TextCritical.net - Bug #2355

Feature # 551 (Closed): Lexicon support: Liddell and Scott (Middle)

Cannot index LSJ

12/22/2018 08:35 PM - Luke Murphey

Status:	Closed	**Start date:**	12/22/2018
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	3.0		

Description

```
python manage.py make_search_indexes -w middlelsj




Creating search indexes for work...
Traceback (most recent call last):
  File "manage.py", line 10, in <module>
    execute_from_command_line(sys.argv)
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/core/management/__init__.
py", line 354, in execute_from_command_line
    utility.execute()
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/core/management/__init__.
py", line 346, in execute
    self.fetch_command(subcommand).run_from_argv(self.argv)
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/core/management/base.py",
 line 394, in run_from_argv
    self.execute(*args, **cmd_options)
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/core/management/base.py",
 line 445, in execute
    output = self.handle(*args, **options)
  File "/Users/lmurphey/git/textcritical_net/src/reader/management/commands/make_search_indexes.py
", line 60, in handle
    WorkIndexer.index_work(work)
  File "/Users/lmurphey/git/textcritical_net/src/reader/contentsearch.py", line 219, in index_work
    cls.index_division(division, commit=False, writer=writer)
  File "/Users/lmurphey/git/textcritical_net/src/reader/contentsearch.py", line 242, in index_divi
sion
    cls.index_verse(verse, division=division, writer=writer, commit=False)
  File "/Users/lmurphey/git/textcritical_net/src/reader/contentsearch.py", line 344, in index_vers
e
    section       = cls.get_section_index_text(division),
  File "/Users/lmurphey/git/textcritical_net/src/reader/contentsearch.py", line 265, in get_sectio
n_index_text
    descriptions.append(unicode(division.get_division_description(use_titles=False).decode("UTF-8"
)))
  File "/Users/lmurphey/venv/django18/lib/python2.7/encodings/utf_8.py", line 16, in decode
    return codecs.utf_8_decode(input, errors, True)
UnicodeEncodeError: 'ascii' codec can't encode characters in position 3-8: ordinal not in range(12
8)
```

Associated revisions

Revision 955 - 12/28/2018 06:08 AM - lukemurphey

Adding test that identifies the error in the lexicon entry encoding

Reference #2355

Revision 956 - 12/29/2018 05:06 PM - lukemurphey

Fixed badly encoded string in Division model

Reference #2355

Revision 957 - 12/29/2018 05:10 PM - lukemurphey

Adding test for bad unicode data

Reference #2355

Revision 958 - 12/29/2018 06:33 PM - lukemurphey

Fixed issue where middle LSJ doesn't index

Reference #2355

History

#1 - 12/26/2018 01:00 AM - Luke Murphey

Is the hanging up on the one without titles titled "*a ἀάατος"

#2 - 12/26/2018 02:39 AM - Luke Murphey

These reproduces the issue:

```
unicode("ἀάατος")
"ἀάατος".encode("ascii").decode("utf-8")
unicode(u"ἀάατος".decode("utf-8"))
u"ἀάατος".decode("utf-8")
str(u"ἀάατος")
```

But this does not:

```
unicode("ἀάατος".decode("utf-8"))
u"ἀάατος".encode("utf-8").decode("utf-8")
str("ἀάατος")
```

#3 - 12/26/2018 03:15 AM - Luke Murphey

Questions:

- What is the division title encoded in?
  - Plain unicode
- Does changing the division titles as unicode fix it?
- Is it the decode call or the unicode construction that causes the error?
  - It fails on the decode, not on the constructor call
- Why does it fail even though the get_division_description() call returns unicode?
  - I confirmed that unicode is returned
- Does telling the unicode() call the encoding help?
  - No
  - See https://gist.github.com/gornostal/1f123aaf838506038710
- Does encoding the join work?

    - No
    - return ",".encode('utf8').join(descriptions)
- Does removing the unicode call help?
    - Causes the blow up later (in get_division_description line 389)
- Does adding an extra encode() call help (.encode("utf-8").decode("UTF-8"))?

- Is *str* on the Division model doing something wrong? It might be not handlings strings correctly

- Why is get_division_description() line 389 trying to force a string?

- Does removing the decode and unicode calls fix it?
    - No
- Is it possible that the unicode data is wrong in the database?
    - Seeing django.utils.encoding.DjangoUnicodeDecodeError: 'ascii' codec can't decode byte 0xe1 in position 0: ordinal not in range(128). You passed in <Division: [Bad Unicode data]> (<class 'reader.models.Division'>)

```
Traceback (most recent call last):
  File "manage.py", line 10, in <module>
    execute_from_command_line(sys.argv)
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/core/management/__init__.py", line 35
4, in execute_from_command_line
    utility.execute()
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/core/management/__init__.py", line 34
6, in execute
    self.fetch_command(subcommand).run_from_argv(self.argv)
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/core/management/base.py", line 394, i
n run_from_argv
    self.execute(*args, **cmd_options)
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/core/management/base.py", line 445, i
n execute
    output = self.handle(*args, **options)
  File "/Users/lmurphey/git/textcritical_net/src/reader/management/commands/make_search_indexes.py", line 60,
in handle
    WorkIndexer.index_work(work)
  File "/Users/lmurphey/git/textcritical_net/src/reader/contentsearch.py", line 219, in index_work
    cls.index_division(division, commit=False, writer=writer)
  File "/Users/lmurphey/git/textcritical_net/src/reader/contentsearch.py", line 242, in index_division
    cls.index_verse(verse, division=division, writer=writer, commit=False)
  File "/Users/lmurphey/git/textcritical_net/src/reader/contentsearch.py", line 348, in index_verse
    section       = cls.get_section_index_text(division),
  File "/Users/lmurphey/git/textcritical_net/src/reader/contentsearch.py", line 270, in get_section_index_text
    descriptions.append(division.get_division_description(use_titles=True))
  File "/Users/lmurphey/git/textcritical_net/src/reader/models.py", line 389, in get_division_description
    title = str(next_division)
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/db/models/base.py", line 503, in __st
r__
    return force_text(self).encode('utf-8')
  File "/Users/lmurphey/venv/django18/lib/python2.7/site-packages/django/utils/encoding.py", line 102, in forc
e_text
    raise DjangoUnicodeDecodeError(s, *e.args)
django.utils.encoding.DjangoUnicodeDecodeError: 'ascii' codec can't decode byte 0xe1 in position 5: ordinal no
t in range(128). You passed in <Division: [Bad Unicode data]> (<class 'reader.models.Division'>)
```

- Does it work with other works?
    - It does with ad-ammaeum

Observations:

- Apparently get_division_description() line 394 gets something other than a string at times

#4 - 12/26/2018 05:31 PM - Luke Murphey

The data in the model appears to be bad:

```
import reader.models
model = reader.models.Division.objects.get(id=54009)
```

This outputs:

```
<Division: [Bad Unicode data]>
```

```
model.descriptor.decode('utf8')
Traceback (most recent call last):
  File "<console>", line 1, in <module>
  File "/Users/lmurphey/venv/django18/lib/python2.7/encodings/utf_8.py", line 16, in decode
    return codecs.utf_8_decode(input, errors, True)
UnicodeEncodeError: 'ascii' codec can't encode characters in position 0-5: ordinal not in range(128)
```

#5 - 12/29/2018 05:19 PM - Luke Murphey

Now that the issue is narrowed down to the

Questions:

- Is the issue in make_division()?
  - It is probably ImportTransforms.convert_descriptors_from_beta_code()
- What field is the problem?
  - descriptor field

Observations: *

#6 - 12/29/2018 05:38 PM - Luke Murphey

This reproduces the issue that I am seeing in content_search:

```
descriptions = []
descriptions.append("main ΆΑΑΤΟΣ")
```

```
descriptions.append(u"ἀάατος")

",".join(descriptions)
```

This does not:

```
descriptions = []
descriptions.append("main ἈΑΑΤΟΣ")
descriptions.append(u"ἀάατος".encode("utf-8"))

",".join(descriptions)
```

#7 - 12/29/2018 05:42 PM - Luke Murphey

No exceptions are thrown in content_search when I do the following:

```
        descriptions.append(division.get_division_description(use_titles=False).encode("utf-8"))
        descriptions.append(division.get_division_description(use_titles=True).decode("utf-8").encode("utf-8")
)
```

However, Whoosh wants unicode.

#8 - 12/30/2018 02:08 AM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 0 to 100