TextCritical.net - Feature #403

Perseus Book Importer

09/02/2012 03:43 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

Create functions necessary to import Perseus works (TEI).

See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html for information about the TEI format. Also, see http://etext.lib.virginia.edu/tei/uvatei3.html for a gentle introduction to the format.

Subtasks:

Bug # 440: Importer fails on works with milestone that do not have an indicator	**Closed**
Bug # 439: Importer fails on works with no state set	**Closed**
Bug # 438: Importer fails on works with no biblStruct	**Closed**
Bug # 437: Importer fails on divisions with no type	**Closed**
Bug # 446: Fix issue where some works import no divisions or verses	**Closed**
Bug # 451: Various import policy changes	**Closed**
Bug # 452: Review all works	**Closed**
Bug # 459: Some works get imported with empty verses	**Closed**

Related issues:

Related to TextCritical.net - Bug #416: Perseus Importer Doesn't Handle Secti...	**Closed**	**10/17/2012**
Related to TextCritical.net - Bug #420: Perseus Importer Gets Nodes Out of Order	**Closed**	

Associated revisions

Revision 164 - 11/17/2012 05:02 PM - Luke Murphey

Updated the scan policy:

- Euripides now ignores division markers
- Sophocles' Sophocles uses line counts for division markers

Reference #403.

Revision 164 - 11/17/2012 05:02 PM - Luke Murphey

Updated the scan policy:

- Euripides now ignores division markers
- Sophocles' Sophocles uses line counts for division markers

Reference #403.

Revision 157 - 11/17/2012 05:02 PM - Luke Murphey

Updated the scan policy:

- Euripides now ignores division markers

- Sophocles' Sophocles uses line counts for division markers

Reference #403.

History

#1 - 09/08/2012 05:16 AM - Luke Murphey

Here is an analysis of the TEI document

Sections Header

The header has a list of the sections chunks:

```
    <encodingDesc>
      <refsDecl doctype="TEI.2">
        <state delim="." unit="book"/>
        <state unit="section"/>
      </refsDecl>
      <refsDecl doctype="TEI.2">
        <state delim="." unit="book"/>
        <state delim="." n="chunk" unit="Whiston chapter"/>
        <state unit="Whiston section"/>
      </refsDecl>
    </encodingDesc>
```

Sections

The sections are broken up by milestones:

```
<milestone n="26" unit="section"/>tre/yomai de\ e)pi\ th\n a)fh/ghsin h)/dh tw=n pragma/twn mnhsqei\s pro/tero
n w(=n peri\ th=s tou= ko/smou kataskeuh=s <pb/>
ei)=pe *mwush=s: tau=ta d' e)n tai=s i(erai=s bi/blois eu(=ron a)nagegramme/na.e)/xei de\ ou(/tws:

<milestone n="1" unit="Whiston chapter"/>
<milestone n="1" unit="Whiston section"/>
<milestone n="27" unit="section"/></p><p>*)en a)rxh=| e)/ktisen o( qeo\s to\n ou)rano\n kai\ th\n gh=n. tau/th
s
d' u(p' o)/yin ou)k e)rxome/nhs, a)lla\ baqei= me\n kruptome/nhs sko/tei,
pneu/matos d' au)th\n a)/nwqen e)piqe/ontos, gene/sqai fw=s e)ke/leusen
o( qeo/s.
```

Books

The div1 node contains the book. The first head contains the name of the book:

```
<div1 type="Book" n="1">
    <note anchored="yes" type="title" place="inline">*prooi/mion peri\ th=s o(/lhs pragmatei/as.
       <list type="toc">
           <head>*ta/de e)/nestin e)n th=| prw/th| tw=n *)iwsh/pou i(storiw=nth=s *)ioudai+kh=s a)rxaiologi/as
.</head>
               <label><num>a</num>.</label><item>h( tou= ko/smou su/stasis kai\ dia/tacis tw=n stoixei/wn.</it
em>
               <label><num>b</num>.</label><item>peri\ tou= ge/nous *)ada/mou kai\ tw=n a)p' au)tou= de/ka gen
ew=n tw=n me/xri tou= kataklusmou=.</item>
```

#2 - 09/10/2012 08:26 PM - Luke Murphey

- Status changed from New to In Progress

#3 - 09/12/2012 03:39 PM - Luke Murphey

Note that if we parse and store individual words *and* allow multiple types of sectioning, then we word need to make sure that same references to the words are used so that we don't have to do the lemmatization twice.

#4 - 09/14/2012 12:03 AM - Luke Murphey

I added a Django manage.py command that allows to import Perseus files from the command-line:

```
python .\manage.py import_perseus --file C:\Users\Luke\Downloads\TextCritical\Perseus\Classics\Josephus\opensource\j.vit_gk.xml
```

#5 - 09/14/2012 03:21 AM - Luke Murphey

The Westcott NT is one of the more complex texts and has a few things that ought to be considered.

encodingDesc

It includes a <encodingDesc> tag seems to have information necessary for adjusting rendering:

```
   <tagsDecl>
    <rendition id="large"> </rendition>
    <rendition id="blockquote"> </rendition>
    <rendition id="header"> </rendition>
    <tagUsage render="header" gi="head"> </tagUsage>
    <tagUsage render="blockquote" gi="q"
     >     Marks poetical language. within the running text.

    </tagUsage>
    <tagUsage render="large" gi="quote"
     >     Marks material quoted from other biblical texts.

    </tagUsage>
   </tagsDecl>
```

para milestones

The text also marks paragraphs:

```
<milestone n="5" unit="verse"/>
     <milestone unit="para"/>*salmw\n de\ e)ge/nnhsen to\n *boe\s e)k th=s *(raxa/b,

<milestone
      unit="para"/>*boe\s de\ e)ge/nnhsen to\n *)iwbh\d e)k th=s *(rou/q,

<milestone unit="para"
      />*)iwbh\d de\ e)ge/nnhsen to\n *)iessai/,
```

These para units are not called out the in the refsDecl:

```
   <refsDecl doctype="TEI.2">
    <state delim=" " unit="book"/>
    <state delim=":" unit="chapter"/>
    <state unit="verse"/>
   </refsDecl>
```

#6 - 09/14/2012 05:16 AM - Luke Murphey

- Subject changed from Perseus Importer to Perseus Book Importer

#7 - 09/14/2012 05:26 AM - Luke Murphey

Some of the texts (such as Plato in Twelve Volumes, Vol. 1) alternate between speakers:

```
<castList><castItem type="role"><role>Euthyphro</role></castItem><castItem type="role"><role><placeName key="t
gn,2674867">Socrates</placeName></role></castItem></castList>
<milestone unit="page" n="2"/><milestone n="2a" unit="section"/><sp><speaker>Euthyphro</speaker><p>What strang
e thing has happened, <placeName key="tgn,2674867">Socrates</placeName>, that you have left your accustomed ha
unts in the Lyceum and are now haunting the portico where the king archon sits?  For it cannot be that you hav
e an action before the king, as I have.</p></sp><sp><speaker>Socrates</speaker><p>Our Athenians, Euthyphro, do
 not call it an action, but an indictment.</p></sp><sp><speaker>Euthyphro</speaker><p>What?  Somebody has, it
seems, brought an indictment against you;
<milestone n="2b" unit="section"/>for I don't accuse you of having brought one against anyone else.</p></sp><s
p><speaker>Socrates</speaker><p>Certainly not.</p></sp><sp><speaker>Euthyphro</speaker><p>But someone else aga
inst you?</p></sp><sp><speaker>Socrates</speaker><p>Quite so.</p></sp><sp><speaker>Euthyphro</speaker><p>Who i
s he?</p></sp><sp><speaker>Socrates</speaker><p>I don't know the man very well myself, Euthyphro, for he seems
 to be a young and unknown person.  His name, however, is Meletus, I believe.  And he is of the deme of Pitthu
s, if you remember any Pitthian Meletus, with long hair and only a little beard, but with a hooked nose.</p></
sp><sp><speaker>Euthyphro</speaker><p>I don't remember him, Socrates.  But
<milestone n="2c" unit="section"/>what sort of an indictment has he brought against you?</p></sp><sp><speaker>
Socrates</speaker><p>What sort?  No mean one, it seems to me;  for the fact that, young as he is, he has appre
hended so important a matter reflects no small credit upon him.  For he says he knows how the youth are corrup
ted and who those are who corrupt them.  He must be a wise man;  who, seeing my lack of wisdom and that I am c
orrupting his fellows, comes to the State, as a boy runs to his mother, to accuse me.  And he seems to me to b
e the only one of the public men who begins in the right way;  for the right way
<milestone n="2d" unit="section"/>is to take care of the young men first, to make them as good as possible, ju
st as a good husbandman will naturally take care of the young plants first and afterwards of the rest.  And so
 Meletus, perhaps, is first
<milestone unit="page" n="3"/><milestone n="3a" unit="section"/>clearing away us who corrupt the young plants,
 as he says;  then after this, when he has turned his attention to the older men, he will bring countless most
 precious blessings upon the State,&mdash;at least, that is the natural outcome of the beginning he has made.<
/p></sp><sp><speaker>Euthyphro</speaker><p>I hope it may be so, Socrates;  but I fear the opposite may result.
  For it seems to me that he begins by injuring the State at its very heart, when he undertakes to harm you.
Now tell me, what does he say you do that corrupts the young?
<milestone n="3b" unit="section"/></p></sp><sp><speaker>Socrates</speaker><p>Absurd things, my friend, at firs
t hearing.  For he says I am a maker of gods;  and because I make new gods and do not believe in the old ones,
 he indicted me for the sake of these old ones, as he says.</p></sp><sp><speaker>Euthyphro</speaker><p>I under
stand, Socrates;  it is because you say the divine monitor keeps coming to you.  So he has brought the indictm
ent against you for making innovations in religion, and he is going into court to slander you, knowing that sl
anders on such subjects are readily accepted by the people.  Why, they even laugh at me and say I am crazy
<milestone n="3c" unit="section"/>when I say anything in the assembly about divine things and foretell the fut
ure to them.  And yet there is not one of the things I have foretold that is not true;  but they are jealous o
f all such men as you and I are.  However, we must not be disturbed, but must come to close quarters with them
.</p></sp><sp><speaker>Socrates</speaker><p>My dear Euthyphro, their ridicule is perhaps of no consequence.  F
or the Athenians, I fancy, are not much concerned, if they think a man is clever, provided he does not impart
 his clever notions to others;  but when they think he makes others to be like himself,
<milestone n="3d" unit="section"/>they are angry with him, either through jealousy, as you say, or for some ot
her reason.</p></sp><sp><speaker>Euthyphro</speaker><p>I don't much desire to test their sentiments toward me
 in this matter.</p></sp><sp><speaker>Socrates</speaker><p>No, for perhaps they think that you are reserved and
 unwilling to impart your wisdom.  But I fear that because of my love of men they think that I not only pour m
yself out copiously to anyone and everyone without payment, but that I would even pay something myself, if any
one would listen to me.  Now if, as I was saying just now, they were to laugh at me, as you say they do at you
, it would not be at all unpleasant
<milestone n="3e" unit="section"/>to pass the time in the court with jests and laughter;  but if they are in e
arnest, then only soothsayers like you can tell how this will end.</p></sp><sp><speaker>Euthyphro</speaker><p>
Well, Socrates, perhaps it won't amount to much, and you will bring your case to a satisfactory ending, as I t
hink I shall mine.</p></sp><sp><speaker>Socrates</speaker><p>What is your case, Euthyphro? Are you defending o
r prosecuting?</p></sp><sp><speaker>Euthyphro</speaker><p>Prosecuting.</p></sp><sp><speaker>Socrates</speaker>
<p>Whom?
<milestone unit="page" n="4"/><milestone n="4a" unit="section"/></p></sp><sp>
```

Oddly, this document also claims to contain two languagse though it clearly contains only English:

```
<langUsage>
<language id="en">English</language>
<language id="greek">Greek</language>
</langUsage>
```

#8 - 09/14/2012 06:29 AM - Luke Murphey

I am getting the faint feeling that I am doing this wrong. I think the best approach would be to break up the XML into chunks and simply save the chunks as verses. Then I could choose how to render the text later. This would be useful because I could add support for other works later on.

Perhaps I could convert the tags to span tags with a class that includes the tagname. Something like <speaker>Euthyphro</speaker> could be stored as Euthyphro. Whatever work is done at view time could be cached to reduce performance impact.

#9 - 09/15/2012 08:31 AM - Luke Murphey

I'm updating the model to support late-binding. In this design, the imports will work like this:

1. Extract the meta-data from the work
2. Chunk the work into chapters; store XML segments in chapter
3. Build sections index
4. Extract verses from chapter XML segments (or process into viewable content at view time)

This way, I only have to get the chunking right. After that, I can rebuild the verse content as I need to support XML nodes necessary for rendering.

#10 - 09/19/2012 01:06 AM - Luke Murphey

Here is a good introduction to the TEI document format: http://etext.lib.virginia.edu/tei/uvatei3.html

#11 - 09/19/2012 01:07 AM - Luke Murphey

- Description updated

#12 - 09/19/2012 05:52 AM - Luke Murphey

Still need to handle divisions. According to http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html#DRDIV the book can be broken into numbered (div, div2, etc.) or unnumbered divs (div).

#13 - 09/21/2012 08:30 AM - Luke Murphey

- Description updated

#14 - 09/22/2012 08:11 AM - Luke Murphey

I am working to import the various works of Josephus. Below are the works that I have successfully imported (along with the stateset used):

- Contra Apione: *
- De bello Judaico libri: 1
- Josephi vita: *
- Antiquitates Judaicae: 1

#15 - 09/23/2012 02:46 AM - Luke Murphey

Antiquitates Judaicae 2 verse 1 seems to include an error. The first verse is: "□περ" which should presumably be "υπερ".

#16 - 09/23/2012 03:25 AM - Luke Murphey

May want to consider converting the content to HTML5 Microdata

#17 - 09/30/2012 12:18 AM - Luke Murphey

Aratus' Phaenomena does not include chapter objects but includes lb objects instead:

```
<body>
   <div1 type="book" n="1">
   <p><pb id="p.380"/>
   <lb type="displayNum" n="1"/>e)k *dio\s a)rxw/mesqa, to\n ou)de/pot' a)/ndres e)w=men
   <lb type="displayNum" n="2"/>a)/rrhton: mestai\ de/ *dio\s pa=sai me\n a)guiai/,
```

The lb instances are linebreaks per http://www.tei-c.org/release/doc/tei-p5-doc/en/html/examples-lb.html.

#18 - 10/15/2012 05:31 AM - Luke Murphey

I'm going to need a better method for processing the original contents for viewing. These are the options:

XLST: use the browser to process the original content into something viewable

- Con: may not work on all browsers
- Con: difficult to employ without using AJAX
- Pro: clean separation between importation and viewing

Microdata: convert content into XML that can be rendered by a browser using Microdata

- Con: will lose some resolution since not all of the data can be represented in Microdata
- Pro: fairy straightforward conversion from original content into something renderable

Markdown: convert the content into Markdown (or something similar) that can be rendered

- Con: will lose some data since not everything can be readily rendered in Microdata
- Pro: the content in the tables will be straightforward to render

Rich HTML: convert the HTML directly into HTML

- Pro: does not lose data
- Con: means that rendering must be done at view time which will impact performance

#19 - 10/15/2012 07:51 PM - Luke Murphey

The HTML5 data attributes would be a good fit. I should be able to migrate most of the content from the XML over into HTML5 compliant content. See below for details:

- http://ejohn.org/blog/html-5-data-attributes/
- http://html5doctor.com/html5-custom-data-attributes/

#20 - 10/15/2012 11:47 PM - Luke Murphey

I have implemented a converter that produces HTML5 compliant snippets from the original TEA document.

Original:

```
<verse>
    <head>*(ikanw=s <num ref="some_ref">d</num> me\n </head>
</verse>
```

Processed:

```
<?xml version="1.0" encoding="utf-8"?><verse><span data-tagname="verse">
    <span data-tagname="head">Ἱκανῶς <span data-ref="some_ref" data-tagname="num">δ</span> μὲν </span>
</span></verse>
```

I did notice however, that some things ought not to be converted from beta-code, such at the <num> node. Will need to handle this.

#21 - 10/16/2012 06:17 PM - Luke Murphey

- % Done changed from 0 to 60

#22 - 11/12/2012 07:46 AM - Luke Murphey

- Start date deleted (09/02/2012)

- % Done changed from 60 to 80

#23 - 11/15/2012 08:12 AM - Luke Murphey

Successfully imported 392 works with 17 that could not be imported errors.

#24 - 11/16/2012 04:42 PM - Luke Murphey

All but one import now:

```
2012-11-16 03:12:26,824 [INFO] reader.importer.PerseusBatchImporter: Import complete, files_imported=408, impo
rt_errors=1, duration=2925
```

The import took 48 minutes.

#25 - 11/17/2012 09:29 AM - Luke Murphey

- Status changed from In Progress to Closed

#26 - 11/17/2012 04:26 PM - Luke Murphey

- Status changed from Closed to In Progress

#27 - 11/17/2012 05:01 PM - Luke Murphey

Updated the import policy:

- Euripides now ignores division markers
- Sophocles' Ichneutae uses line counts for division markers

These changes have been verified as successful.

#28 - 11/23/2012 06:28 PM - Luke Murphey

- Status changed from In Progress to Closed