

NAME

sortIntoDirectories.pl

SYNOPSIS

% perl sortIntoDirectories.pl

DESCRIPTION**SUMMARY**

Script to organize (.sgml) texts by manuscript. The script creates folders and places each text within its respective folder. Individual files (saved within the new respective folders) are modified/saved to hold only the text (specifically, the words between the sgml <s> tags), thus, all header and most <tag> information has been removed. sgml tags for Old English characters are kept, e.g., ð for eth. In this script, no consolidation of characters is performed; for example, no converting from eth to thorn (see folder "2_count_words" for consolidation options).

Note: this script currently only handles A[poetry] and B[prose].
All texts within other manuscripts are ignored.

Note: see additional information in file "editsByHand.txt"

INPUT

(see code labeled "PATH and FILE NAMES" in code)

This script assumes you are a member of and have legal access to: The Dictionary of Old English Web (DOE) Corpus, "an online database consisting of at least one copy of every Old English text. Compiled as part of the Dictionary of Old English project at the University of Toronto, the texts in the Corpus are SGML encoded and are fully conformant with the 1994 Text Encoding Initiative (TEI) Guidelines." (<http://www.doe.utoronto.ca/>)

The DOE Web Corpus is available by institutional site license through the Dictionary of Old English project.

<https://tir.doe.utoronto.ca/store/index.php?page=corpus>

Once you have purchased the Dictionary of Old English Corpus, place the DOE's sgml-corpus folder found in the directory:

:\oecorpushtml\sgml-corpus

in place of the sgml-corpus folder in this zip folders' directory.

Text files (.sgml) are input one at a time based on those listed in the file:

/sgml-corpus/textlist.sgml

```
<!-- Entity Cameron number Short title -->
<!ENTITY T00010 SYSTEM "T00010.sgml"> <!-- A1.1 GenA,B -->
```

This script snags a line at a time from textlist.sgml and then determines the file to open (e.g., T00010.sgml) and the appropriate manuscript (e.g., 'A01'), text# (e.g., 01.001), and short title (e.g., GenA,B) from the Cameron number and short title. For example, the contents of the file /sgml-corpus/T00010.sgml will be stored in the A01 manuscript folder within the A_poetry folder.

OUTPUT

A new folder called "sorted_texts" that holds: "A_poetry" and "B_prose", each holding subdirectories for each manuscript (e.g., A01, A02, ...). Each manuscript folder holds (new) filenames: A<n>.<m>.<o>.<p>_<title>

Example: A_poetry/A01/A01.001_GenA_B.txt
 /A01.002_Ex.txt

AUTHORS

Mark D. LeBlanc
Christina L. Nelson

MODIFICATION HISTORY

June 4, 2007 (mdl) --

back from fishin', just getting started

Sept 14, 2007 (mdl) --

no longer keep only A1, A2, A3, A4, A5, A6, A10;
now keeping all poetry (all As) since so many fewer
compared to prose(B)

May 12, 2008 (cIn) --

walkthrough, commenting

May 22, 2008 (cIn) --

writing pod

May 28, 2008 (mdl, cIn) --

work on changing files from A1 to A01, etc.
to make them more web-friendly.

May 29, 2008 (cIn) --

finish making the files more web-friendly.
Ex. A1.1 changed to A01.001

May 30, 2008 (cIn) --

finished pod

June 2, 2008 (cIn) --

finished commenting

May 11, 2009 (mdl, cIn) --

updated documentation

COPYRIGHT INFORMATION

=====
Copyright (C) 2009 Wheaton Lexomics Research Group, Norton MA

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.