“With One Voice:”

A Modular Approach to Streamlining Character Data for Tokenization

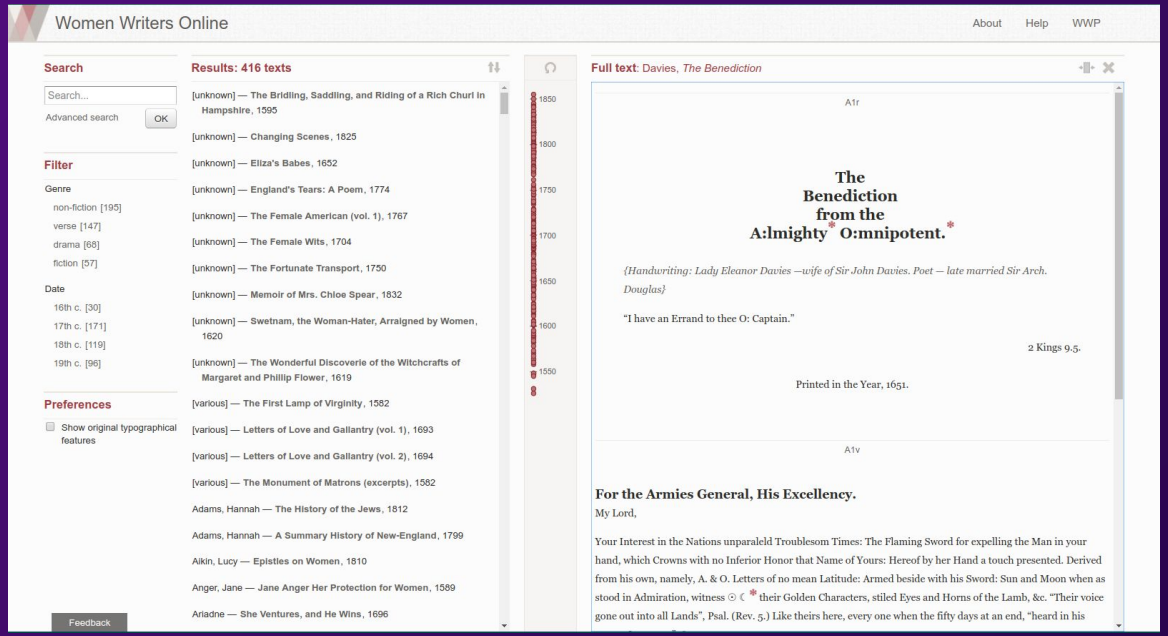
Ashley M. Clark
Northeastern University Women Writers Project
July 31, 2019
Balisage: The Markup Conference

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://wwp.northeastern.edu/wwo



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>oel</persName>
```

- <choice>

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

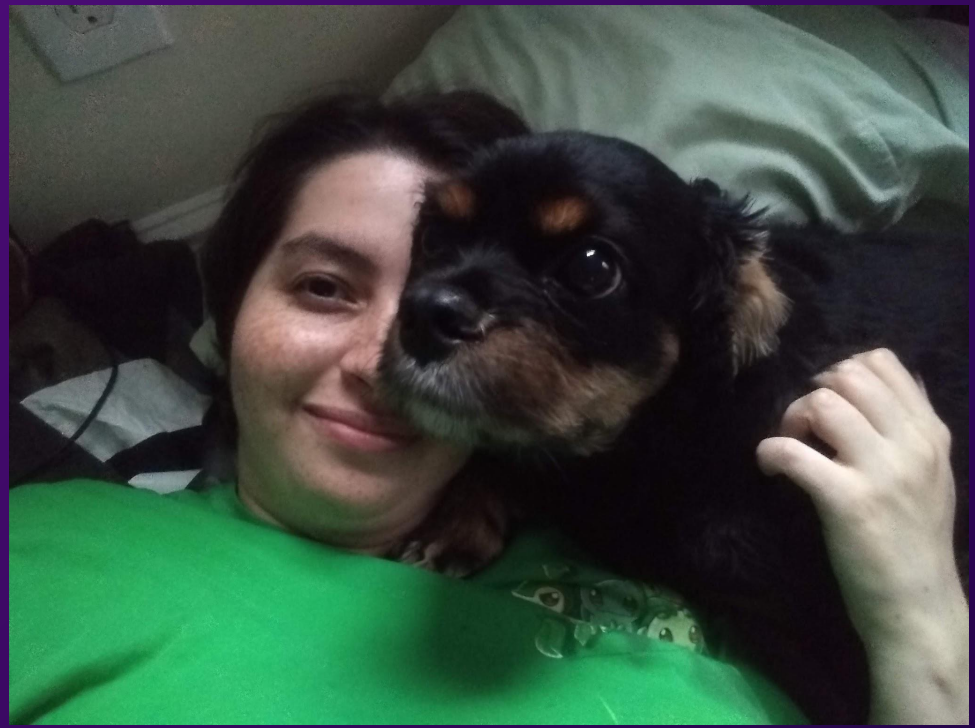
“Soft” (end-of-line) hyphens

```
witneſs <seg xml:id="a003" corresp="#n003">⊙ ☾</seg> their Golden Cha-

<lb/>racters, ſtiled 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://wwp.northeastern.edu/research/publications/documentation/internal
- Statements about project methods, principles
  - https://wwp.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

```
Inventory
---------
   files/                        contains the individual source XML files[1]
   files/personography.xml       subset of the WWP prosopography file[2]
   aggregate/                    contains derived aggregations of the source files[3]
   aggregated/corpusDriver.xml   driver <teiCorpus> file; XIncludes the individual files
   aggregated/corpus.xml         corpusDriver.xml after XInclude processing[4]
   common-boilerplate.xml        contains common elements included in textbase files in files/
   words/                        contains lists of "words"[5]
   words/ALL.txt                 words of all files in the corpus, all on one line[6,4]
   words/ranked.txt              one word per line with the number of times it occurs
   words/no_punc_ranked.txt      same, but punctuation stripped out first
   words/caseless_ranked.txt     same, but also case-folded (to lowercase) first
   schema/                       contains the WWP storage schemas
   schema/wwp-store.odd          TEI ODD customization file
   schema/wwp-store.rng          RELAX NG with Schematron
   schema/wwp-store.rnc          RELAX NG compact syntax
   schema/wwp-store.isosch       Schematron
   schema/wwp-store.nvdl         NVDL script for using both RNG and Schematron[7]
   schema/wwp-store.html         generated documentation
   schema/shyTest.sch            Schematron to test soft hyphen placement
   schema/charDecl.xml           character declaration (for non-Unicode chars)

Notes
-----
[1] The classification declaration has been removed from each file.
    It looked like
        <classDecl>
            <xi:include href="taxonomy.xml"/>
        </classDecl>
    and appeared in the <encodingDesc> after the <tagsDecl> (typically
    the last element of the <encodingDesc>).

[2] Those explicitly flagged as "non-textbase" (typically student employees)
    have been removed.

[3] The entire set of aggregations is experimental. The resulting
    files, while interesting and perhaps useful, have problems I know
    of (e.g., are invalid due to repeated ID values on id=
    attributes), and may have problems I don't know about.

[4] *Very* large file, so you may not want to open this in your
    editor.
```

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

Demystifying the process

“FulltextBot” XSLT guiding principles:

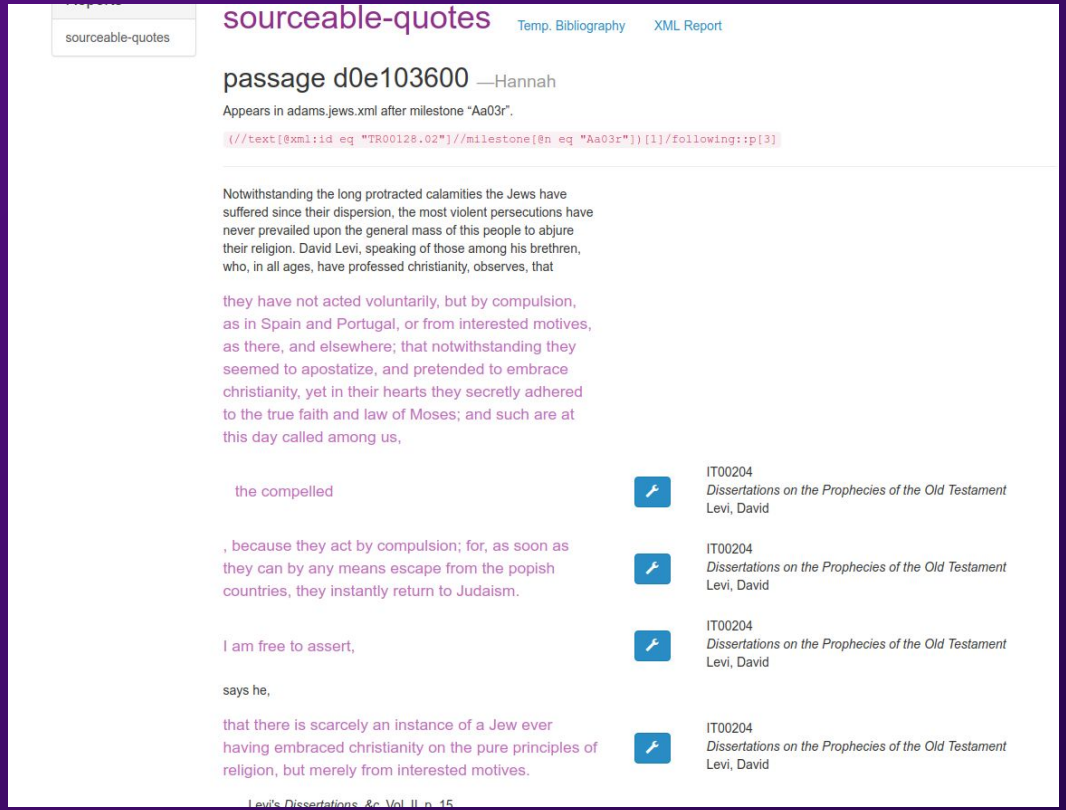
- Some normalization processes are almost always beneficial
- Preserve markup for as long as it is valuable
- Textual interventions are marked for human decipherability

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



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