TextCritical.net - Bug #446

Feature # 403 (Closed): Perseus Book Importer

Fix issue where some works import no divisions or verses

11/17/2012 06:38 PM - Luke Murphey

Status:	Closed	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	0.1		

Description

Associated revisions

Revision 172 - 11/18/2012 07:11 AM - Luke Murphey

Chapters are now treated as chunks. Reference #446.

Revision 172 - 11/18/2012 07:11 AM - Luke Murphey

Chapters are now treated as chunks. Reference #446.

Revision 165 - 11/18/2012 07:11 AM - Luke Murphey

Chapters are now treated as chunks. Reference #446.

Revision 178 - 11/19/2012 04:22 AM - Luke Murphey

Fixed issue where the importer did not import some works which had a head tag. Reference #446.

Revision 178 - 11/19/2012 04:22 AM - Luke Murphey

Fixed issue where the importer did not import some works which had a head tag. Reference #446.

Revision 171 - 11/19/2012 04:22 AM - Luke Murphey

Fixed issue where the importer did not import some works which had a head tag. Reference #446.

Revision 179 - 11/19/2012 05:26 AM - Luke Murphey

Changed the import policy to account to allow import of plut.068_teubner_gk.xml. Reference #446.

Revision 179 - 11/19/2012 05:26 AM - Luke Murphey

Changed the import policy to account to allow import of plut.068_teubner_gk.xml. Reference #446.

Revision 172 - 11/19/2012 05:26 AM - Luke Murphey

Changed the import policy to account to allow import of plut.068_teubner_gk.xml. Reference #446.

History

#1 - 11/18/2012 12:09 AM - Luke Murphey

The following works are affected:

```
1 xen.mem_gk.xml
2 xen.hell_gk.xml
3 xen.cyrop_gk.xml
4 xen.anab_gk.xml
5 char_gk.xml
6 idylls_gk.xml
```

```
7 plut.083_loeb_gk.xml
8 plut.068_teubner_gk.xml
9 pind_gk.xml
10 paus_gk.xml
11 nonnos_03.xml
12 nonnos_02.xml
13 nonnos_01.xml
14 lys_gk.xml
15 lyc_gk.xml
16 71_gk.xml
17 hyp_gk.xml
18 hp.littre_gk.xml
19 dl_gk.xml
20 diochr01_gk.xml
21 din_gk.xml
22 dem51-61_gk.xml
23 dem41-50_gk.xml
24 dem31-40_gk.xml
25 dem21-30_gk.xml
26 dem11-20_gk.xml
27 dem01-10_gk.xml
28 demad_gk.xml
29 call_02.xml
30 nt_gk.xml
31 bacchyl_gk.xml
32 aret_gk.xml
33 andoc_gk.xml
```

#2 - 11/18/2012 05:04 AM - Luke Murphey

The problem is that these works are chunked by chapters but the chapters are not indicated as chunks so the importer doesn't know to break them into divisions.

#3 - 11/18/2012 05:04 AM - Luke Murphey

- % Done changed from 0 to 50

#4 - 11/18/2012 07:12 AM - Luke Murphey

- % Done changed from 50 to 70

#5 - 11/18/2012 03:18 PM - Luke Murphey

Making chapters chunks reduced the number of works with this error to 27 (from 33).

Theophrastus' Characters is not importing because it has no verses. The division content looks correct but the milestones are not being detected. Below is an example that does not import:

```
<chapter>
<head>*ei)rwnei/as *a'</head>
<p><milestone n="1" unit="section"/>*(h me\n ou)=n ei)rwnei/a do/ceien a)\n ei)=nai, w(s tu/pw| labei=n,
prospoi/hsis e)pi\ xei=ron pra/cewn kai\ lo/gwn, o( de\ ei)/rwn
<milestone n="2" unit="section"/>toiou=to/s tis, oi(=os proselqw\n toi=s e)xqroi=s e)qe/lein lalei=n,
ou) misei=n: kai\ e)painei=n paro/ntas, oi(=s e)pe/qeto la/qra, kai\
tou/tois sullupei=sqai h(ttwme/nois: kai\ suggnw/mhn de\ e)/xein
toi=s au(to\n kakw=s le/gousi kai\ e)pi\ toi=s kaq' e(autou= legome/nois.
<milestone n="3" unit="section"/>kai\ pro\s tou\s a)dikoume/nous kai\ a)ganaktou=ntas pra/ws
diale/gesqai: kai\ toi=s e)ntugxa/nein kata\ spoudh\n boulome/nois
<milestone n="4" unit="section"/>prosta/cai e)panelqei=n. kai\ mhde\n w(=n pra/ttei o(mologh=sai,
a)lla\ fh=sai bouleu/esqai: kai\ prospoih/sasqai a)/rti paragegone/nai
<milestone n="5" unit="section"/>kai\ o)ye\ gene/sqai au)to\n kai\ malakisqh=nai. kai\
pro\s tou\s daneizome/nous kai\ e)rani/zontas . . . w(s ou) pwlei=, kai\
mh\ pwlw=n fh=sai pwlei=n: kai\ a)kou/sas ti mh\ prospoiei=sqai,
kai\ i)dw\n fh=sai mh\ e(orake/nai, kai\ o(mologh/sas mh\ memnh=sqai:
kai\ ta\ me\n ske/yasqai fa/skein, ta\ de\ ou)k ei)de/nai, ta\ de\
<pb/>qauma/zein, ta\ d' h)/dh pote\ kai\ au)to\s ou(/tws dialogi/sasqai.
</p>

</chapter>
```

#6 - 11/18/2012 03:36 PM - Luke Murphey

The problem has to do with failing to recurse down nodes after head nodes. The recurse flag gets changed for all following nodes. I'm surprised that this didn't prevent importation of other works. All unit tests pass after fixing this though.

#7 - 11/19/2012 04:35 AM - Luke Murphey

The following two works are not being imported:

```
1    plut.068_teubner_gk.xml
2    71_gk.xml
```

The error is:

```
2012-11-18 11:03:52,864 [ERROR] reader.importer.PerseusBatchImporter: Exception generated when attempting to p
rocess file="plut.068_teubner_gk.xml"
Traceback (most recent call last):
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/PerseusBatchImporter.py",
line 326, in process_directory
    if self.__process_file__( os.path.join( root, f) ):
```

```
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/PerseusBatchImporter.py",
line 277, in __process_file__
    return self.process_file(file_path, document_xml, title, author, language)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/PerseusBatchImporter.py",
line 455, in process_file
    perseus_importer.import_file(file_path)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/Perseus.py", line 143, in
import_file
    return self.import_xml_document(doc)
  File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/site-packages/django/db/transaction.py
", line 209, in inner
    return func(*args, **kwargs)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/importer/Perseus.py", line 643, in
import_xml_document
    raise Exception("No divisions were discovered, title=%s" % (self.work.title) )
Exception: No divisions were discovered, title=Quomodo adolescens poetas audire debeat
```

#8 - 11/19/2012 05:26 AM - Luke Murphey

- % Done changed from 70 to 90

#9 - 11/22/2012 09:19 AM - Luke Murphey

71_gk.xml (Podagra) is the only file that cannot be imported (complains that not divisions were discovered).

#10 - 11/23/2012 06:27 PM - Luke Murphey

All works successfully import now. The import took 4606 seconds (76 minutes) and the resulting file is 330 MB.

#11 - 11/23/2012 06:27 PM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 90 to 100