

# “With One Voice:”

## A Modular Approach to Streamlining Character Data for Tokenization

Ashley M. Clark

Northeastern University Women Writers Project

July 31, 2019

Balisage: The Markup Conference

# Women Writers Online

- Corpus of TEI documents
  - WWO uses a TEI customization with added elements, constraints, descriptions
- Texts authored\* by women, published between 1526 and 1850
  - (where authorship can include contributions such as translation)
- Transcribed from facsimile by Women Writers Project encoders

# Women Writers Online

- Published online for subscribing institutions and individuals
- <https://www.northeastern.edu/www>

The screenshot displays the Women Writers Online website. The top navigation bar includes 'About', 'Help', and 'WWP'. The main content area is divided into three sections: Search, Results, and Full text.

**Search:** A search bar with the text 'Search...' and an 'OK' button. Below it is a 'Filter' section with a 'Genre' dropdown menu showing 'non-fiction [195]', 'verse [147]', 'drama [68]', and 'fiction [57]'. There is also a 'Date' section with a dropdown menu showing '16th c. [30]', '17th c. [171]', '18th c. [119]', and '19th c. [96]'. A 'Preferences' section at the bottom left has a checkbox for 'Show original typographical features'.

**Results: 416 texts**

- [unknown] — The Bridling, Saddling, and Riding of a Rich Churt in Hampshire, 1595
- [unknown] — Changing Scenes, 1625
- [unknown] — Eliza's Babes, 1652
- [unknown] — England's Tears: A Poem, 1774
- [unknown] — The Female American (vol. 1), 1767
- [unknown] — The Female Wits, 1704
- [unknown] — The Fortunate Transport, 1750
- [unknown] — Memoir of Mrs. Chloe Spear, 1632
- [unknown] — Swetnam, the Woman-Hater, Arraigned by Women, 1620
- [unknown] — The Wonderful Discoverie of the Witchcrafts of Margaret and Phillip Flower, 1619
- [various] — The First Lamp of Virginitie, 1582
- [various] — Letters of Love and Gallantry (vol. 1), 1693
- [various] — Letters of Love and Gallantry (vol. 2), 1694
- [various] — The Monument of Matrons (excerpts), 1582
- Adams, Hannah — The History of the Jews, 1812
- Adams, Hannah — A Summary History of New-England, 1799
- Alkin, Lucy — Epistles on Women, 1810
- Anger, Jane — Jane Anger Her Protection for Women, 1589
- Ariadne — She Ventures, and He Wins, 1696

**Full text: Davies, The Benediction**

The full text view shows a poem titled 'The Benediction from the A:Imighty O:mnipotent.' The poem is attributed to 'Handwriting: Lady Eleanor Davies — wife of Sir John Davies. Poet — late married Sir Arch. Douglas'. The text of the poem is as follows:

"I have an Errand to thee O: Captain."

2 Kings 9.5.

Printed in the Year, 1651.

The poem is followed by a section titled 'For the Armies General, His Excellency. My Lord,' which begins with the text: 'Your Interest in the Nations unparaleld Troublesom Times: The Flaming Sword for expelling the Man in your hand, which Crowns with no Inferior Honor that Name of Yours: Hereof by her Hand a touch presented. Derived from his own, namely, A. & O. Letters of no mean Latitude: Armed beside with his Sword: Sun and Moon when as stood in Admiration, witness ☉ ☽ their Golden Characters, stiled Eyes and Horns of the Lamb, &c. "Their voice gone out into all Lands", Psal. (Rev. 5.) Like theirs here, every one when the fifty days at an end, "heard in his

# Women Writers Online

- Maintained privately under version control using Subversion
- The corpus is regularly given to researchers
  - The whole thing
  - Subsets of documents (by author, by genre, by time period, &c.)
  - One document
- XML? Only the text content?
  - Tokens (characters, words, sentences, &c.)
  - Word-based analysis is popular, at least at Northeastern

# Tokenization and Regularization

- How much difference should be captured by a single unique form?
- Old-style letterforms

The Prophet

```
<persName rend="slant(upright)"><vuji>I</vuji>oe1</persName>
```

- <choice>

```
<persName rend="slant(upright)" xml:id="cromwel2">  
  <choice>  
    <abbr>O:</abbr>  
    <expn>Oliver</expn>  
  </choice> Cromwel</persName>
```

If regularization is a personal standard,  
why not give the researcher XML and  
let them choose?

# Learning to manipulate XML is not enough

```
<hyphenation eol="all">
```

```
    <p>The encoding distinguishes between hard and soft hyphens. Hard hyphens, whether at the end-of-line or not, are recorded as a plain-text hyphen character. Soft hyphens are recorded using the SOFT HYPHEN character (U+00AD). Hyphens in catchwords are transcribed as hard hyphens.</p>
```

```
</hyphenation>
```

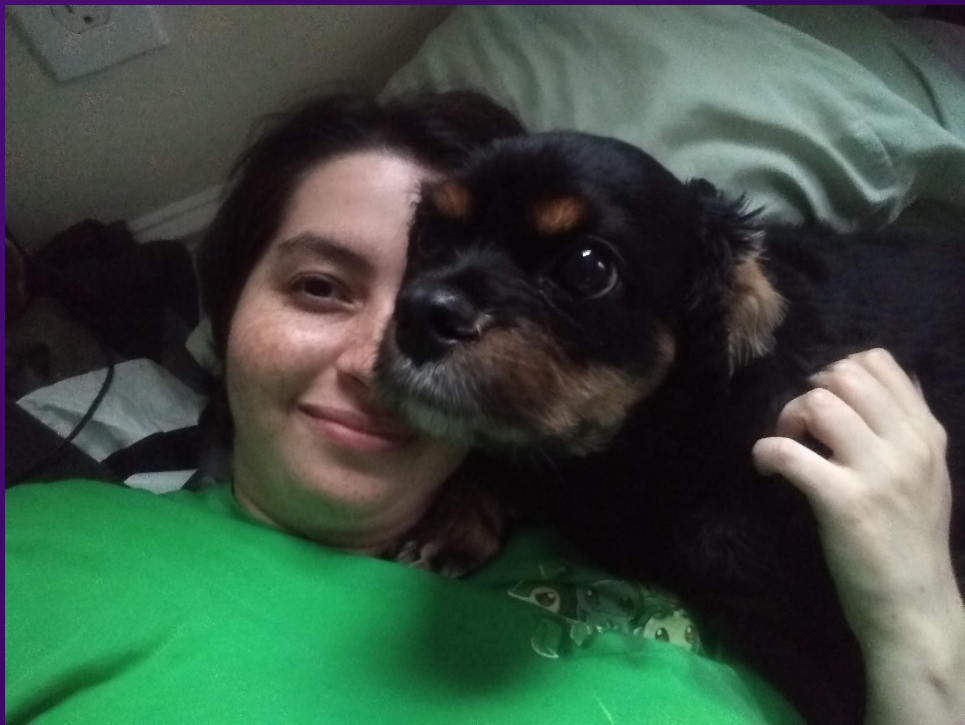
From the WWO <editorialDecl>.

# “Soft” (end-of-line) hyphens

```
witnefs <seg xml:id="a003" corresp="#n003">⊙ ㉔ </seg> their Golden Cha-  
<lb/>racters, tiled Eyes and Horns of the  
<lb/>Lamb
```

```
text()[  
  preceding::text()  
    [not(parent::sic)]  
    [not(normalize-space(.)='')]  
    [1]  
    [contains(.,'&#xAD;')] ]  
|  
text()[  
  preceding::*  
    [1]  
    [not( self::anchor | self::cb |  
self::mw | self::gb | self::lb |  
self::milestone | self::pb )]  
  /  
  preceding::text()  
    [not(normalize-space(.)='')]  
    [1]  
    [contains(.,'&#xAD;')]
```





# How does one *know* a corpus?

- TEI ODD file
  - Schema, Schematron
  - Lists of allowed values
  - Descriptions
- Documentation
  - <https://www.northeastern.edu/research/publications/documentation/internal>
- Statements about project methods, principles
  - [https://www.northeastern.edu/about/methods/editorial\\_principles.html](https://www.northeastern.edu/about/methods/editorial_principles.html)
- Querying
- Experience!

# XML



By Popo le Chien - Own work, CC0,  
<https://commons.wikimedia.org/w/index.php?curid=38271720>

# Plain text



By Adina Bogert-O'Brien. Own work assumed (based on  
copyright claims), Public Domain,  
<https://commons.wikimedia.org/w/index.php?curid=603597>



```

1 Inventory
2 -----
3 ...files/.....contains the individual source XML files[1]
4 ...files/personogrophy.xml.....subset of the WWP prosopography file[2]
5 ...aggregate/.....contains derived aggregations of the source files[3]
6 ...aggregated/corpusDriver.xml driver <teiCorpus> file; XIncludes the individual files
7 ...aggregated/corpus.xml.....corpusDriver.xml after XInclude processing[4]
8 ...common-boilerplate.xml.....contains common elements included in textbase files in files/
9 ...words/.....contains lists of "words"[5]
10 ...words/ALL.txt.....words of all files in the corpus, all on one line[6,4]
11 ...words/ranked.txt.....one word per line with the number of times it occurs
12 ...words/no_punc_ranked.txt.....same, but punctuation stripped out first
13 ...words/caseless_ranked.txt.....same, but also case-folded (to lowercase) first
14 ...schema/.....contains the WWP storage schemas
15 ...schema/wwp-store.odd.....TEI ODD customization file
16 ...schema/wwp-store.rng.....RELAX NG with Schematron
17 ...schema/wwp-store.rnc.....RELAX NG compact syntax
18 ...schema/wwp-store.isosch.....Schematron
19 ...schema/wwp-store.nvdl.....NVDL script for using both RNG and Schematron[7]
20 ...schema/wwp-store.html.....generated documentation
21 ...schema/shyTest.sch.....Schematron to test soft hyphen placement
22 ...schema/charDecl.xml.....character declaration (for non-Unicode chars)]
23
24 Notes
25 -----
26 [1] The classification declaration has been removed from each file.
27 ...It looked like
28 .....<classDecl>
29 .....<xi:include href="taxonomy.xml"/>
30 .....</classDecl>
31 ...and appeared in the <encodingDesc> after the <tagsDecl> (typically
32 ...the last element of the <encodingDesc>).
33
34 [2] Those explicitly flagged as "non-textbase" (typically student employees)
35 ...have been removed.
36
37 [3] The entire set of aggregations is experimental. The resulting
38 ...files, while interesting and perhaps useful, have problems I know
39 ...of (e.g., are invalid due to repeated ID values on id=
40 ...attributes), and may have problems I don't know about.
41
42 [4] *Very* large file, so you may not want to open this in your
43 ...editor.

```

Excerpt from the README Syd provides to people who want a copy of WWO.

# Demystifying the process

“FulltextBot” XSLT guiding principles:

- Some normalization processes are almost always beneficial
- Preserve markup for as long as it is valuable
- Textual interventions are marked for human decipherability

# Some normalization processes are almost always beneficial

Prerequisites for soft hyphen processing:

- Add whitespace around elements which imply breakage, or which are stated to break
- Group paginal artifacts together
- “Dehydrate” text nodes which are unique to a document's manifestation rather than expression (catch words, page numbers, signature marks)

# Preserve markup for as long as it is valuable

sourceable-quotes

sourceable-quotes

sourceable-quotesTemp. BibliographyXML Report

passage d0e103600 —Hannah

Appears in adams.jews.xml after milestone "Aa03r".

((//text[@xml:id eq "TR00128.02"]//milestone[@n eq "Aa03r"])[1]/following::p[3])

Notwithstanding the long protracted calamities the Jews have suffered since their dispersion, the most violent persecutions have never prevailed upon the general mass of this people to abjure their religion. David Levi, speaking of those among his brethren, who, in all ages, have professed christianity, observes, that

they have not acted voluntarily, but by compulsion, as in Spain and Portugal, or from interested motives, as there, and elsewhere; that notwithstanding they seemed to apostatize, and pretended to embrace christianity, yet in their hearts they secretly adhered to the true faith and law of Moses; and such are at this day called among us,


the compelled


, because they act by compulsion; for, as soon as they can by any means escape from the popish countries, they instantly return to Judaism.


I am free to assert,


says he,

that there is scarcely an instance of a Jew ever having embraced christianity on the pure principles of religion, but merely from interested motives.

IT00204  
*Dissertations on the Prophecies of the Old Testament*  
Levi, David

IT00204  
*Dissertations on the Prophecies of the Old Testament*  
Levi, David

IT00204  
*Dissertations on the Prophecies of the Old Testament*  
Levi, David

IT00204  
*Dissertations on the Prophecies of the Old Testament*  
Levi, David

Levi's *Dissertations*, &c. Vol. II. p. 15.

# Textual interventions are marked for human decipherability

```
220     let $teiDoc := $teiDoc/descendant-or-self::TEI
221     (: Change $ELEMENTS to reflect the elements for which you want full-text
222        representations. :)
223     let $ELEMENTS := $teiDoc/text
224     (: Below, add the names of elements that you wish to remove from within $ELEMENTS.
225        For example,
226        ('castList', 'elision', 'figDesc', 'label', 'speaker')
227     :)
228     let $ELEMENTS2OMIT := ()
229     let $fulltext :=
230         let $wordSeq := for $element in $ELEMENTS
231             let $abridged := local:omit-descendants($element, $ELEMENTS2OMIT)
232             return
233                 if ( $is-morphadorned ) then
234                     local:get-morphadorned-text($abridged, $morphadorner-text-type)
235                 else local:get-text($abridged)
236         let $wordSeparator := ' '
237         return normalize-space(string-join(($wordSeq), $wordSeparator))
238     (: The variable $optionalMetadata will be empty if $return-only-words is 'true()'. :)
239     let $dataSeq := ( $optionalMetadata, $fulltext )
240     order by $file
241     return
```



# Thank you!

<https://github.com/NEU-DSG/wwp-public-code-share/tree/master/fulltext>