CMPU 187 · Introduction to Digital Humanities

Textual Data

23 April 2021

The world of text

Most text *corpora* are either:

- “born digital”, e.g.,
  - Blog posts
  - Social media activity
  - Stories on fanfiction.net
- Converted from handwritten / typed writing on paper, e.g.,
  - a collection of 19th-century British novels
  - letters written by Seneca
  - issues of *The North Star*

Conversion from physical texts to digital ones can be done by

manual transcription,

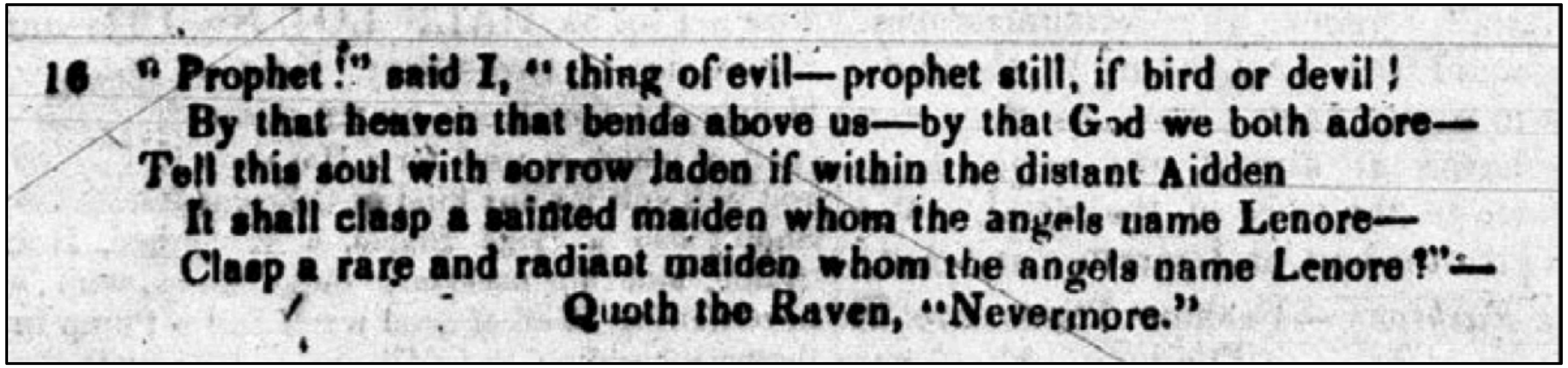
automatic optical character recognition (OCR), or

a hybrid, e.g., OCR with manual editing

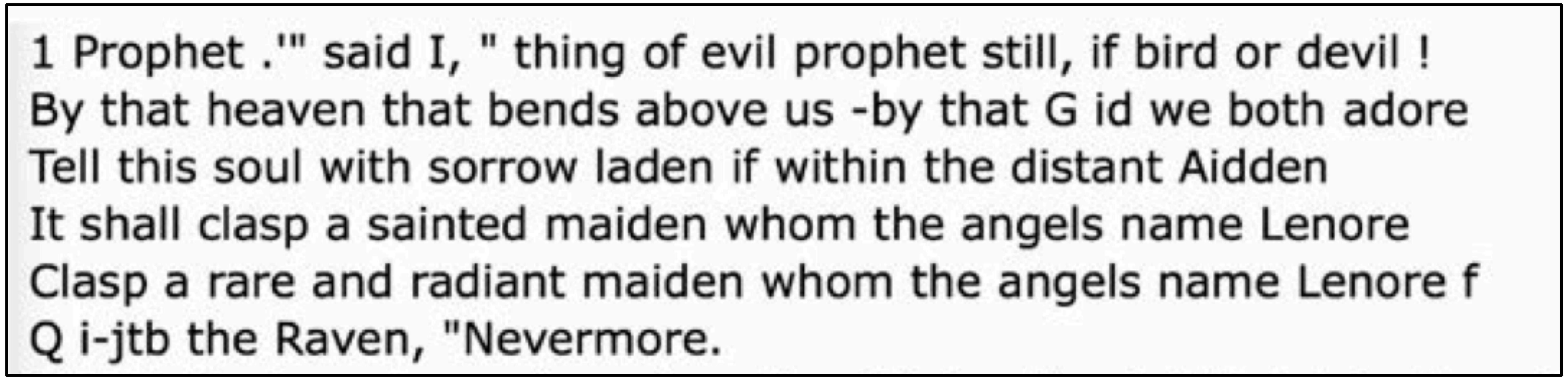
“Where human transcription would be prohibitively expensive and slow, through OCR words printed on thousands or millions of physical texts become, almost immediately to scholarly timelines, machine readable data that can be identified and computationally analyzed.”

Ryan Cordell, “ ‘Q i-jtb the Raven’: Taking Dirty OCR Seriously”

OCR without manual verification and correction can be problematic – a noisy channel:



16 " Prophet !" said I, " thing of evil—prophet still, if bird or devil !
By that heaven that bends above us—by that God we both adore—
Tell this soul with sorrow laden if within the distant Aidden
It shall clasp a sainted maiden whom the angels uame Lenore—
Clasp a rare and radiant maiden whom the angels name Lenore !"—
Quoth the Raven, "Nevermore."



1 Prophet .'" said I, " thing of evil prophet still, if bird or devil !
By that heaven that bends above us -by that G id we both adore
Tell this soul with sorrow laden if within the distant Aidden
It shall clasp a sainted maiden whom the angels name Lenore
Clasp a rare and radiant maiden whom the angels name Lenore f
Q i-jtb the Raven, "Nevermore.

Ryan Cordell,
" 'Q i-jtb the Raven': Taking Dirty OCR Seriously"

OCR

There are many options, including

- Tesseract (free, but less friendly)
- OmniPage (commercial)
- ABBYY FineReader (commercial; popular in digital humanities)

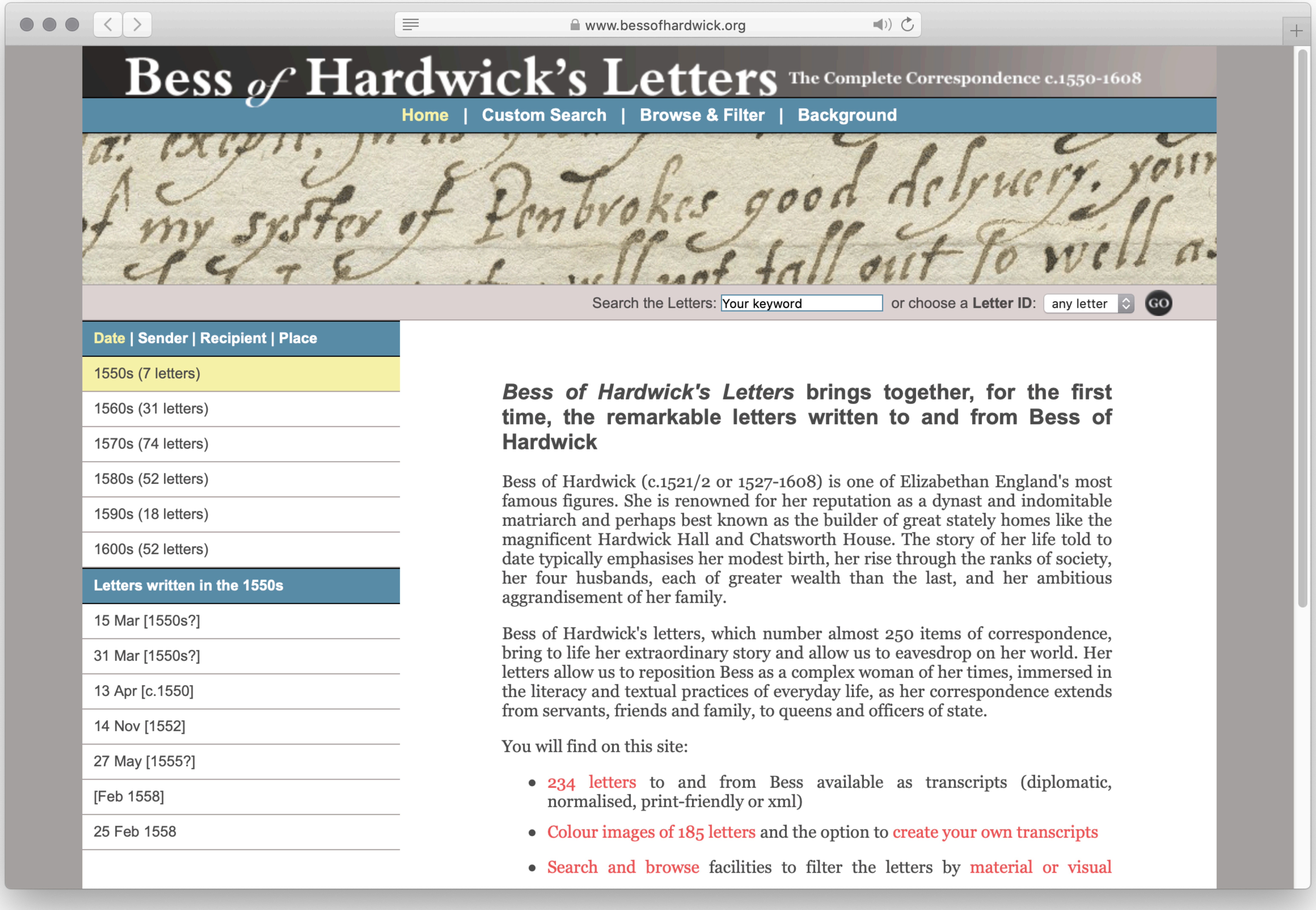
Tesseract

Requires image as input rather than PDFs.

- Given a PDF, first convert it into a directory of images using pdf2png.com

FineReader

Encoding and annotation



www.bessofhardwick.org
Home | Custom Search | Browse & Filter | Background
Search the Letters: Your keyword or choose a Letter ID: any letter GO
Date | Sender | Recipient | Place

Bess of Hardwick's Letters The Complete Correspondence c.1550-1608

1550s (7 letters)
1560s (31 letters)
1570s (74 letters)
1580s (52 letters)
1590s (18 letters)
1600s (52 letters)

Letters written in the 1550s

15 Mar [1550s?]
31 Mar [1550s?]
13 Apr [c.1550]
14 Nov [1552]
27 May [1555?]
[Feb 1558]
25 Feb 1558

Bess of Hardwick's Letters brings together, for the first time, the remarkable letters written to and from Bess of Hardwick

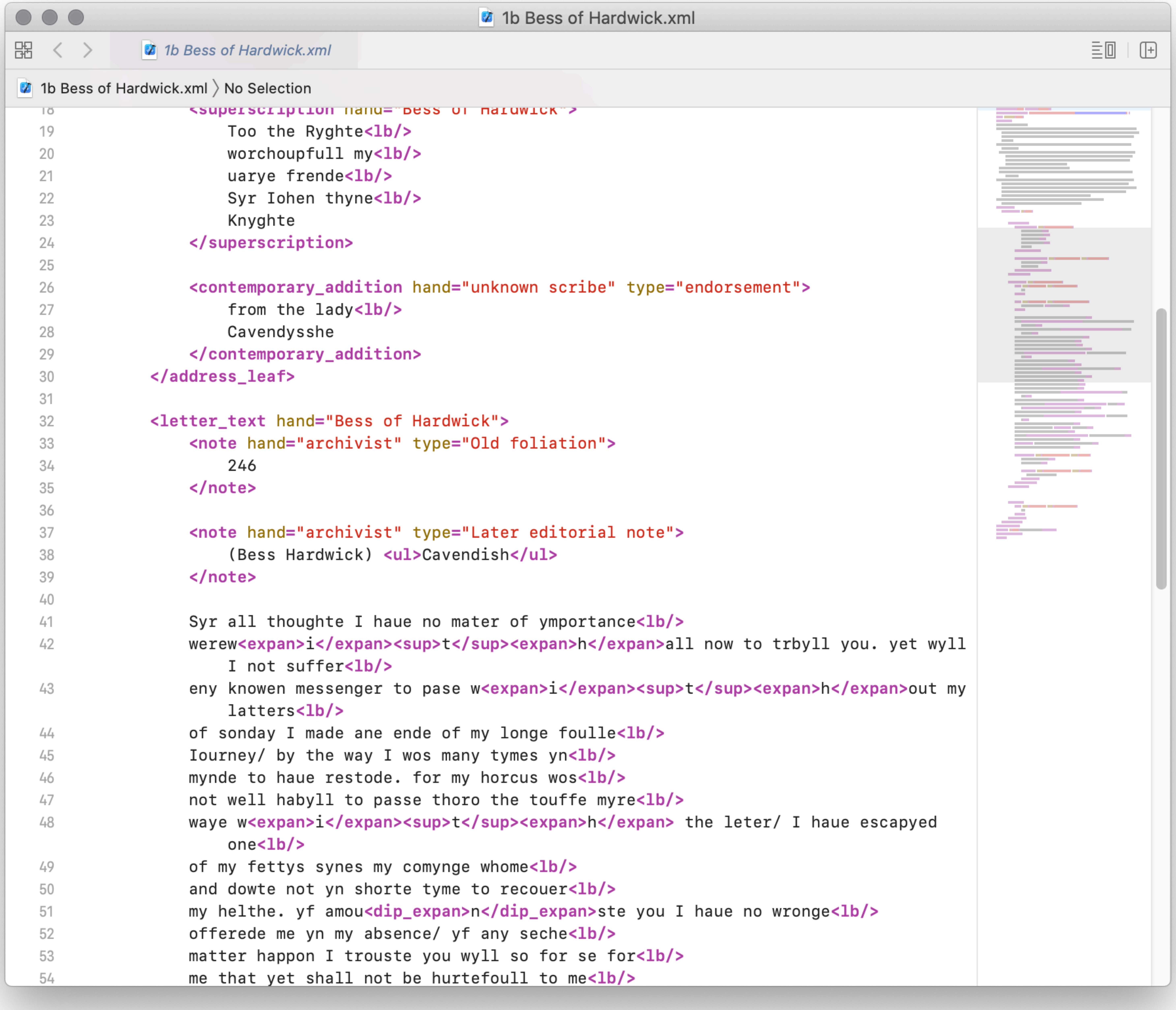
Bess of Hardwick (c.1521/2 or 1527-1608) is one of Elizabethan England's most famous figures. She is renowned for her reputation as a dynast and indomitable matriarch and perhaps best known as the builder of great stately homes like the magnificent Hardwick Hall and Chatsworth House. The story of her life told to date typically emphasises her modest birth, her rise through the ranks of society, her four husbands, each of greater wealth than the last, and her ambitious aggrandisement of her family.

Bess of Hardwick's letters, which number almost 250 items of correspondence, bring to life her extraordinary story and allow us to eavesdrop on her world. Her letters allow us to reposition Bess as a complex woman of her times, immersed in the literacy and textual practices of everyday life, as her correspondence extends from servants, friends and family, to queens and officers of state.

You will find on this site:

- 234 letters to and from Bess available as transcripts (diplomatic, normalised, print-friendly or xml)
- Colour images of 185 letters and the option to create your own transcripts
- Search and browse facilities to filter the letters by material or visual

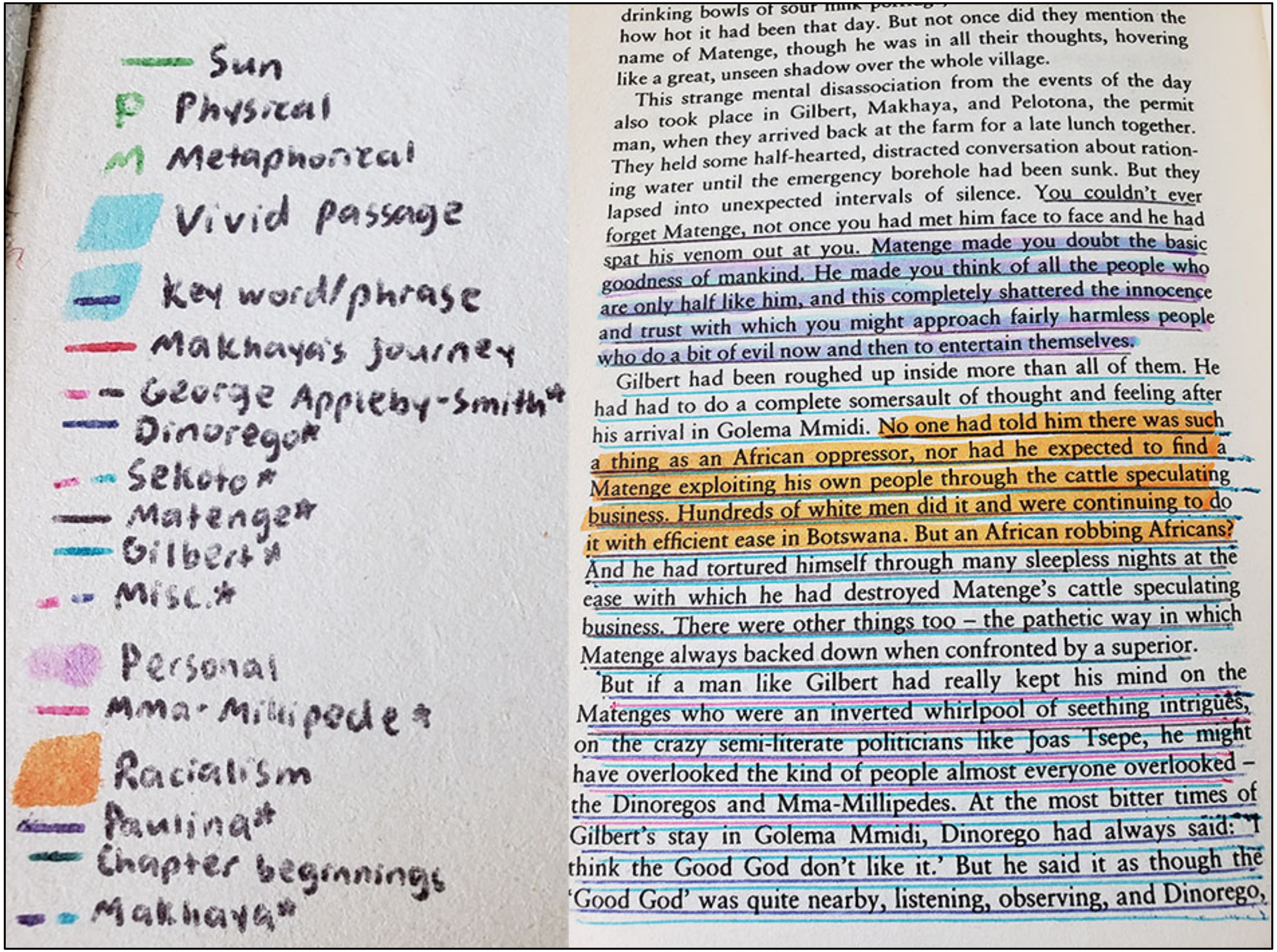
bessofhardwick.org



```
            <superscription hand="Bess of Hardwick">
                Too the Ryghte<lb/>
                worchoupfull my<lb/>
                uarye frende<lb/>
                Syr Iohen thyne<lb/>
                Knyghte
            </superscription>

            <contemporary_addition hand="unknown scribe" type="endorsement">
                from the lady<lb/>
                Cavendysshe
            </contemporary_addition>
        </address_leaf>

        <letter_text hand="Bess of Hardwick">
            <note hand="archivist" type="Old foliation">
                246
            </note>

            <note hand="archivist" type="Later editorial note">
                (Bess Hardwick) <ul>Cavendish</ul>
            </note>

            Syr all thoughte I haue no mater of ymportance<lb/>
            werew<expan>i</expan><sup>t</sup><expan>h</expan>all now to trbyll you. yet wyll I not suffer<lb/>
            eny knowen messenger to pase w<expan>i</expan><sup>t</sup><expan>h</expan>out my latters<lb/>
            of sonday I made ane ende of my longe foulle<lb/>
            Iourney/ by the way I wos many tymes yn<lb/>
            mynde to haue restode. for my horcus wos<lb/>
            not well habyll to passe thoro the touffe myre<lb/>
            waye w<expan>i</expan><sup>t</sup><expan>h</expan> the leter/ I haue escapyed one<lb/>
            of my fettys synes my comynge whome<lb/>
            and dowte not yn shorte tyme to recouer<lb/>
            my helthe. yf amou<dip_expan>n</dip_expan>ste you I haue no wronge<lb/>
            offerede me yn my absence/ yf any seche<lb/>
            matter happon I trouste you wyll so for se for<lb/>
            me that yet shall not be hurtefoull to me<lb/>
```

“…markup is less like jotting down notes in the margins of a book, and more like going through a book with a set of highlighters, using one color to mark speech acts by women, and another color to mark metaphors, and a third color to mark allusions to Emily Dickinson. If your book contains a speech act by a woman that contains a metaphor, some of the text might be highlighted with two colors…your ‘highlighters’ are what’s called ‘elements’, and ‘attributes’ that modify those elements”

Beshero-Bondar et al., *DSC 5*



— Sun
P Physical
M Metaphorical
▨ Vivid passage
▨ Key word/phrase
— Makhaya's journey
– – George Appleby-Smith*
— Dinorego*
– – Sekoto*
— Matenge*
— Gilbert*
– – Misc.*
● Personal
— Mma-Millipede*
▨ Racialism
— Paulina*
— Chapter beginnings
– – Makhaya*

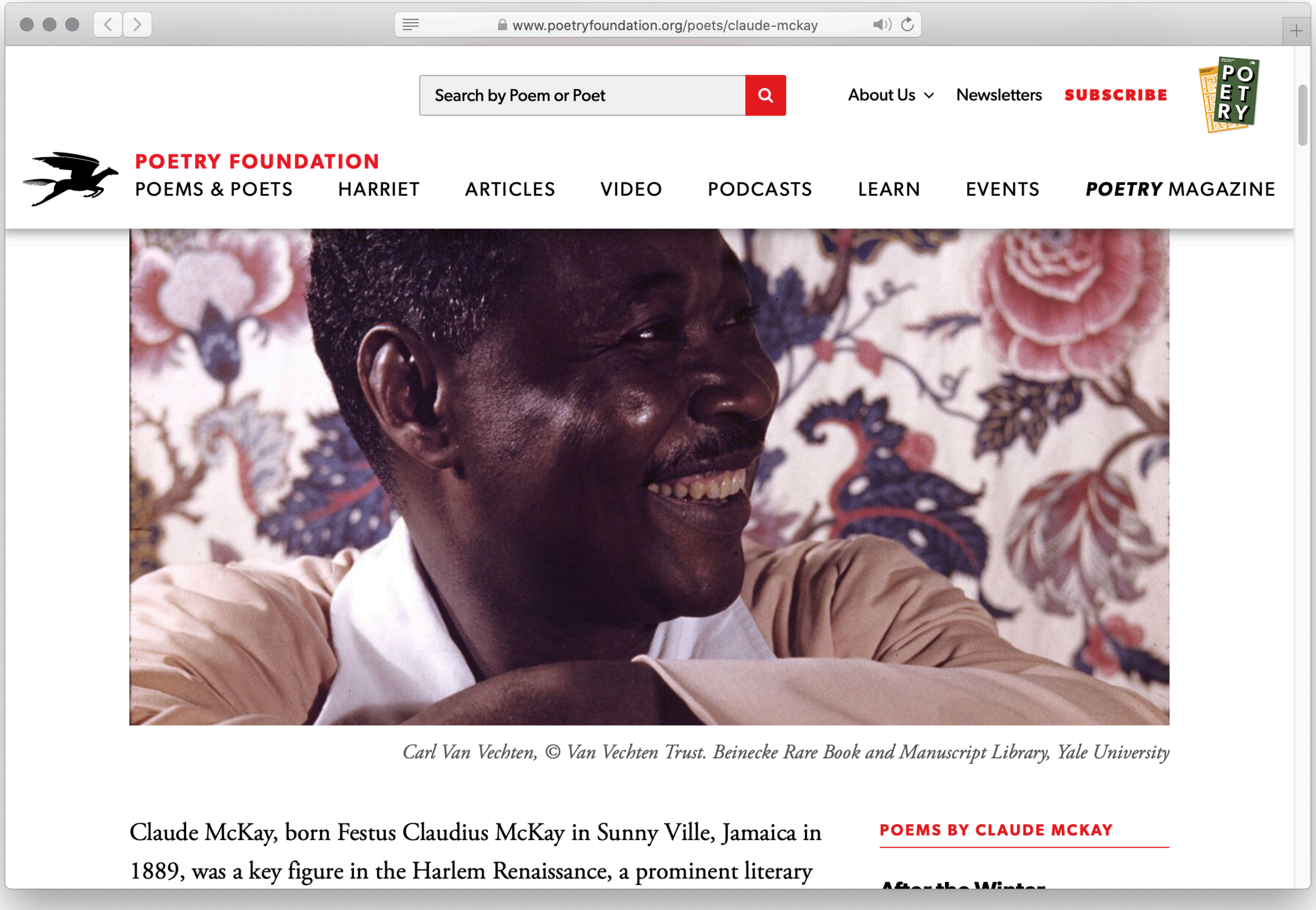
drinking bowls of sour milk por how hot it had been that day. But not once did they mention the name of Matenge, though he was in all their thoughts, hovering like a great, unseen shadow over the whole village.

This strange mental disassociation from the events of the day also took place in Gilbert, Makhaya, and Pelotona, the permit man, when they arrived back at the farm for a late lunch together. They held some half-hearted, distracted conversation about rationing water until the emergency borehole had been sunk. But they lapsed into unexpected intervals of silence. You couldn't ever forget Matenge, not once you had met him face to face and he had spat his venom out at you. Matenge made you doubt the basic goodness of mankind. He made you think of all the people who are only half like him, and this completely shattered the innocence and trust with which you might approach fairly harmless people who do a bit of evil now and then to entertain themselves.

Gilbert had been roughed up inside more than all of them. He had had to do a complete somersault of thought and feeling after his arrival in Golema Mmidi. No one had told him there was such a thing as an African oppressor, nor had he expected to find a Matenge exploiting his own people through the cattle speculating business. Hundreds of white men did it and were continuing to do it with efficient ease in Botswana. But an African robbing Africans? And he had tortured himself through many sleepless nights at the ease with which he had destroyed Matenge's cattle speculating business. There were other things too – the pathetic way in which Matenge always backed down when confronted by a superior.

But if a man like Gilbert had really kept his mind on the Matenges who were an inverted whirlpool of seething intrigues, on the crazy semi-literate politicians like Joas Tsepe, he might have overlooked the kind of people almost everyone overlooked – the Dinoregos and Mma-Millipedes. At the most bitter times of Gilbert's stay in Golema Mmidi, Dinorego had always said: 'I think the Good God don't like it.' But he said it as though the 'Good God' was quite nearby, listening, observing, and Dinorego,

Beshero-Bondar et al., DSC 5



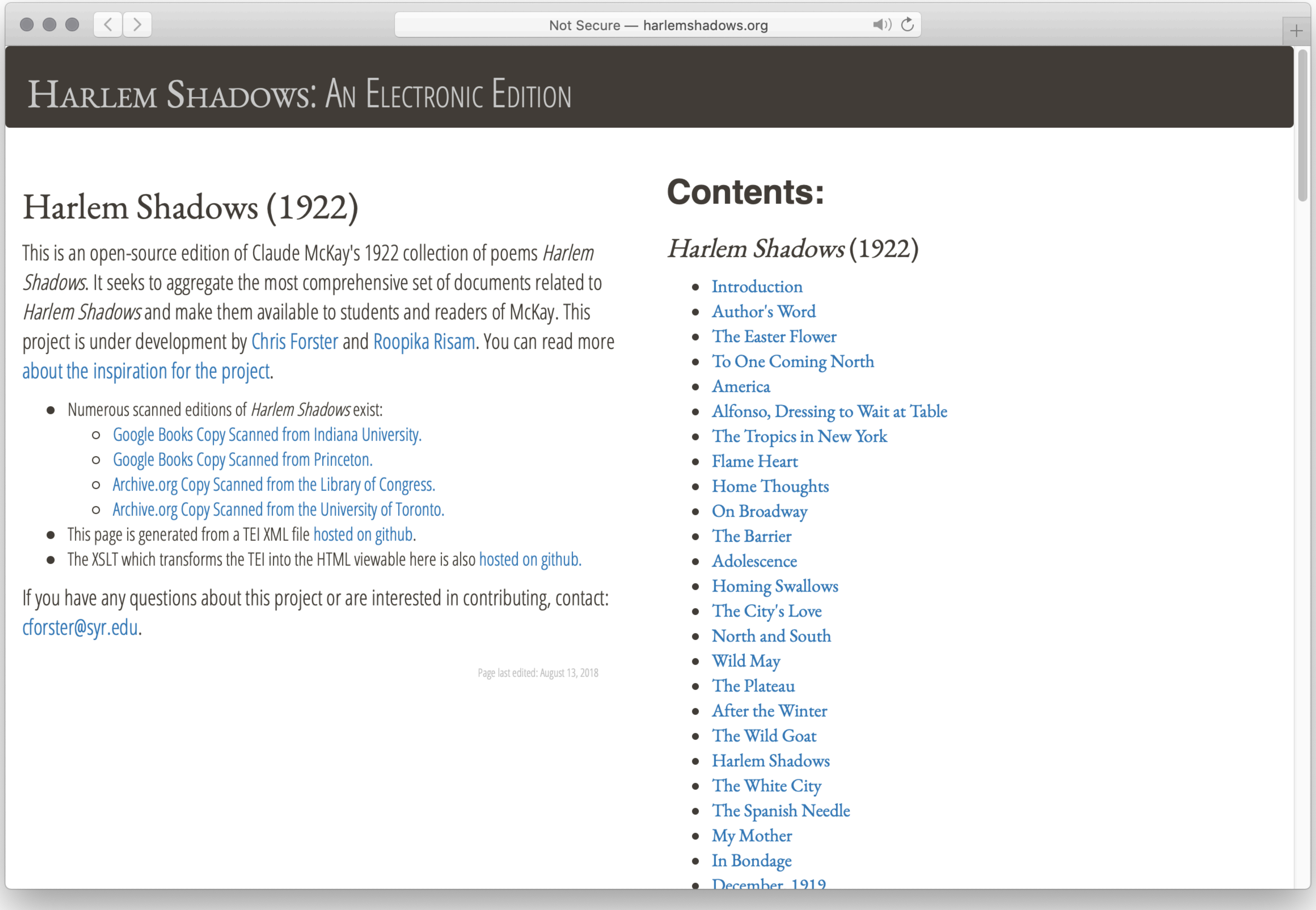
Carl Van Vechten, © Van Vechten Trust. Beinecke Rare Book and Manuscript Library, Yale University

Claude McKay, born Festus Claudius McKay in Sunny Ville, Jamaica in 1889, was a key figure in the Harlem Renaissance, a prominent literary

POEMS BY CLAUDE MCKAY

After the Winter

poetryfoundation.org/poets/claude-mckay



Harlem Shadows: An Electronic Edition

Harlem Shadows (1922)

This is an open-source edition of Claude McKay's 1922 collection of poems *Harlem Shadows*. It seeks to aggregate the most comprehensive set of documents related to *Harlem Shadows* and make them available to students and readers of McKay. This project is under development by Chris Forster and Roopika Risam. You can read more about the inspiration for the project.

- Numerous scanned editions of *Harlem Shadows* exist:
  - Google Books Copy Scanned from Indiana University.
  - Google Books Copy Scanned from Princeton.
  - Archive.org Copy Scanned from the Library of Congress.
  - Archive.org Copy Scanned from the University of Toronto.
- This page is generated from a TEI XML file hosted on github.
- The XSLT which transforms the TEI into the HTML viewable here is also hosted on github.

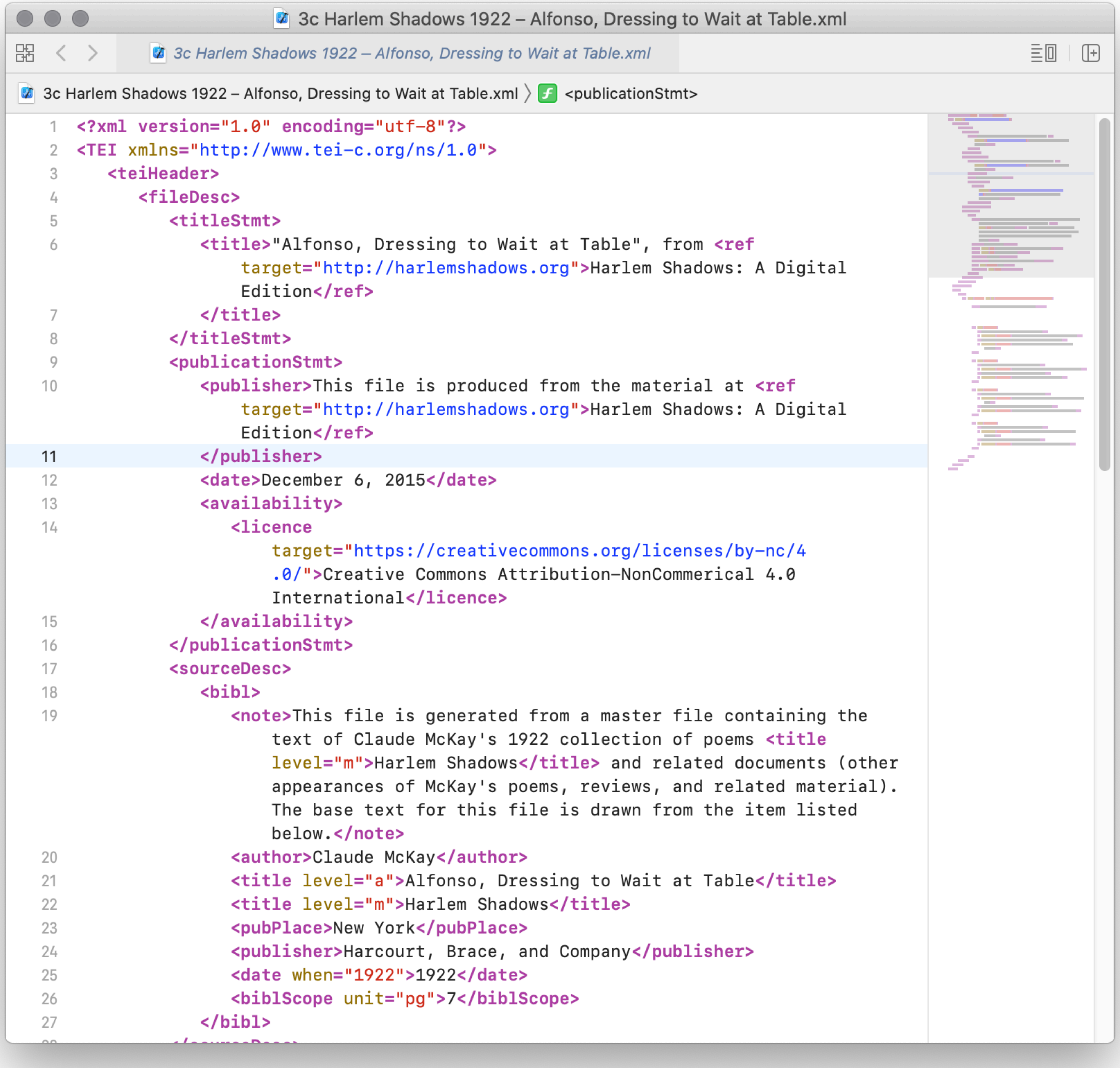
If you have any questions about this project or are interested in contributing, contact: cforster@syr.edu.

Page last edited: August 13, 2018

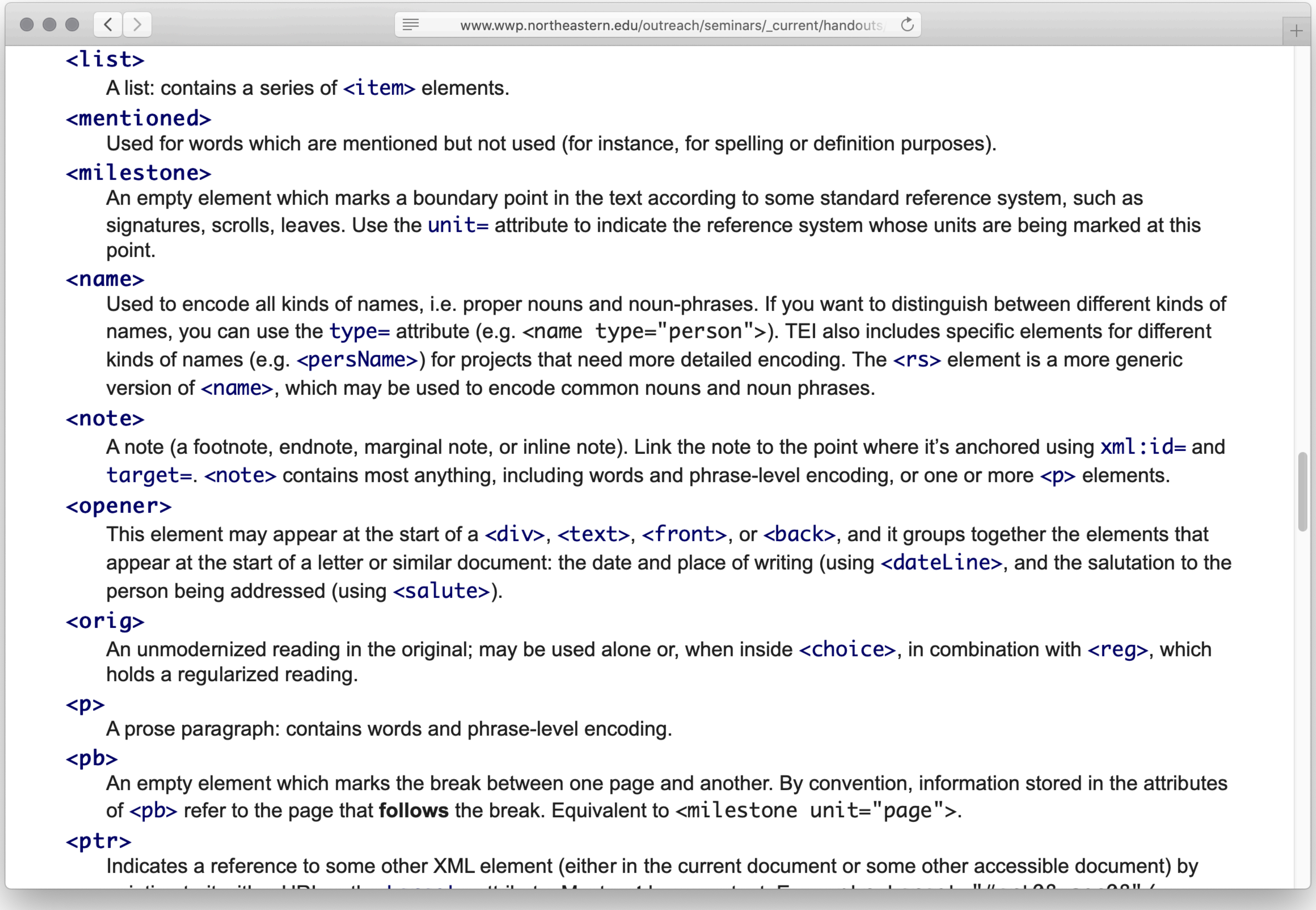
Contents:

Harlem Shadows (1922)

harlemshadows.org


3c Harlem Shadows 1922 – Alfonso, Dressing to Wait at Table.xml
3c Harlem Shadows 1922 – Alfonso, Dressing to Wait at Table.xml
3c Harlem Shadows 1922 – Alfonso, Dressing to Wait at Table.xml 〉 <publicationStmt>

```
<?xml version="1.0" encoding="utf-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>"Alfonso, Dressing to Wait at Table", from <ref
                target="http://harlemshadows.org">Harlem Shadows: A Digital
                Edition</ref>
            </title>
         </titleStmt>
         <publicationStmt>
            <publisher>This file is produced from the material at <ref
                target="http://harlemshadows.org">Harlem Shadows: A Digital
                Edition</ref>
            </publisher>
            <date>December 6, 2015</date>
            <availability>
               <licence
                  target="https://creativecommons.org/licenses/by-nc/4
                  .0/">Creative Commons Attribution-NonCommerical 4.0
                  International</licence>
            </availability>
         </publicationStmt>
         <sourceDesc>
            <bibl>
               <note>This file is generated from a master file containing the
                  text of Claude McKay's 1922 collection of poems <title
                  level="m">Harlem Shadows</title> and related documents (other
                  appearances of McKay's poems, reviews, and related material).
                  The base text for this file is drawn from the item listed
                  below.</note>
               <author>Claude McKay</author>
               <title level="a">Alfonso, Dressing to Wait at Table</title>
               <title level="m">Harlem Shadows</title>
               <pubPlace>New York</pubPlace>
               <publisher>Harcourt, Brace, and Company</publisher>
               <date when="1922">1922</date>
               <biblScope unit="pg">7</biblScope>
            </bibl>
```


www.wwp.northeastern.edu/outreach/seminars/_current/handouts

<list>

A list: contains a series of <item> elements.

<mentioned>

Used for words which are mentioned but not used (for instance, for spelling or definition purposes).

<milestone>

An empty element which marks a boundary point in the text according to some standard reference system, such as signatures, scrolls, leaves. Use the unit= attribute to indicate the reference system whose units are being marked at this point.

<name>

Used to encode all kinds of names, i.e. proper nouns and noun-phrases. If you want to distinguish between different kinds of names, you can use the type= attribute (e.g. <name type="person">). TEI also includes specific elements for different kinds of names (e.g. <persName>) for projects that need more detailed encoding. The <rs> element is a more generic version of <name>, which may be used to encode common nouns and noun phrases.

<note>

A note (a footnote, endnote, marginal note, or inline note). Link the note to the point where it's anchored using xml:id= and target=. <note> contains most anything, including words and phrase-level encoding, or one or more <p> elements.

<opener>

This element may appear at the start of a <div>, <text>, <front>, or <back>, and it groups together the elements that appear at the start of a letter or similar document: the date and place of writing (using <dateLine>, and the salutation to the person being addressed (using <salute>).

<orig>

An unmodernized reading in the original; may be used alone or, when inside <choice>, in combination with <reg>, which holds a regularized reading.

<p>

A prose paragraph: contains words and phrase-level encoding.

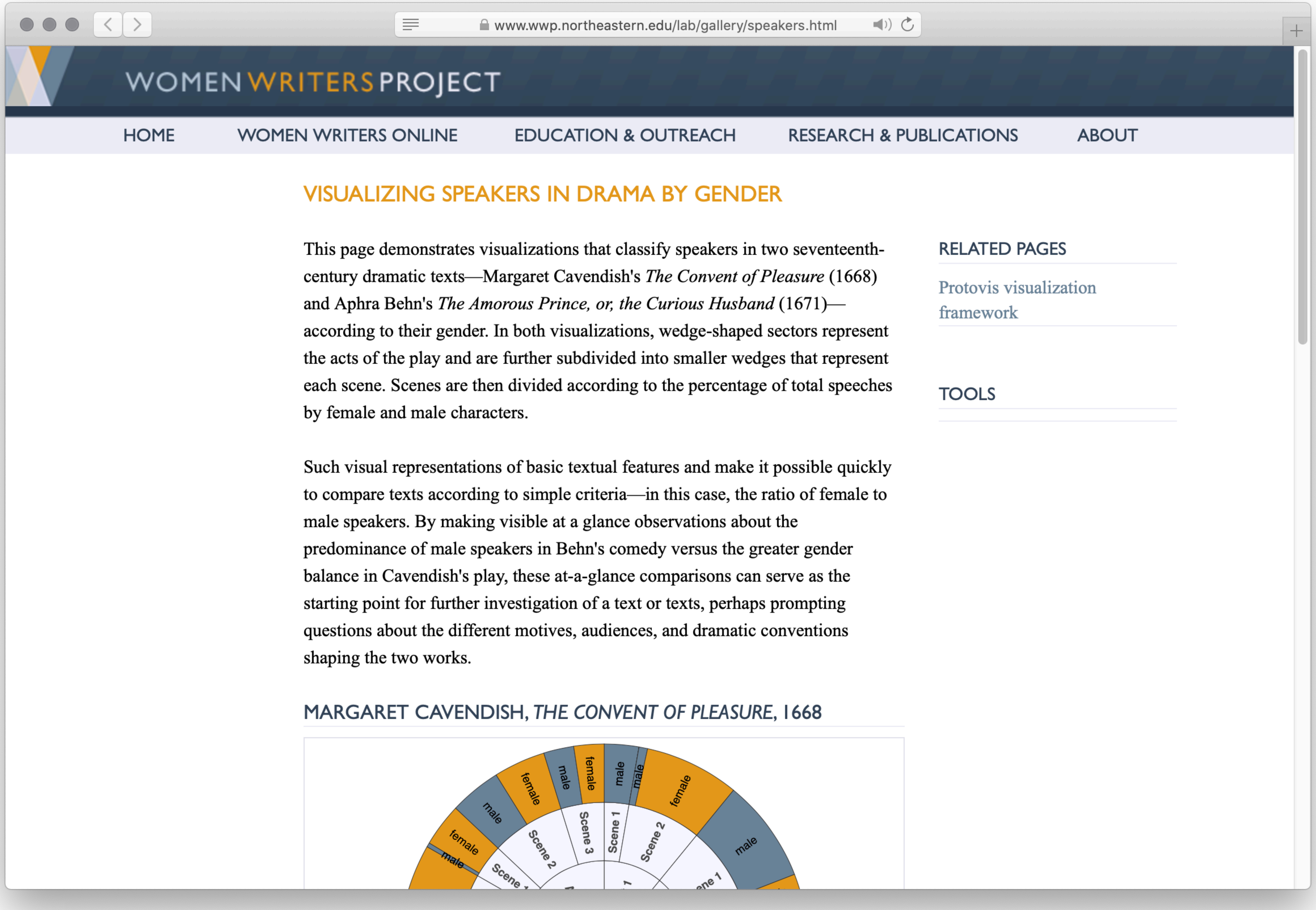
<pb>

An empty element which marks the break between one page and another. By convention, information stored in the attributes of <pb> refer to the page that **follows** the break. Equivalent to <milestone unit="page">.

<ptr>

Indicates a reference to some other XML element (either in the current document or some other accessible document) by

bit.ly/32H8YJl



VISUALIZING SPEAKERS IN DRAMA BY GENDER

This page demonstrates visualizations that classify speakers in two seventeenth-century dramatic texts—Margaret Cavendish's *The Convent of Pleasure* (1668) and Aphra Behn's *The Amorous Prince, or, the Curious Husband* (1671)—according to their gender. In both visualizations, wedge-shaped sectors represent the acts of the play and are further subdivided into smaller wedges that represent each scene. Scenes are then divided according to the percentage of total speeches by female and male characters.

Such visual representations of basic textual features and make it possible quickly to compare texts according to simple criteria—in this case, the ratio of female to male speakers. By making visible at a glance observations about the predominance of male speakers in Behn's comedy versus the greater gender balance in Cavendish's play, these at-a-glance comparisons can serve as the starting point for further investigation of a text or texts, perhaps prompting questions about the different motives, audiences, and dramatic conventions shaping the two works.

RELATED PAGES

Protovis visualization framework

TOOLS

MARGARET CAVENDISH, *THE CONVENT OF PLEASURE*, 1668

wwp.northeastern.edu/lab/gallery/speakers.html

Thoughts on the reading?

“In digital humanities, it seems like the underlying mechanism for discovery is ‘let’s try it and see what happens’, which is different than what most traditional literary scholarship does.”

Maria Sachiko Cecire, quoted in *DSC 2.*

But a good way to have fun and find the unexpected!