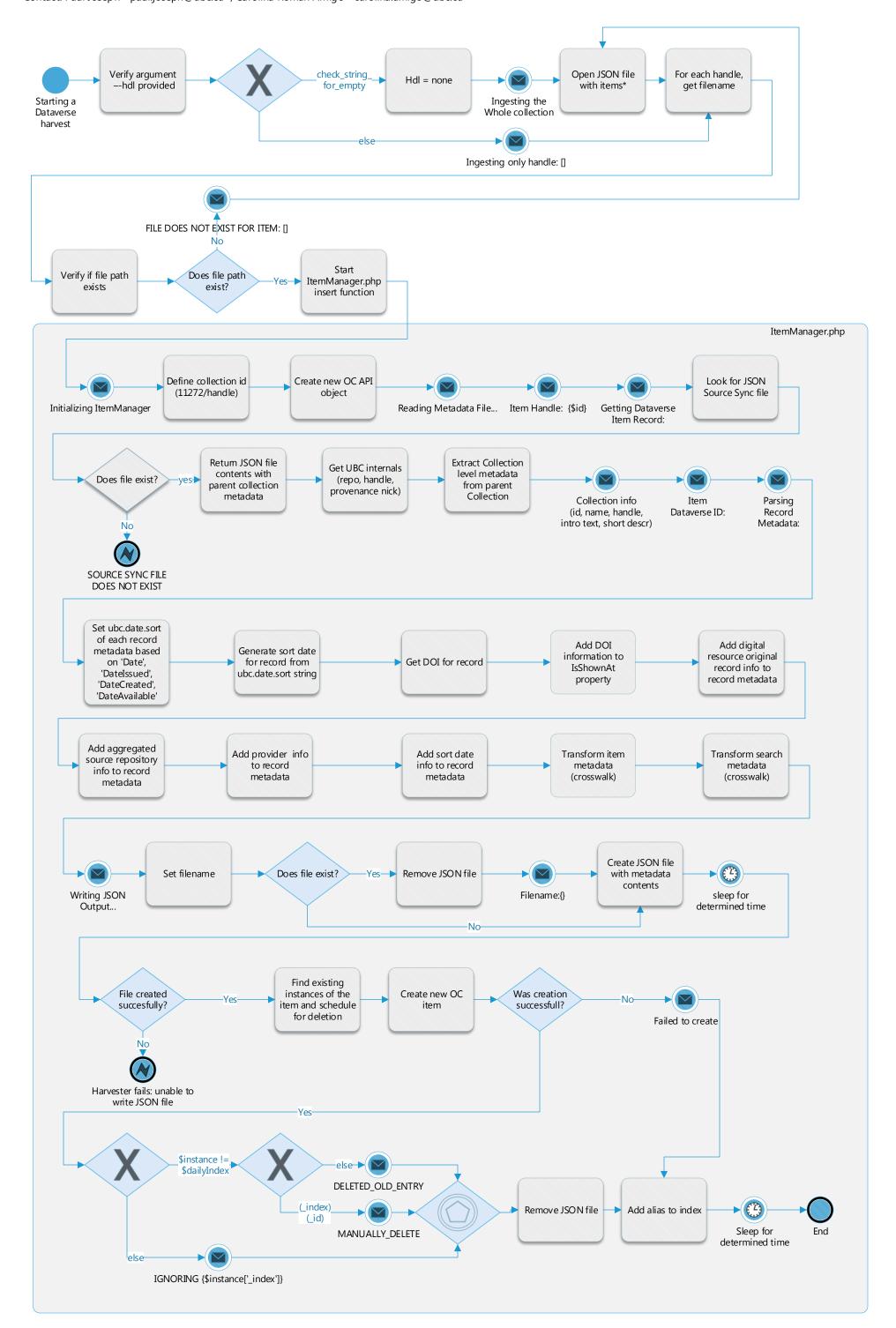
Are there duplicates?



### **Harvesting Process Dataverse - October 16<sup>th</sup> 2017**

Business Process Model and Notation (BPMN) diagram of the following code: https://git.library.ubc.ca/open-collections/harvester/blob/master/bin/dataverse/process-collection.py https://git.library.ubc.ca/open-collections/harvester/blob/master/src/Dataverse/ItemManager.php

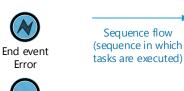
Contact: Paul Joseph <paul.joseph@ubc.ca>, Carolina Roman Amigo <carolina.amigo@ubc.ca>



# Legend



Timer











<sup>\*&#</sup>x27;build-collections-items-list.py' script gets items metadata using dataverse OAI feed. It maps item metadata to OC format and stores it a JSON file.