Data and Culture

Prof. Andrew Goldstone (andrew.goldstone@rutgers.edu)
Prof. Meredith McGill (meredith.mcgill@rutgers.edu)

September 29, 2022. Edited and/or Plain Text?

motivation

- Why digitize Rutgers War Service Bureau letters?

motivation

- Why digitize Rutgers War Service Bureau letters?
- Why *edit* them?

upcoming assignments

- Exercise 2: short Cordell response (Canvas)
- Exercise 3: finished XML
  - valid XML
  - correctly annotate at least five persons, places, terms!
  - write a separate interpretive statement: so what?
    - short letter? write a little more
  - submit both files on Canvas

demo: Oxygen features

- Document: Validate: Validate
- Document: Validate: Check Well-Formedness
- previewing HTML output with XSLT
  - Follow Francesca Giannetti's instructions
  - Document: Transformation: Configure Transformation Scenarios
  - Document: Transformation: Apply Transformation Scenario(s)

demotivation

- Why *wouldn't* you create digital editions of historical or cultural documents?

Cordell warm-up

- Consider the handout excerpt from "The Raven" as it is available from the *Chronicling America* "text" representation of the *Lewisburg Chronicle, and the West Branch Farmer*, November 28, 1849, 1. Point out some notable features of this excerpt and write a few sentences about their significance.

some key terms from Cordell

- OCR
- digitized (vs. born-digital)
- machine reading
- surrogate
- provenance
- bonus: provenance

See also

- Cordell et al., Viral Texts Project, viraltexts.org
- Edgar Allan Poe Society, *Collected Works of Edgar Allan Poe*, www.eapoe.org/works

showdown!

- In your groups, brainstorm some of the differences *for scholarship* between "mass-digitized" and hand-digitized documents.

hand	mass
exact transcription	dirty OCR
small-scale	large-scale
editorial markup	high-level features not encoded
detailed metadata	questionable metadata
facilitates search	facilitates search
expensive	less expensive
highly selective	applied to large collections
individual treasures	lots of stuff
process may be visible	process largely opaque
edits primary sources	often remediates earlier surrogates
could be paywalled	could be paywalled
unrepresentative	unrepresentative
idiosyncratic display	overlay images for web display

Cordell's motive

- Why does Cordell think the bibliography of "dirty OCR" matters? Identify his scholarly *motive*.

Cordell's motive

- Why does Cordell think the bibliography of "dirty OCR" matters? Identify his scholarly *motive*.

The digital medium has already transformed humanistic research…. The predominant public interfaces of large-scale archives (focused on page images) and common modes of representing those materials in scholarship (a citation to the historical newspaper itself) encourage a fundamental misrecognition of the machine reading in which we are all engaged. (193)

describe!

- Locate one digital item representing a publication from before 1926
  - examine it in image form
  - find the text "edition"
- Describe it bibliographically as best you can (cf. Cordell)
  - what kind of format is it?
  - who made it?
  - what were the rationales for the digitization choices?
- Database assignments
  - Tables 1–3: Chronicling America: a page of a New Jersey paper
  - Tables 4–7: HathiTrust: any book you have read or heard of
  - Tables 8–10: Internet Archive: any book or magazine you have read or heard of

next

- Exercise 2: short response to Cordell due on Canvas, Monday 11:50 p.m.
- read Adler chapters on Canvas