ELTeC-eng zoom party report

ELTeC-eng

- Responsible: Just me, and the internet
- Current status: 99 texts (12.2 million words), E5C score 97.69

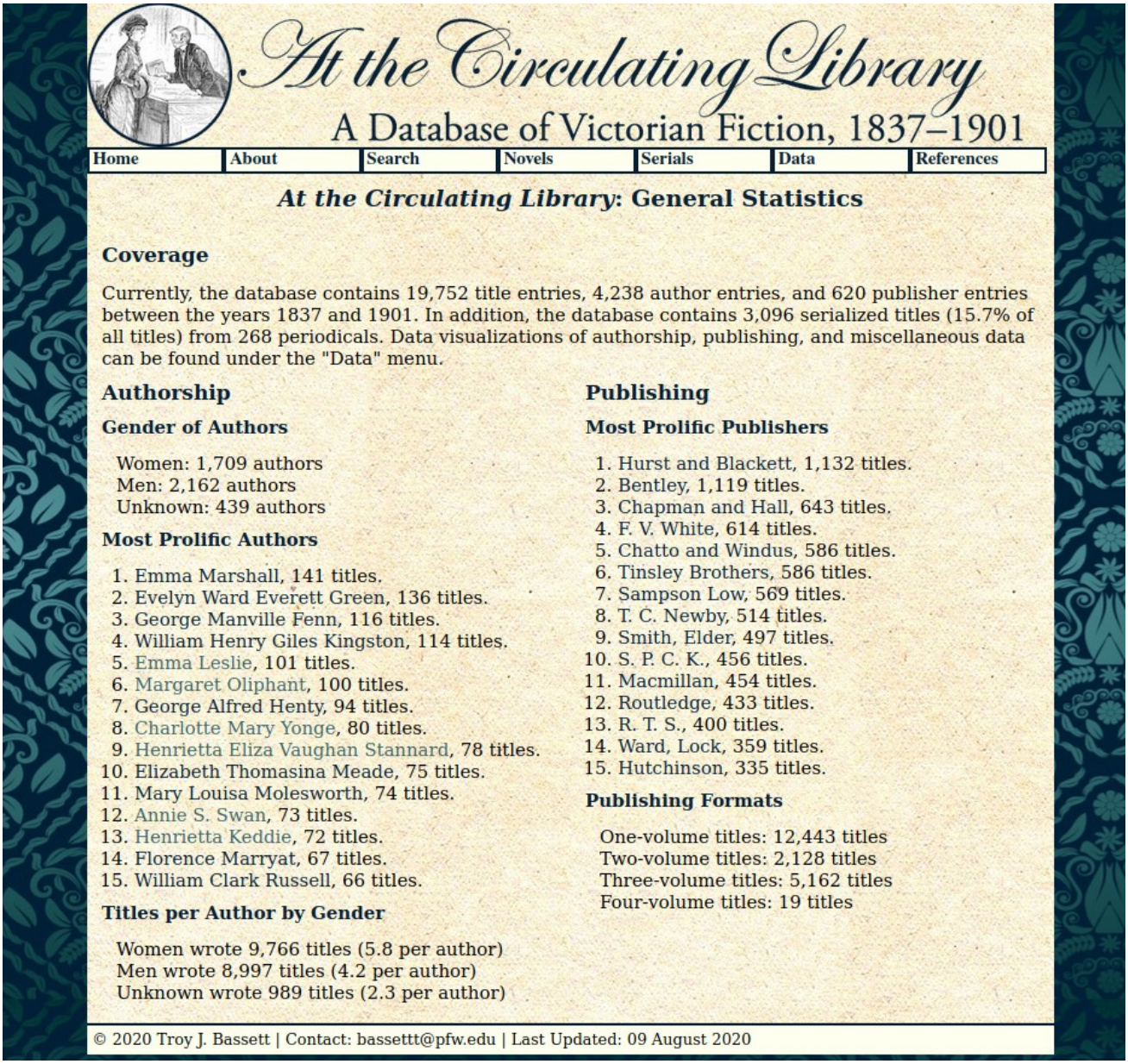
English has some (slightly) unfair advantages

- English literature is much studied and readily available in digital form
- Copyright deposit in major national collections (Bodleian, British Library, NSS etc.)
- Extensive digitization projects, though generally only page images

Finding the metadata

I relied mostly on Troy J Bassett's At the Circulating Library

- for bibliographic metadata
- as an approximation to the population being sampled



At the Circulating Library

A Database of Victorian Fiction, 1837–1901

Home | About | Search | Novels | Serials | Data | References

At the Circulating Library: General Statistics

Coverage

Currently, the database contains 19,752 title entries, 4,238 author entries, and 620 publisher entries between the years 1837 and 1901. In addition, the database contains 3,096 serialized titles (15.7% of all titles) from 268 periodicals. Data visualizations of authorship, publishing, and miscellaneous data can be found under the "Data" menu.

Authorship

Gender of Authors

Women: 1,709 authors
Men: 2,162 authors
Unknown: 439 authors

Most Prolific Authors

1. Emma Marshall, 141 titles.
2. Evelyn Ward Everett Green, 136 titles.
3. George Manville Fenn, 116 titles.
4. William Henry Giles Kingston, 114 titles.
5. Emma Leslie, 101 titles.
6. Margaret Oliphant, 100 titles.
7. George Alfred Henty, 94 titles.
8. Charlotte Mary Yonge, 80 titles.
9. Henrietta Eliza Vaughan Stannard, 78 titles.
10. Elizabeth Thomasina Meade, 75 titles.
11. Mary Louisa Molesworth, 74 titles.
12. Annie S. Swan, 73 titles.
13. Henrietta Keddie, 72 titles.
14. Florence Marryat, 67 titles.
15. William Clark Russell, 66 titles.

Titles per Author by Gender

Women wrote 9,766 titles (5.8 per author)
Men wrote 8,997 titles (4.2 per author)
Unknown wrote 989 titles (2.3 per author)

Publishing

Most Prolific Publishers

1. Hurst and Blackett, 1,132 titles.
2. Bentley, 1,119 titles.
3. Chapman and Hall, 643 titles.
4. F. V. White, 614 titles.
5. Chatto and Windus, 586 titles.
6. Tinsley Brothers, 586 titles.
7. Sampson Low, 569 titles.
8. T. C. Newby, 514 titles.
9. Smith, Elder, 497 titles.
10. S. P. C. K., 456 titles.
11. Macmillan, 454 titles.
12. Routledge, 433 titles.
13. R. T. S., 400 titles.
14. Ward, Lock, 359 titles.
15. Hutchinson, 335 titles.

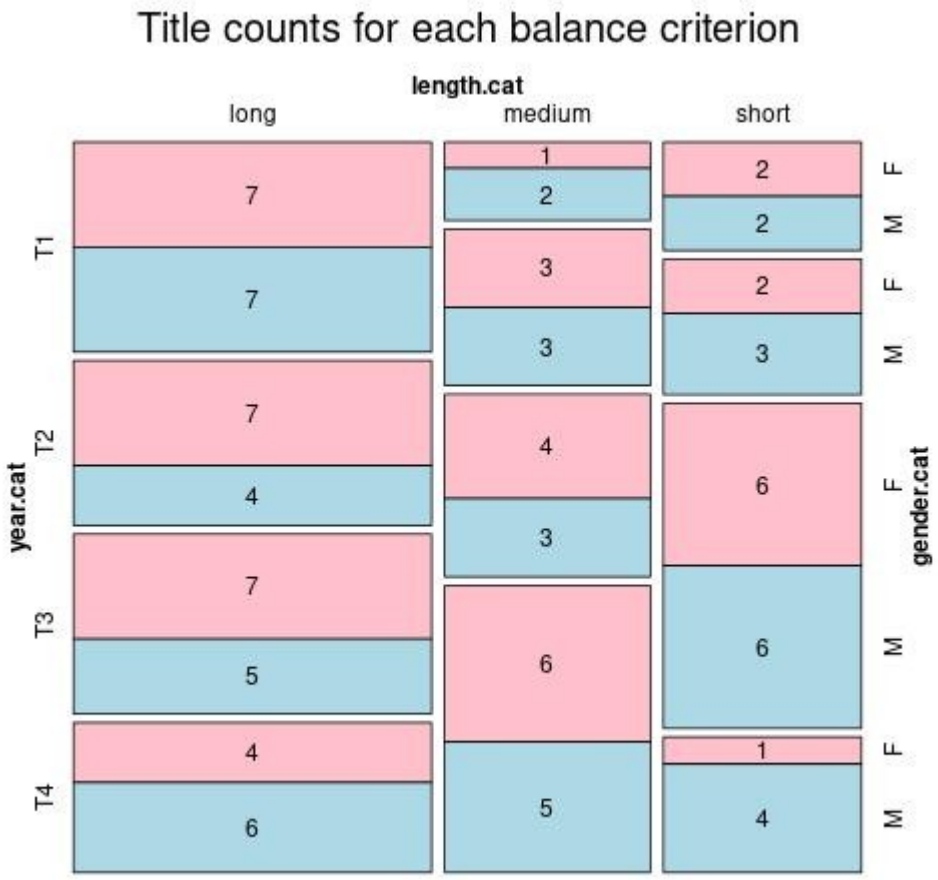
Publishing Formats

One-volume titles: 12,443 titles
Two-volume titles: 2,128 titles
Three-volume titles: 5,162 titles
Four-volume titles: 19 titles

Achieving balance

- Mostly a trial-and-error procedure

- I selected titles to represent each category in the four selection criteria (date, size, authorship, canonicity) equally
- I also tried to maximize variability in other respects (genre, subject matter, etc.)
- The mosaic graphic helps a lot, but doesn't show reprintCount



The sources I pilfered

- 13 from 19th c Fiction (Chadwyck Healey) -- TEI P2
- 14 from Victorian Women Writers Project -- TEI P5
- 62 from Project Gutenberg -- varieties of HTML
- 10 "other" :
    - 7 various flavours of HTML
    - 1 Oxford Text Archive XML
    - 1 Wikisource XHTML
    - 1 TEI P5 from Github
- Inconclusive experiments with OCR of some PDFs

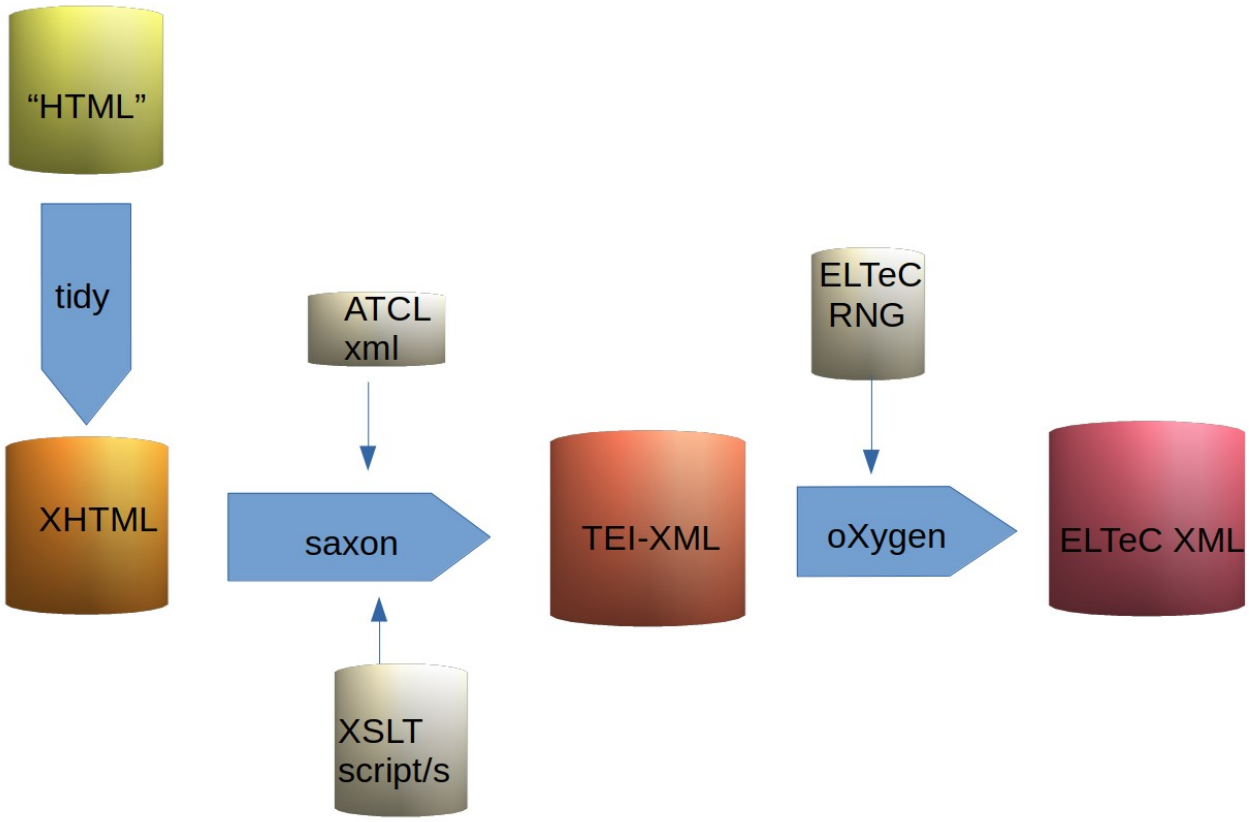
Dealing with data in the wild

- Sometimes the HTML provided was valid, quite often it was not (sometimes not even well-formed) : the preprocessing step was essential
- The TEI formats found also varied quite a lot, but were easier to down-translate
- The biggest problem in the HTML was the lack of consistency : the same feature (e.g. page number, chapter division) might appear in many different, and sometimes mutually incompatible ways
- Consequently, it proved impossible to write a one-size-fits-all converter (though I came close)
- Of course, the texts all looked much the same when rendered, but there were many different ideas about what encoding should be used to achieve that look
- Moreover, different decisions about *what* to capture in an encoding were commonplace ...

For example...

- not content with <emph>, some Gutenberg texts put ITALIC CONTENT INTO CAPITALS
- some decided to preserve highlighting of first words in a paragraph
- some texts had "semantic styles" aka meaningful @class values -- but they were all different
- chapter structure was usually not represented directly but had to be inferred from placement of headings, or special divs containing headings only
- footnotes and footnote references were represented in many different ways
- etc.

Semi-automatic workflow



Challenges

- Selection of texts for balance is *HARD*
  - I found it hard to resist obscure and amusing titles

        - and even harder not to be influenced by the availability of tractable digital versions
    - Coping with the messiness of existing digitizations is time consuming and sometimes frustrating

Future plans

- TEI Publisher instance
- Genre identification
- Other corpora