Yuri Rubinsky would have been 65 years old today.

[ Items in these notes that are enclosed in square brackets were added ~3 weeks after the presentation. ]
Pointy Brackets for Poets

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

s.bauman@northeastern.edu

How many in the audience are humanities majors? English majors?
“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)
    - If time: Tech savvy French professor
    - Tangent if 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
- Plurality English students (both undergraduates majoring in English and graduate students)
- A smorgasbord of other departments as well
  - Classics, Environmental Science, Linguistics, Physics, Literature & Society, Psychology, Education Studies, International Relations, Religious Studies, Women's Studies, American Civilization, Health and Society, Anthropology, Development Studies, Applied Math, ...

[ By “advertising” I mean things intended to convince the student that this is a great job; by “posting” I mean a notice that says little more than there is a job availability, the job title, and contact information.

By “of ’em” I mean of student employees, not job postings :-)

The list of other departments is not scientific at all — I simply went through the list of previous student employees, and in no particular order looked up where which department each of the first 40 or so came from.
]
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

- [I completely made up the percentages]
- OCR not really an option [Probably hard to OCR our older stuff even today, but in 1991? No.]
- Outsourcing to a vendor defeats one of our purposes: to teach students DH methods and XML
- [Major point here is that students are part of the decision-making process of our markup system. We’re not a democracy — staff still has the final say. But most any significant decision gets input from the students first.]
- [Those in bold have presented at Balisage, its pre-conference symposium, or its ancestor conference. The one in italics is on the TEI Technical Council.]
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

[By “SGMLish” I mean we had files that would be SGML after a few programmable tweaks.

- Waterloo Script is a Touring-complete language; Steve DeRose (in the audience) wrote a CS final project in it just to prove he could. REXX is a wonderful scripting language for VM, and CMS Pipelines — well, I don’t have enough good things to say about CMS Pipelines. They’re like Unix pipe commands on steroids.]

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 (1\textsuperscript{st} Emacs/psgml, then Emacs/nxml)
- Tools continued to improve
- Reasonable level of expertise attained
- High volume of good data generated

[ Official conversion from SGML to XML was on 2003-06-11. ]

[ The cultivation of a thriving intellectual community among the student employees was absolutely deliberate, at least on my part. ]
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

[ I don’t know why this happened. Perhaps there was less money, perhaps we (the staff) were putting less effort into the students, perhaps the stars were mis-aligned. ]
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-heartedely and rely on documentation because there was no one else between us novices and [Syd & Julia].” —SS

[ I had little to no interaction with actual English professors, but it seems to me from a 3rd or 4th-hand distance that the attitude of the Brown English department was “we are training you to be an English scholar, a professor; of course, work at the WWP because you need money, but don’t let it deflect you from your true course”; whereas the attitude of the Northeastern English department seems to be “You have the chance to work on digital humanities? With Julia Flanders? Go!” Sorta like Captain James T. Kirk towards the end of Star Trek (2009): “Do it, do it, do it!”.

[ JF = Julia Flanders; SS = Sarah Stanley. ]
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

• 1988–1999 XEDIT
• ~1996–2010-01-29 Emacs/psgml
• 2010-01-29–2013-01-28 Emacs/nxml
• 2013-01-28–present oXygen
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 in China. Upon my return from the first voyage, I was greeted with great adoration and seemed most fortunate by those who admired me. I was presented with great honor and position, and invited to dine with them, and accorded the highest honors as the Greek and the Roman emperors had bestowed.

A jealous person, out of rivalry or ambition, accosting me of being a foreigner who had served their Spanish Majesty faithfully and vigorously, and asked 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 me, and asked him and all those present to make the 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.

Copyright 1993, Win Robinson and Marc Gassonell

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

[ On bullet point #2: My dad told me this story. Not sure if it was his high school physics teacher or college physics professor, but he stood in front of the room with a small quantity of RDX in his palm, struck a match, and in a relaxed manner, touched it to the RDX. Nothing happened. He then took a roughly equivalent amount of gunpowder (black powder not modern gunpowder; i.e., sulfer, charcoal, and saltpeter) and put it in his palm. He struck a match, and holding his hand as far away from himself as he could, turning his head away and grimacing, touched it to the gunpowder. **POP!** He had a small 1\textsuperscript{st} degree burn on his hand. The point is that although the RDX is really powerful (~1.5 x more powerful than TNT), it takes a *lot* to set it off, to use it. Same w/ Emacs. Really powerful, but it takes a *lot* to use it. ]
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 …

So, we had 7 or so encoders in each group, after 3 months we went to switch and the A/E kids moved over to using Emacs, and the Emacs kids … refused to switch.
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
Mon: complains about doing X, a task that requires a half-dozen keystrokes, repeatedly; I show him how to record a macro

Tue: now wants to do both X and Y; I show him how to name a macro so he can have more than 1

Thu: I discover he starts each day by typing in half-a-dozen macros; I show him how to save a macro across sessions

Fri: he is using macros that are nearly a hundred keystrokes long

Mon: he has been hacking all weekend, and now has a dozen macros, including macros that write macros
Sarah Connell has a doctorate in English Literature and was one of our student encoders. She is now the manager of the student encoders. She wanted to know what elements in the works of Margaret Cavendish were not in English. So she wrote this XPath.

It’s not perfect. Arguably the first “/” should be just “/”, as the one and only <text> element in a TEI file is a child of <TEI>. And you might think she would get false positives from cases where @xml:lang = "en", but there are no such cases in Cavendish, and she knows that.
Teaching: stage 2 and on

- Audio: click above to play ~/Music/03 T-E-I.ogg Of ~/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

[ The song was not played during the presentation, but rather later during Balisage Bard. ]
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

[ On panicking: I try to imbue the students with an attitude that considers ill-formedness a sin against all humanity, but invalidity merely an annoyance. When my kids were little I taught them about the CO and fire alarms. “If this alarm {BEEP} goes off, it means there is too much CO. Get up, get dressed, make sure your sister is up, kick the dogs out, and grab the phone on your way out to the meeting point to call 9-1-1. Elapsed time 2–3 minutes. If this alarm {BUZZ} goes off, it means there is a fire. Get up, get out. Do not get your shoes, do not worry about your sister, do not get the phone. If the dogs are still in the house, they die. You’re only responsibility is to get yourself out to the meeting point rapidly. Elapsed time: 0.2–0.3 minutes (12–18 s).

So it is with WF vs validity. If it’s invalid and you have the time to fix it, great. But don’t even think about checking-in that ill-formed file! ]
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

[Presentations like this are often filled with pithy quotations from smart people that support one’s position or argument. Charles Petzold is a smart guy who has written several books that I understand are quite good. This quotation is from the preface to CODE: The Hidden Language of Computer Hardware and Software, in which he is describing his approach to explaining computing in said book. I am told it is an excellent book, and while I believe that, I could hardly disagree with this quotation more. Metaphors and similes and analogies, especially when your students analyze such literary devices for a living, are the bedrock of good teaching. ]
[ Apparently having a oXygen screenshot is obligatory, so here’s mine. 2 points to anyone who recognizes the episode. ]
The size and shape of the room matters. In general, a seminar room with a rectangular or oval (or round, I suppose) table that fosters discussion is better than people sitting in rows. People sitting in rows *behind computer monitors* so I can’t see their faces is the worst.

This image is related to the previous one — it is from the 1981 film *Excalibur*, and Patrick Stewart is one of those knights.

The background image of these slides is of the sun rising *directly behind* the U.S. Capitol building, taken Fri 2012-03-23 at 06:19:27 ET from right near the Washington Monument with a Nikon D5100 using a 55–200 mm lens set at 200 mm, 1/500 @ f/5.6. It was then cropped and processed in `gimp` using the `Levels…` dialog (from the `Colors` menu), but I don’t remember the details.