
E-Text: Creating, Encoding and Delivering

Perry Willett

Associate Director, DLP

Feb. 6, 2004

Overview

- Short history of etext
- How to create etext
- Encoding theory
- Encoding practice
- Delivery

Beginnings

“Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, “memex” will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.”

Vannevar Bush, “As We May Think,” *Atlantic Monthly* July 1945: 101-108

First Etext

“Announcements

“[... Father Busa] requests of scholars: [...]

b) any information they can supply about such mechanical devices as would serve to achieve the greatest possible accuracy, with a maximum economy of human labor. (Father Busa has been in contact with IBM in New York, the RCA laboratories in Princeton, the Library of Congress and the Library of the Department of Agriculture, in Washington.”

Speculum, 25.3 (1950): 424-425.

Through 1980s

- Character encoding (ASCII, EBCDIC)
 - Text encoding for input and output
 - No standards / all home grown
 - Not meant for human readability
 - Used for research, publishing
- “T*C1HE *C0P*C1UPIL” translates to “The Pupil”

Beginnings of standards

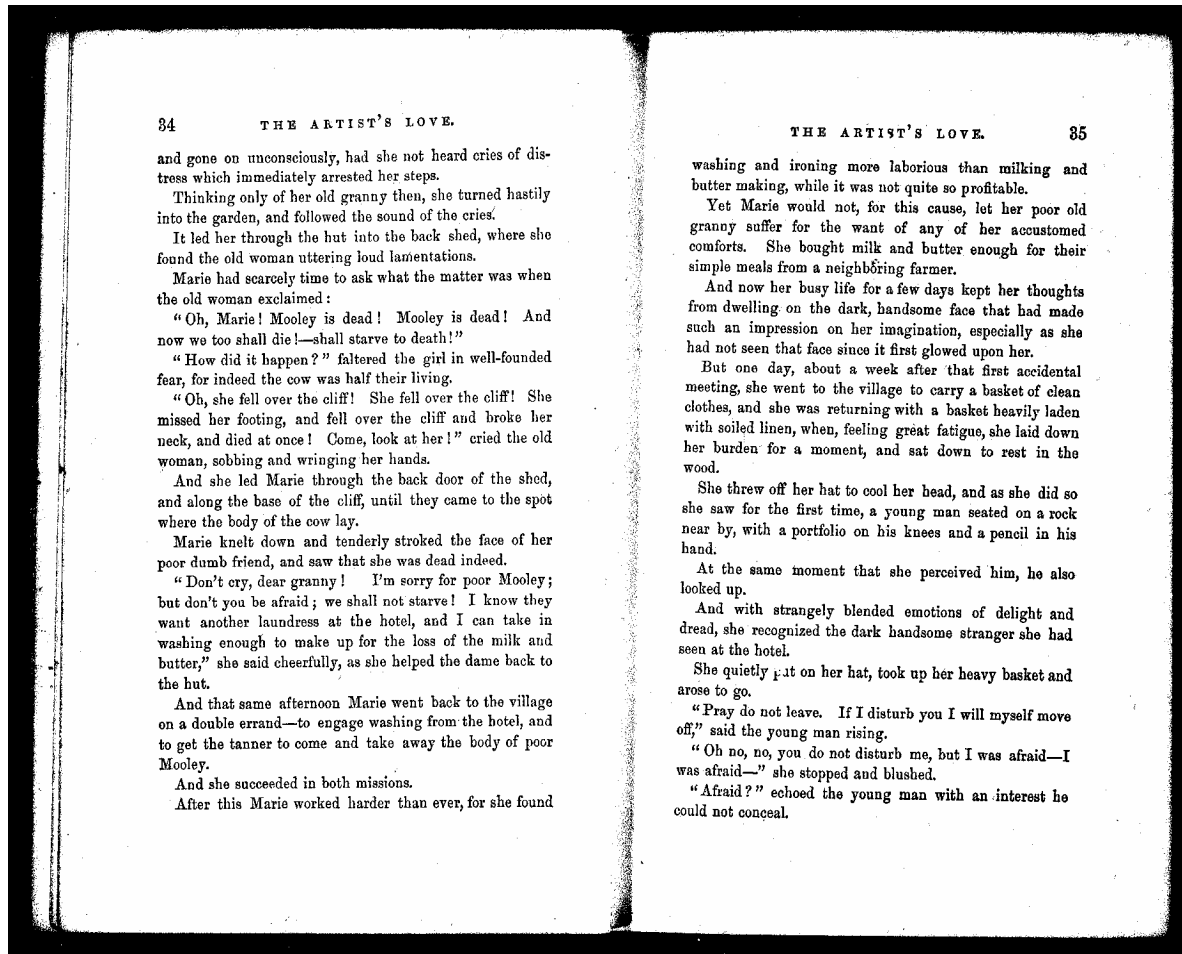
(Or at least, standard applications)

- GML=**G**oldfarb, **M**osher and **L**orie OR **G**eneralized **M**arkup **L**anguage
- TeX
- troff

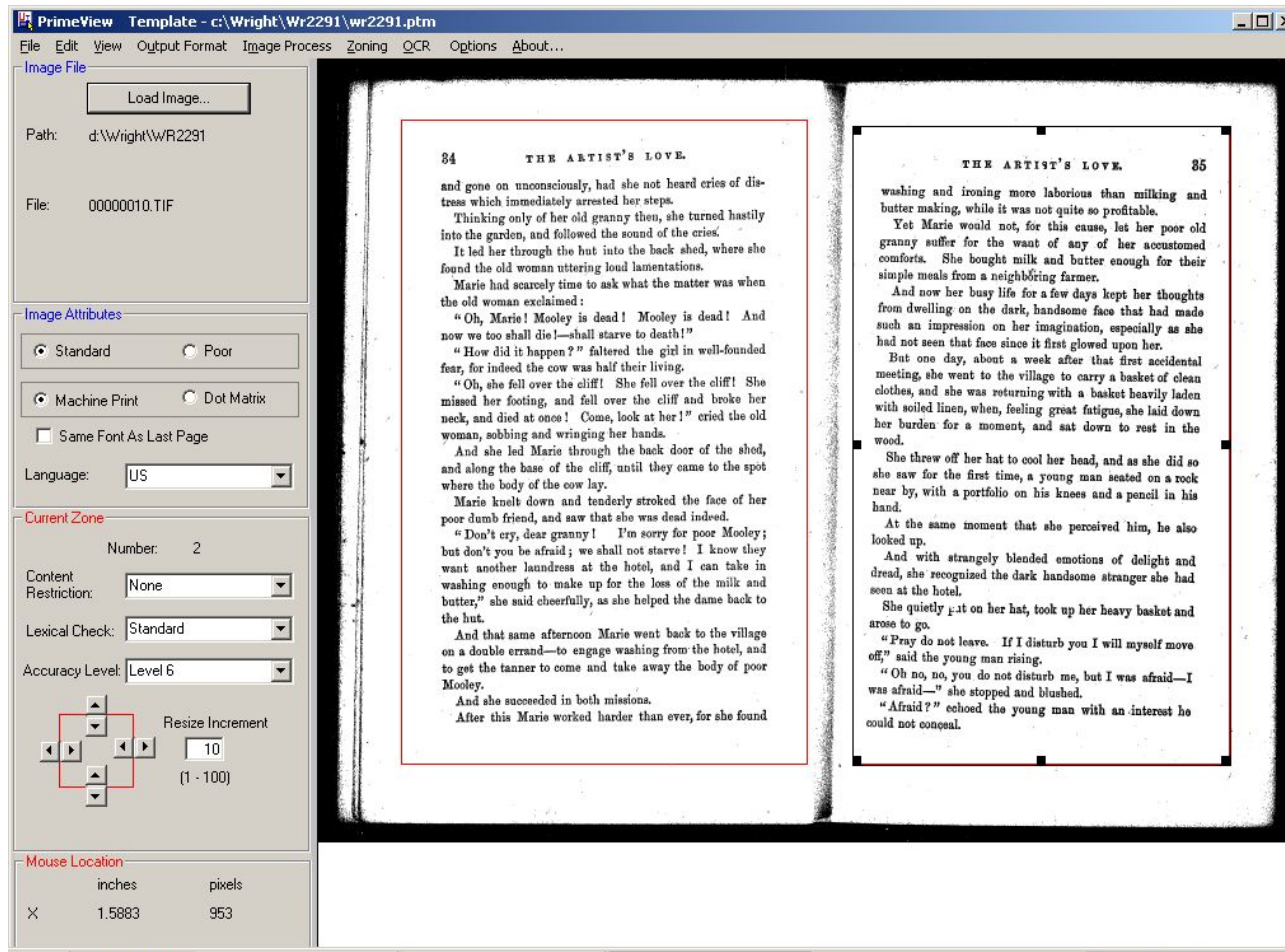
Creating Electronic Text

- Scanning / Optical Character Recognition (OCR)
- Transcription

Typical page



OCR program



Results

34 THE ARTIST'S LOVE.

and gone on unconsciously, had she not heard cries- of distress which immediately arrested her steps.

Thinking only of her old granny then, she turned hastily into the garden, and followed the sound of the cries.

It led her through the hut into the back shed, where she found the old woman uttering loud lamentations.

Marie had scarcely time to ask what the matter was when the old woman exclaimed:

"Oh, Marie! Mooley is dead! Mooley is dead! And now we too shall die!-shall starve to death!"

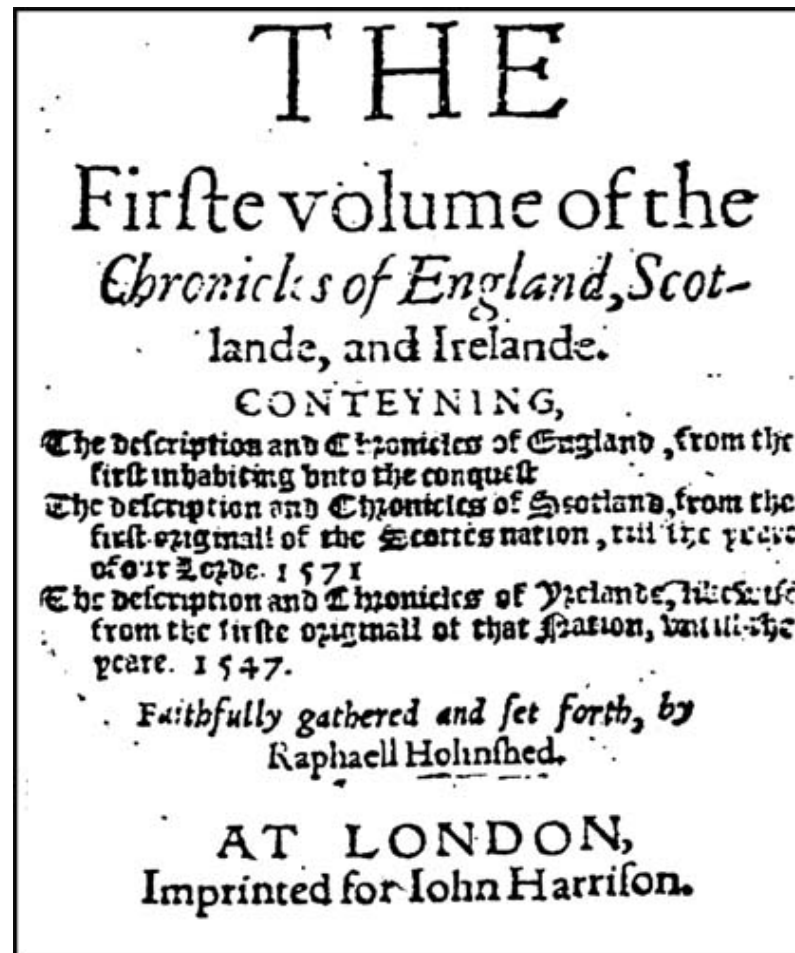
"How did it happen?" faltered the girl in well-founded fear, for indeed the cow was half their living.

"Oh, she fell over the cliff! She fell over the cliff! She missed her footing, and fell over the cliff and broke her neck, and died at once! Come, look at her!" cried the old woman, sobbing and wringing her hands.

And she led Marie through the back door of the shed, and along the base of the cliff, until they came to the spot where the body of the cow lay.

Marie knelt down and tenderly stroked the face of her poor dumb friend, and saw that she was dead indeed.

Page from Early English Books Online:



OCR output

```
~ ~k ~  
  
~ l I ~ li ~]J]O DmU~ov O~ii |  
  
~ ~1l ~ ~ -\O~Si~\r<,St~5,o t%,\~t,\~ ~ ~  
  
~' .-bnEIs~l br~; <~5n~1 ~  
  
~1 1~t ~3mo71~k~7noostI3o~rsd  
~i~mlm87il fif ~s ~  
  
~' 3,Ilmo~l.6n3~nm/l7~io\~ ~7g ~i  
  
...~ -,~. ;lIl~1B~ ]8 ~ . ~ ~ '  
  
'~`~@~ ~ ~`~pA til Sns t' - b~ ~I\U\ `i:~]  
~ ~ ~  
  
I I noin~Hodol~o]bsJni~qml '~1 11  
  
1~.1 ~ ~1 11  
  
" ~ ~ |?' ~ 9~ 9~] boO \~  
  
,,---. ~13 ~ ~ ~  
  
-: ~__ 1  
  
.
```

A word about accuracy

- No standard for calculating accuracy
- Rates determined by the software itself are suspect
- Expectations for accurate text are very high
- Error rates generally calculated per character, yet we search by word:

Quick brown fox jumps over the lazy dog.

1234567890123456789012345678901234567890

1

2

3

4

95% of characters accurate, while only 75% of the words.

Problems with OCR

- Time consuming and expensive to correct (perhaps not necessary)
- Languages other than English, particularly non-Western languages
- Handwriting:

http://www.dlib.indiana.edu/collections/nuer/scanned_nouns/094.jpg

Transcription / Typing

- Low tech solution—requires only a computer
 - high concentration required
- Double-keying provides higher accuracy without needing to proofread every word
 - Double-keying can achieve accuracy rate of 99.995% (1 error per 20,000 characters)
 - Triple-keying 99.999% (1 error per 100,000 characters)

Encoding

- Procedural vs. Descriptive
- Encoding standards

Procedural vs. Descriptive

■ Procedural

- ❑ specifies how to process text—“when you see this symbol, do this.”
- ❑ concerned with appearance and display
- ❑ has little to say about content
- ❑ HTML example:
<i>Italicized text here</i>

Descriptive markup

- Meant to separate content and display issues
- Indicates hierarchical structure
- OHCO theory: Text is an “Ordered Hierarchy of Content Objects”
 - sections within chapters within books
 - problems accounting for multiple hierarchies

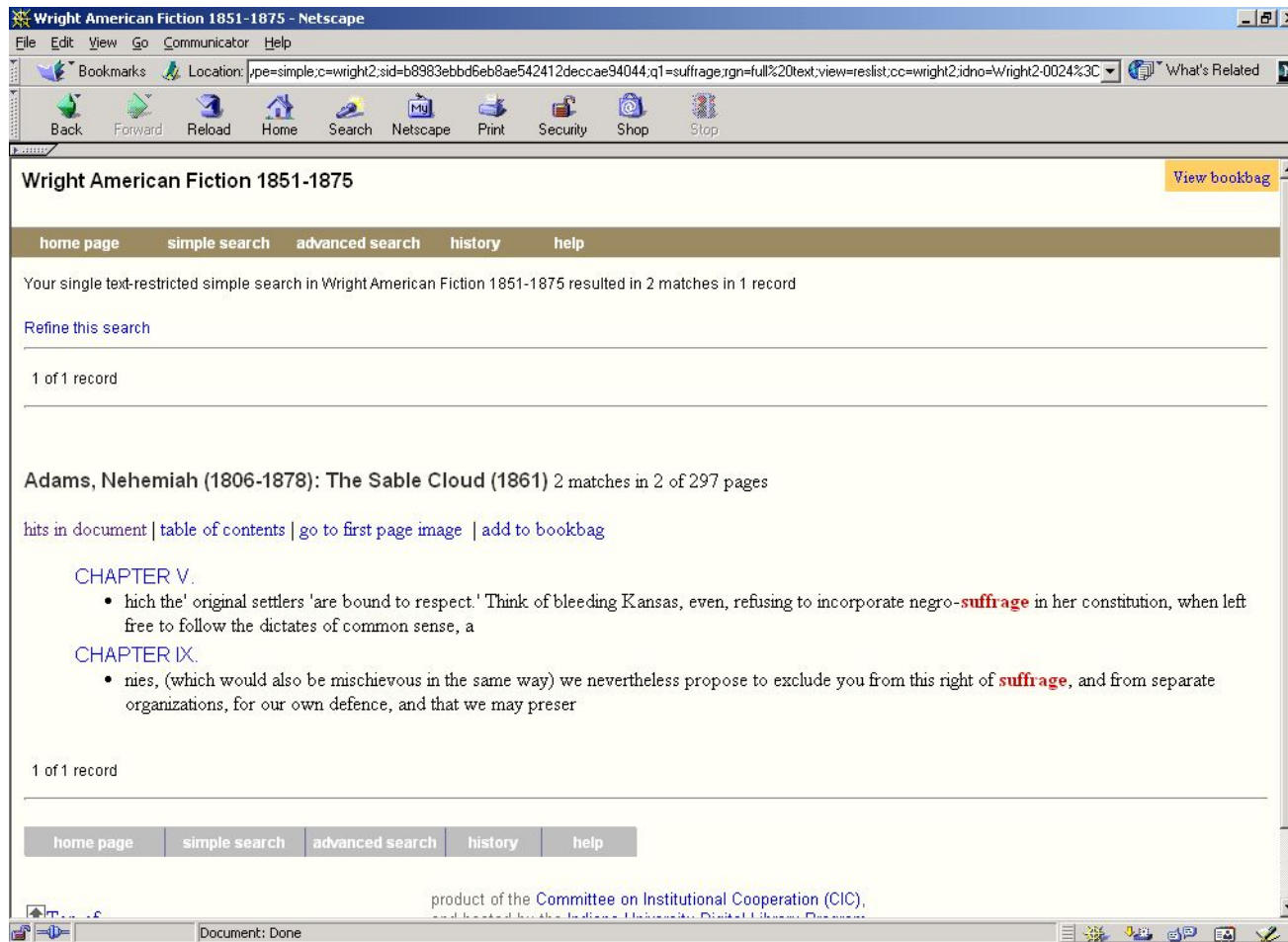
SGML

- ISO standard adopted in 1986
- Defines a “grammar” for markup languages
 - Elements
 - Attributes
 - Entity References
 - Document Type Definitions

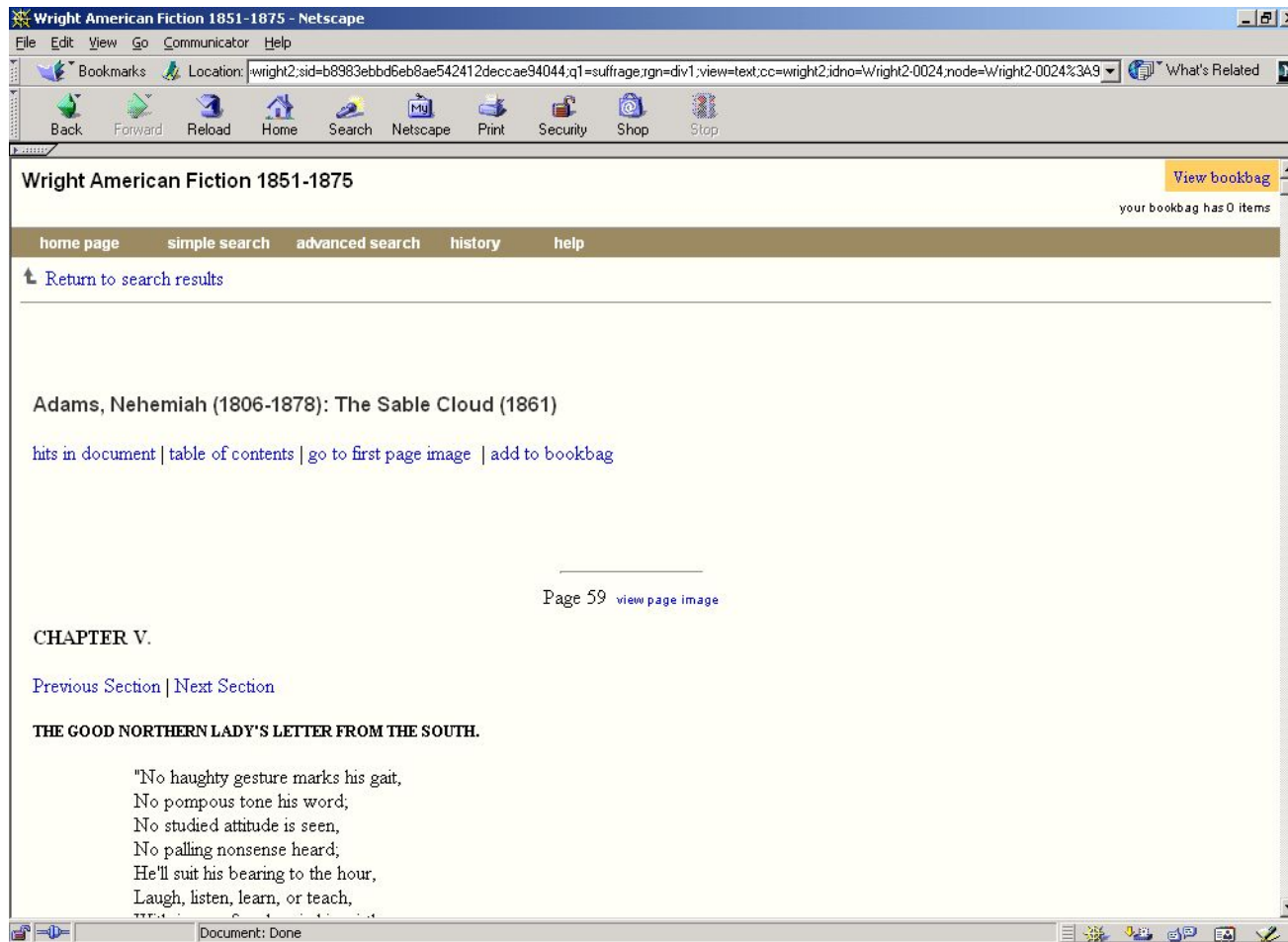
SGML [2]

- Should allow maximum flexibility in creating and using documents
- Multiple hierarchies? Use CONCUR
- Advantages:
 - different views for different people
 - <http://www.marthaonline.com/content/movie.html>
 - formatting separate from content
 - structured search and display

Structured search



Display



XML

- SGML was designed before the WWW
- Browsers did not display SGML
- The goals of XML:
 - Eliminate complexity
 - Facilitate use in WWW
- XML is a subset of SGML
- Unicode is the character set

Markup Languages

- XML <http://www.xml.org>
- Text Encoding Initiative (TEI)
 - <http://www.tei-c.org>
- HTML
- DocBook
 - <http://www.docbook.org>
- OpenEbook
 - <http://www.openebook.org>

What's the point?

- Non-proprietary, well-documented markup languages present the best opportunity for consistent, wide-spread adoption, and for long-term preservation of electronic text
- However, there hasn't been much software available for search and display of SGML/XML full text

The good news

- Growing number of open source digital library software for electronic text:
- Greenstone: <http://www.greenstone.org>
- eXist: <http://sourceforge.net/projects/exist/>
- Xindice: <http://xml.apache.org/xindice/>
- DLXS Lite: <http://www.dlxs.org>

Remaining Issues

- Wider adoption of Unicode
- Completion of XML components such as XLink and XML Query
- Better understanding of how people use (and want to use) text online

Bibliography

- Burnard, Lou, Michael Sperberg-McQueen and Syd Bauman. "A Gentle Introduction to XML." IN: Burnard and Sperberg-McQueen, eds. *Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium: 2001. <<http://www.tei-c.org/P4X/SG.html>>
- DeRose, Steven, David Durand, Elli Mylonas, and Allen Renear. "What is Text, Really?" *Journal of Computing in Higher Education* (1990) 1.2: 3-26. (OHCO Theory)
- Farrell, Elizabeth F. and Florence Olsen. "A New Front in the Sweatshop Wars?" *Chronicle of Higher Education*, Oct. 26, 2001: A35. (Ethics of data conversion industry)
- McGann, Jerome. "Radiant Textuality," *Victorian Studies* 39.3 (1995): 379-390. (Use of electronic texts by scholars)
- Willett, Perry. "Encoding Texts in the Humanities." In *Libraries, the Internet and Scholarship*. Thomas, C. Franklin, ed. New York: Marcel Dekker, 2002: 133-154.