

TextCritical.net - Bug #439

Feature # 403 (Closed): Perseus Book Importer

Importer fails on works with no state set

11/14/2012 06:39 AM - Luke Murphey

Status:	Closed	Start date:	
Priority:	High	Due date:	
Assignee:	Luke Murphey	% Done:	100%
Category:		Estimated time:	0.00 hour
Target version:	0.1		
Description			
The following works have no state sets and cannot be imported:			
<div>1 aristoph.wasps_gk.xml</div> <div>2 aristoph.thes_gk.xml</div> <div>3 aristoph.pl_gk.xml</div> <div>4 aristoph.peace_gk.xml</div> <div>5 aristoph.lys_gk.xml</div> <div>6 aristoph.kn_gk.xml</div> <div>7 aristoph.frogs_gk.xml</div> <div>8 aristoph.eccl_gk.xml</div> <div>9 aristoph.cl_gk.xml</div> <div>10 aristoph.birds_gk.xml</div> <div>11 aristoph.ach_gk.xml</div>			
<div>reader.importer.PerseusBatchImporter: Exception generated when attempting to process file="aristoph.wasps_gk.xml"</div> <div>Traceback (most recent call last):</div> <div>File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/PerseusBatchImporter.py", line 326, in process_directory</div> <div>if self.__process_file__(os.path.join(root, f)):</div> <div>File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/PerseusBatchImporter.py", line 277, in __process_file__</div> <div>return self.process_file(file_path, document_xml, title, author, language)</div> <div>File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/PerseusBatchImporter.py", line 453, in process_file</div> <div>perseus_importer.import_file(file_path)</div> <div>File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/Perseus.py", line 139, in import_file</div> <div>return self.import_xml_document(doc)</div> <div>File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/site-packages/django/db/transaction.py", line 209, in inner</div> <div>return func(*args, **kwargs)</div> <div>File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/Perseus.py", line 571, in import_xml_document</div> <div>current_state_set = state_sets[self.state_set]</div> <div>IndexError: list index out of range</div>			

History

#1 - 11/14/2012 06:41 AM - Luke Murphey

- Description updated
- Parent task set to #403

#2 - 11/15/2012 07:09 AM - Luke Murphey

- Status changed from New to In Progress

#3 - 11/15/2012 07:16 AM - Luke Murphey

- *Description updated*

The following methods of PerseusTextImporter are likely to need to be modified to accept documents that do not have state sets:

- is_milestone_chunk
- get_state_for_milestone
- is_milestone_in_state_set

#4 - 11/15/2012 07:38 AM - Luke Murphey

Looks like this was a problem with the import policy which indicated that state set one ought to be used (which doesn't exist).

#5 - 11/15/2012 06:09 PM - Luke Murphey

- *Status changed from In Progress to Closed*

- *% Done changed from 0 to 100*