Image to Text: TEI + XML … or PITA?

David Denison
Linguistics and English Language

for 'Going Digital with Humanities Research'

Mary Hamilton papers

(Anson & Anson 1925: frontispiece)

- Family and friends
  - Mary Hamilton (1756-1816), courtier and diarist
  - old aristocratic family of military/political background
  - 1777-1782, governess to George III's daughters
  - 1783-1785, London, friends with the Bluestocking circle
  - 1785, married John Dickenson; daughter Louisa b.1787

Title	Mary Hamilton Papers
Dates	1743-1826
Extent	3 series; 2496 items
	Correspondence (1743-1826): 22 sub-series (+53), 2474 items
	Diaries (1776-1797): 16 autograph diaries
	Manuscript volumes (c.1779-c.1791): 6 volumes
Ref	GB 133 HAM
Held at	The University of Manchester, The John Rylands University Library

2

Project 'Image to Text'

- Mainly concerned with Mary Hamilton Papers
- Website **http://man.ac.uk/wm3Ws2**
- SALC website, but may one day appear on Rylands site

- Students edit one or two letters each in original spelling and layout, and code them in XML using TEI.
- Research assistant processes file, then all(!) three project members check it.
  - So far 161 letters = 70,722 text words

3

Purpose of transliterations

- For students
  - Engage with genuine texts some 200 years old
  - Solve puzzles in letters from knowledge of history of English language, plus reference works, etc.
  - *Learn* language history from examples found in letters
- For others (historians, general public, …)
  - Convenient rendition of historical documents which are often hard to read, sometimes hard to understand
  - Reliable text in standard format

4

Working assumptions of project

- For others
  - Full text as file in compliant TEI/XML form
  - Transliteration on-screen to help reader negotiate their way through the original
- For students
  - Minimise purely technical demands
  - Use modest subset of TEI guidelines (4 sides A4) for mark-up; our list modified if new problems encountered
  - Leave complex XML codes (e.g. for hyphenation) and TEI header to be added by research assistant

5

HAM/1/1/2/2, HAM/1/10/1/24

- Note treatment of
  - page layout, lineation and underline, superscript, etc.
  - words broken at end of line
  - abbreviations
  - corrections, additions, deletions, gaps
  - long-s
  - notes
- Website sets transliteration side by side with original
  Letter from Queen Charlotte to Mary Hamilton
  Letter from Anna Maria Clarke to John Dickenson

6

Physical vs. logical layout

- Desire for side-by-side display suggests
  - Preserve page- and column-breaks
  - Preserve line-breaks
- Keep all blocks of text together that appear on same page of original
- But then, what to do with
  - address panel written in middle of page?
  - insertions squeezed into any available space?
  - blocks of text written at 90° or upside-down or 'crossed'?

7

Layout problems

- Each XML line begins with <**lb**/> [= linebreak] tag
- Must other tags be opened and closed within line?
- Our pragmatic decision: No
  - Would cause wild proliferation of tags
  - Impossible e.g. for a name tag where forename and surname or title and name separated across break
- What about words broken at end of line?
  - XML contains both original layout (including 0, 1 or 2 hyphens!) and reconstructed whole word
  - Broken word displayed in blue, mouseover recombines

8

Layout problems

- Should L-to-R positioning be preserved?
- Our pragmatic choice:
  - Ignore indent for start of paragraph
  - But if line starts more than halfway across page or column, mark as align-right
- Thus just two options (and no use of tabs or spacing)
  - Conceals paragraph breaks unless writer left noticeable vertical gap – but para breaks often uncertain anyway
  - Preserves familiar position of date and place of writing at top right, and closing salutation at bottom right

9

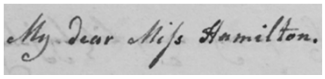
Layout problems

- Address is on page of letter (no envelopes then); when unfolded, address appears in centre of page.
- Should address interrupt XML text or not?
  - Eventual pragmatic decision: only if clean break in letter content with no text on both sides of address
- Always a compromise between
  - rearranging all blocks of text **on a single page** in logical order – letter content, postscripts, address, insertions at side, etc. – and
  - guiding reader to relevant parts in handwritten original

10

Representation of text

- Show long-s? If so, how?
- Yes, using Unicode long-s character ſ:
  - miſs, paſsion, etc.
- Good for TEI compliance ✓
- Screen character lacks descender (except when italicised *miſs, paſsion*!) and has crossbar
- 'Latin small letter esh' looks much better:
  - miʃs, paʃsion
- but non-compliant! ×



11

Representation of text

- Correct or modernise spelling?
- Our choice: keep original spelling throughout
  - If spelling differs from present-day standard, leave unmarked if current at the time (as shown in *OED*)
  - Otherwise mark obvious slips or idiosyncrasies for correction – with on-screen pop-up
- But eventual linguistic tagging and parsing may require a different normalisation

12

Plain text version

- For users such as linguists who want content only without reference to handwritten original
  - Text rearranged to maximise continuity
  - Corrections, unclear text, recombined broken words silently included
  - No long-s or macrons, but otherwise original spelling
  - Barely any mark-up: only filename, change of hand, discontinuity, gap (these tags within carets, not XML)
- Users advised to check examples against XML text

13

TEI/XML advantages

- International standard, independent of O.S.
- Allows pleasing on-screen display using only desired parts of text and mark-up. We choose blue colour to indicate where mouseover would produce a pop-up.
- Metadata from file can be presented elegantly.
  - display via server-side XSLT, CSS and javascript.
- Mark-up allows sophisticated search in XML files or by collation of fields in a database, e.g. of
  - names, places, dates, foreign words, corrections, etc.

14

TEI/XML disadvantages

- Almost no end to what can be tagged
  - We decided to tag all names, places and dates
  - We mark some structural features important in letters, e.g. <dateline>, <opener>, <salute>, <closer>
  - Inevitable arbitrariness at times
- Official TEI guidelines long and daunting
- Some inconsistencies in the documentation
- Tiny, inconspicuous errors can block processing of file
  - We use Notepad++ for editing and oXygen to review for errors

15

Residual questions (and 1 answer)

- Whether to code with screen display or archival correctness as prime concern
  - For screen display, where is best compromise between helpfulness and fussiness
  - Whether to assume 'screen' = large monitor, tablet or smartphone
- How many users who request archive are actually using TEI/XML version, and how
- We were never in doubt that TEI/XML was the right choice for master-copy of text.

16

Acknowledgements

- Nuria Yáñez-Bouza (joint project director)
- George Bailey, Donald Morrison (research assistants)
- Fran Baker (archivist at JRL)
- Lisa Crawley (archivist for Hamilton Papers)
- Carol Burrows and the Digitisation Steering Group
- enthusiastic students at Manchester and Vigo
- various expert advisers on XML, TEI, XSL and CSS, eighteenth-century London, sea silk, ...

17

URLs

- This presentation available from my downloads page:

http://tinyurl.com/UMan-DD

- **Image to Text** project website:

http://man.ac.uk/wm3Ws2

- Comments welcome! Thank you.

18