Pointy Brackets for Poets

Syd Bauman

2017-08-02

Brought to you by
Northeastern University Libraries
Digital Scholarship Group
Women Writers Project

Balisage 2017
Pointy Brackets for Poets

• Q: Can an English major use XML?
• A: Yes!

• Questions?
  s.bauman@northeastern.edu
“Please, sir, I want some more”

- OK, but caveats:
  - Some of what follows is more impressionistic memoir than fact-checked history
  - I don’t plan to differentiate

- Plan:
  - Overview of WWP
  - Tools we used for transcription
  - Tech savvy of English majors: 2 examples
  - How we teach XML (to humanists)
    - Iff time: Tech savvy French professor
    - Tangent iff time: public safety message version of well-formedness
    - (Obligatory oXygen screenshot included)
Women Writers Project, aka “WWP”

• ~1986: English professors looking at the *Norton Anthology of Literature by Women: The Tradition in English* thought “they’re missing ...”.

• They imagined creating an add-on anthology of works by authors not represented in *Norton*.

• Allen Renear came along and said “why not use the computer?”.

• So — the purpose was to correct the canon of English literature by publishing a computer textbase of works written by women in English, published between 1330 and 1830.
WWP hx

- Grant submitted in 1987
- Funding started in 1988
- I was hired 1990-07-01
- Brown University
  - 1994–2009: Computing and Information Services
  - 2009–2013: Library

- Northeastern University
  Library / Digital Scholarship Group
Student Employees

- Very little advertising, if any (1st job posting in 1997-05)
- ~200 of ’em over the years
- Plurarity English students (both undergraduates majoring in English and graduate students)
- A smorgasbord of other departments as well
What do they do?

- **85%**:  
  - Mostly transcription, nowadays in XML  
  - Proofreading, both markup and content  
  - Editing (i.e., entering corrections)

- **15%**:  
  - Helping develop our markup system & practices  
  - Writing (documentation, newsletter articles or blog posts, papers, etc.)

- Several have gone on to markup or digital humanities careers  
  - Carole Mah, **Jula Flanders**, **Paul Caton**, Jacque Wernimont, John Melson, Alicia Peaker, **Sarah Connell**, and **Sarah Stanley** all jump to mind
Growing pains

- WWP went through roughly 5 stages of markup growth
  - Fumbling beginnings
  - Into stride
  - Dark ages
  - [transition year]
  - Renaissance
Stage 1: Fumbling

- Waterloo Script → GML → SGMLish
- IBM VM/SP mainframe system
- Programs written in Script, REXX, and CMS Pipelines
- Students use XEDIT, little to no knowledge of markup
- Decent documentation for systems, but almost none for encoding
- Almost no support to speak of
  - Have a problem? Ask Syd
- Profoundly broken, but not unlike many modern projects started by faculty who do not know better
Stage 2: Into stride

- SGML → XML
- Development of markup training materials
- Documentation improved
- Cultivated a thriving intellectual community
- Moved from XEDIT on mainframe to Emacs on a Unix system (1st Emacs/psgml, then Emacs/nxml)
- Tools continued to improve
- Reasonable level of expertise attained
- High volume of good data generated
Stage 3: the Dark Ages

- Mostly undergrad students, often for only 1–2 semesters; and fewer of them at one time
- Thus loss of continuity and intellectual intensity
- Switched from Emacs/psgml to using oXygen on Mac OS X; data in Subversion
- Despite tighter schema and simpler system, less student expertise, thus encoding less competent
Stage 4: Transition

• We moved from Brown to Northeastern
  – Northeastern wanted Julia, took the whole project in order to get her

• Very different institutional attitude
  – Eagerness and expectation, not just tolerance

• Nonetheless, moving takes time (Solaris to RedHat, new bureaucracy, entire new crop of students, etc.)
Stage 5: Renaissance

- Expert students — almost all English grad students
- Students are remaining longer — several are pushing 4 years
- Enthusiasm and institutional moral support
- Better documentation & training
- Once again a cultivated thriving intellectual community
Why so much better?

- Departmental attitude
- Student’s research interests more aligned with DH
- “‘Digital Humanities’ has more cultural cache than ‘Humanities Computing’ ever had.” —JF
- Improved documentation
- “[We] had to dive in and go at it whole-heartedly and rely on documentation because there was no one else between us novices and [Syd & Julia].” —SS
Transcription Tools

• 1988–1999: XEDIT
• 1996–2010: Emacs/psgml
  – Mostly by Syd, but temporary help from a student — more later
• 1997: Great Author/Editor vs Emacs/psgml Experiment
  – Next slide(s)
• 2010–2013: Emacs/nxml
• 2013–present: oXygen
  – Note: we were teaching others using jEdit then oXygen since 2003, but internally using Emacs until 2013-01-28
Author/Editor

- Commercial product from SoftQuad
- “the first specialized SGML editor”
- A GUI with menus and buttons, etc
- Represented tags with icons
  - Eliminates whole categories of `<error></error>`
- At least Mac & Windows (Unix?)
- I think we used it on Mac with Unix disk mounted for data
That others in Spain were as capable as I in undertaking my mission.

I remind you of the story of the egg. Upon my return from the first voyage, I was greeted with great acclamation, and named Admiral by Their Majesties, Pedro González de Mendoza, Grand Cardinal of the Spanish Empire, a man of great wisdom, piety, and position, invited me to dine with him, and accorded me such honours as he might have bestowed upon King Ferdinand himself.

A jealous person, of no value or station, accusing me of being a foreigner (I, who had served Their Spanish Majesties faithfully and vigorously), asked me if there were not other men in Spain who could have accomplished the discovery of the Indies.

I did not reply, but took an egg from the plate in front of me, and passed it to the man beside, inviting him and all there present to make this egg stand on its end.

Each guest tried this test, and the egg was passed on to the next, each as unsuccessful as the one before. After some time, the egg was passed back to me. I held it in one hand and struck it hard on one end, so as to break the shell. I left it standing on its end on the table.

From Christopher Columbus Answers All Charges.
Emacs/psgml

- “the extensible, customizable, self-documenting, real-time display editor”
- Steep learning curve; like a high explosive
- Ran on Unix, thus via an X-window on encoder’s Mac
- Very powerful, keystroke-oriented, non-intuitive
  - C-x v = SGML validation
  - C-C v = extra validation (“supraSGML”)
  - C-v = down page
- Psgml-mode, an add-on, specialized it for SGML
Experiment

• The Plan:
  - ½ of encoders use A/E
  - ½ of encoders use Emacs/psgml
  - After 3 months we switch
  - After 3 more months we interview encoders, decide which to use forevermore, and maybe get a paper out of it

• The best laid plans of mice and men …
Student programmer extraordinaire

- Tom Hinkle, comparative literature: literary translation ’02 (MAT ’03)
- IIRC, joins us in early 2002
- **No** tech background *at all*
- Has what it takes:
  - Smart
  - Logical
  - Lazy
  - Does not like repetative boring tasks
Automation

‘Automating’ comes from the roots ‘auto-’ meaning ‘self-’, and ‘mating’, meaning ‘screwing’.

— https://xkcd.com/1319/
English and XPath

- TEI[
  teiHeader
  /fileDesc
  /titleStmt
  /author
  /persName[
    @ref[contains(.,'mcavendi')]
  ]
]//text//*[@xml:lang]/local-name()
Teaching: stage 2 and on

Audio: click above to play ~/Music/03 T-E-I.ogg OR ~/Music/iTunes/iTunes Music/Ursa Minors/Songs to Tag By/03 T-E-I.mp3

Link to original archived posting

Song and lyrics also available in the TEI Songbook

Local, corrected version of posting is at ./TEI-L_1996-10-13_updated.txt
Teaching English Majors XML (1)

• Self-selected audience
  – Yes, in some sense we’re cheating :-)

• Start from the problem space:
  – XML is for organizing data
  – Encoding is about representing texts
    • As much about what we throw away as what we keep
  – Goal is for the computer to help us with our research
Teaching English Majors XML (2)

- We do not treat encoding and XML mechanistically, but rather as an intellectual exercise.
Teaching English Majors XML (3)

- An intellectual exercise that we find interesting, important, useful ...

Students don’t care how much you know until they know how much you care.

—Anonymous, clearly based on a quote attributed to both John C. Maxwell and Theodore Roosevelt
Teaching English Majors XML (4)

• We treat it as interesting and important, not as “don’t worry your pretty little head about that”
• Conversely, we don’t panic if we don’t cover all the details
• But issues that are elided are treated as interesting stuff we don’t have time to cover
Teaching English Majors XML (5)

- There are things to memorize, but we firmly believe that:

The test of a good teacher is not how many questions he can ask his pupils that they will answer readily, but how many questions he inspires them to ask him which he finds it hard to answer.

—Alice Wellington Rollins
anti-quote

Metaphors and similes are wonderful literary devices but they do nothing but obscure the beauty of technology.

—Charles Petzold
In my experience, communication is a matter of patience, imagination. I would like to believe that these are qualities that we have in sufficient measure.

Cpt. Jean-Luc Picard, ST:tng (5)2 [45047.2]
Room Matters