Intellectual life in the age of Google

Gregory Crane
Winnick Family Chair in Technology and Entrepreneurship
Thanks

- Elli Mylonas
- Steve (Nick) Derose
- David Durand
- Allen Renear
  - “What is text, really?”, ACM DL 1992
Welcome to the New World

• The digital world is a separate space
• There are no indigenous peoples
  – We are the first settlers
• What would you like to see evolve?
  – Canada?
  – Or an alternative south of us?
How are doing?

- Where have the finest minds in the humanities brought us?
- What could we accomplish if we were able to inspire the leaders in our fields?
What’s wrong with this picture?
What’s wrong with this picture?
Democratic Valuations

• What do budgets say?
• What do we think?
• What do we want to change?
• What kind of world do we want?
• What are we doing about it?
• The world is changing.
Beware the “coiled fish”

• Separate tendentious anecdotes from whatever real values they advance
  – If thousands read and no one corrects, how important are those errors?

• Should I walk home to Boston because planes have crashed?
  – What are the risks of walking?
Wagging rights?

• Does the print dog wag a digital tail?
  – Where in society is that still true?
  – What happens when the Biodiversity Heritage Lib etc. pull the last scientists out of print?

• Phase shift as the dog turns digital
Digital Priorities?

• How many projects assume their users have access to a huge research library?
• Should we build add-ons for specialized research or a new space for the intellectual life of society?
  – Academics live simultaneously in the old and new worlds
• Do we want to change the world? How?
Age of Google

• Massive, global audiences
  – cross culture and language
  – customize and personalize

• Massive collections
  – Many producers, languages, cultures
  – Subsumes all traditional media
  – Supports new intellectual production: Wikip
Age of Google

• Mass expansion of access
  – You can get an answer to that question…
• Best minds solving simple problems
  – Search, translation, doc. understanding
• Power beyond/revolt against the elite*
  – Wikipedia and the “Moron from California”

* Elite = Everyone at the TEI conference
Age of Google

• Changing measures:
  – Collections acquire structure as
    • They get bigger (more stuff)
    • They get smarter (better DM/ML)
  – Statistical processes *minimize* errors
  – Good and flexible beats better and static
    • Decentralize corrections (Distributed Proofer)
    • Let the audience decide if it cares
    • If no one fixes N errors per page....?
Age of Google

- Machine generated texts = initial state
  - Flexibility of the system
    - How easily can you fix things?
  - Energy of the environment
    - Does anyone care enough to notice *and* fix?

- Publication as evaluation of your field
  - If texts don’t rapidly improve -->
    - What do we conclude about the seriousness of the field?
Million Book Libraries

Or, picking the bones of the academic world.
Major Efforts

• Google and Microsoft: open access
  – Semi-open source
  – 1-2 million books already processed (Google)
• Open Content Alliance: open source
  – Yahoo, Microsoft(?), also Sloan
  – Internet Archive as clearing house
  – Scanning centers in Toronto, California. Others?
    • 1,000 books/machine/week
• Alouette -- Canadian effort/ EU i2010?
• Other third party scanning: c. $.10/page
Mellon Million Books Study

• Call for Papers: December 15, 2006
  – Millionbooks@perseus.tufts.edu
  – http://geryon.perseus.tufts.edu/data/millionbooks/callforpapers.htm

• Workshop at Tufts University:
  – Medford, MA: May 22-24, 2007
Core Technologies

• Core processes
  – Page image to text
  – One language to another
  – Text to data
  – Customization/Personalization
“Million Book” Libraries

- Magnitude - OM $10^3$ bigger
- Structure - very little markup
- Content - unbounded
- Errors - no manual checking
- Access - 24/7 to any point on earth
- Languages -- 500 languages in Widener
- Audience -- global and massive

8 features changes by > 1 order of magnitude
Outcome space

- High end: aggregate 10.5+ million library
  - Best lib in history of human race
  - Universally accessible
  - Mass market vs. prestige elite model

- Low end: lots of public domain books
  - Systematic collection building
    - Sabin American Collection
    - Extensive public domain editions
    - Machine actionable reference works
Artisanal collections?

- EEBO/Text Creation Partnership
  - 20,000 books ~ 50 OCA scanners/1 month
  - New images from source (not Microfilm)
  - *Can be* unencumbered

- What is the role of miniscule collections with a few *thousand* curated texts?

- How scalable is markup?
Oh my God!

Did we all waste the last 20 years?
Is TEI to Google Books as Intermedia/Xerox Notecards was to the World Wide Web?
One Collection’s response
Redefine institutional role

• How do we benefit from behemoth collections with increasingly sophisticated services?
  – Don’t compete where we will lose
  – Identify our strengths
Classics as example

• Open up existing classical materials
  – Use markup to enhance value
  – Encourage use by Google/MS etc.

• Expand existing materials
  – Create Open Content Infrastructure
    • Image books evolving to XML
    • Exploit automation and decentralization
    • Rethink roles and workflows
1) Open Access/Open Source World
We get religion: power to the people

• Rights regimes
  – vocal and emphatic demand for unencumbered XML
    • Better for the field to let texts circulate
    • seems universal among classicists under 50 who care enough to comment

• Perseus Source Materials
  – March 2006 7.5m words of Greek and Latin, 55m words of American TEI Compliant texts
  – Released under CC license
QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
Sustainability?

- Existing work enters Institut. Repository
  - Gradually Fedora/Google etc. add services
  - Projects/collections too small for infrastruct.
- If new work can’t justify funding …
  - How valuable is it?
- If old work can be freely extended
  - What labor will we attract?
- Hypothesis: increasing value of previous work increases support for future work
Capital Formation

• Text Creation Partnership model
  – Limited embargo *pour encourager les autres*
  – Data reverts to the public domain in a finite period of time (e.g., 5 years).
QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
Ι. Δίδυμος ὁ γραμματικός ἐν τῇ περὶ τῶν ἁζώνων τῶν Σόλωνος ἀντιγραφῇ πρὸς Ἀσκληπιάδην Φιλόκλεος πτῶς τέθεικε λέξιν, ἐν ὧν τῶν Σόλωνα πατρὸς Ἑυφρότιον ἀποφαίνει παρὰ τὴν τῶν ἄλλων δόξαν, ὡσα μέμνησται Σόλωνος. Εἰς τετελεῖσθαι γὰρ αὐτῶν ὀπάντες ὀμαλῶς γεγο-νέα λέγουσιν, ἀνδρὸς οὐδὲς μὲν, ὡς φασί, καὶ ὑπάνως μέσῳ τῶν πολιτών, οἷος δὲ πρῶτος κατὰ γένος ἦν ὁ γὰρ Κορινθίας ἀνέκαθεν. τὴν δὲ μητέρα τοῦ Σόλωνος Ἡρακλείδης ὁ Ποσειδίκος ἱστορεῖ τῆς Πεισίστρατος μητρὸς ἀνέργειας γε-νέσθαι, καὶ φιλία τὸ πρῶτον ἦν αὐτοῦ πολλὴ μὲν διὰ τὴν ανυγχέντα, πολλὴ δὲ διὰ τὴν ἐυφυήν καὶ ἄραν, ὡς ἐνώτικον, ἐρωτικὸν τῶν Πεισίστρατος ἀπαλληλούμενον τοῦ Σόλωνος. θεῖον ύπερτέρων, ὡς ἐρυθείς, εἰς διαφοράν αὐτῶν ἐν τῇ πολιτείᾳ καταστάσεως οὐδὲν ἤγευσεν ἡ ἐχθρία σκληρόν ὑπὸ ἄργου τάδιος, ἀλλὰ παρέμενεν ἐκεῖνα τὰ δίκαια ταῖς ὕψωσις, καὶ παρεδίδατος.

Τυφάμενα Δίων πυρός ἐτί ζοῦσαν φλόγα,

3 τὴν ἐρωτικὴν μνήμην καὶ χάριν. ὅτι δὲ πρὸς τοὺς καλοὺς οὓς ἦν ἔχωρον ὁ Σόλων οὐδὲ ἔρωτι παρα-βαλέος ἀηταναστήτην πῦκτης ὅπως ἐς χείρας.1

1 Ἐρωτὶ μὲν ἵνα δυτικὰ ἀντανακλησίναι πῦκτῆς ὅπως ἐς χείρας, ἐς καλὰς φυτεῖς.

(Sophocles, Trachiniae, 441 f.)
Aggregate Error Correction Rate

• Some perspective:

  – Data entry service companies often guarantee accuracy rates on blind double-keyed input of 99.95%:

    “Your project is handled by our experienced data entry staff to deliver 99.95% accuracy, using a variety of quality techniques…” – CPI Data Services (http://www.compupacific.com/services/data-entry.shtml)

  – Our character-level accuracy rate on Greek of 99.93%, which relies on primarily automated correction techniques, compares favorably to this industry standard.
Revolution!

- We can index vast amounts of Greek and redefine the role that Greek plays in the digital world
- Classic example of how communities in the long tail can radically enhance the value of collections, piece by piece
Google Usage Guidelines

• “Refrain from automated querying Do not send automated queries of any sort to Google’s system: If you are conducting research on machine translation, optical character recognition or other areas where access to a large amount of text is helpful, please contact us. We encourage the use of public domain materials for these purposes and may be able to help”
Rights Regimes

• Google and Microsoft: Semi-open
  – Have not yet developed a collaborative rel.
  – Obvious logic to do so in public domain
    • Each digitizes half and all is open source

• Open Content Alliance (Intern. Archive)
  – Library centric, massive dig strategy
  – Possible source for purposeful collections
    • Space in which to share and to develop
2) TEI as output format
TEI as output format

How many books could we tag by hand?
Human life span ~ c. 33,000 days
  Working career ~ c. 10,000 days
1,000,000 books @ 1 book a day
  --> 100 careers, 4,000 years of labor
TEI as output format

TEI provides the best target tagset in which to encode many categories of data
Auto-tag unstructured text
  e.g., named entity analysis -->
    up to 1 billion words a day
TEI as output format

- The DL reads its gazetteers/encycl’s
  - Enumerates named entities
  - Analyzes heuristics associated with each
    - Coarse “Bag of words” heuristics
    - Information heuristics (e.g., death date, population)
- Assumes that DL can recognize start and end of entries….
TEI as output format

About the same time <persName n="Alcibiades, , , , , " id="n-0001.0000.00000.01912"
reg="mostcommon:Alcibiades, nomatch:0"> <surname>Alcibiades</surname> </persName>
returned with his <num value="13">thirteen</num> ships from <placeName key="perseus,Caunos" >Caunos</placeName> and <placeName reg="Phaselis, Antalya Ili, Akdeniz kiyisi" key="tgn,7002612">Phaselis</placeName> to <placeName key="tgn,7002673">Samos</placeName>, bringing word that he had prevented the
TEI as output format

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
TEI as output format

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
3) TEI as training data
TEI as training data

• TEI markup and emerging ontologies allow automated systems to mine texts
  – Use knowledge about Victorias in 1873 to analyze a document from 1879
  – Use one good edition to align N others
TEI as training data

• Open Content Scholarly Editions
  – “Elevator approach”
    • Start with image books and add structure
      – How much structure is enough?
      – Who cares and how much?
TEI as training data

Part of a page image of the Roman poet Statius
TEI as training data

- Compare a TEI text of Statius with the OCR output
  - Identify words that show up 1X in both the TEI text and the OCR output
  - Find those dual-unique words that define a sequence
    - “A B C D E” in Edition 1
    - “C A B D E” in Edition 2
    - Use “A B D E” as anchors to align the two texts
TEI as training data

Note the words in bold -- these are points where we can align OCR output to a TEI text.

This works even when comparing original/modernized spelling.
TEI as training data

• What can you do?
  – Help filter out non-textual elements (e.g., textual notes)
  – Import structural metadata from TEI text to the OCR text, e.g., add line numbers!
  – Perform scalable error correction:
    • If a morphological analyzer recognizes word N in text 1 but not in text 2, then word N in text 1 is probably an error
TEI as training data

Automatic corrections in Livy 21.1:

bellum maxime omnium memorabile, quae umquam gesta sint, me scripturum, quod Hannibale duce <corr sic="Uarthagmienses">Carthaginienses</corr> cum populo <corr sic="nomano">Romano</corr> gessere.
TEI as training data

- How far can these techniques go?
  - Pretty far -- Bioinformatics has developed *excellent* tools for string matching

- Are these techniques perfect?
  - No -- there is always an error rate.

- Do they let us do things we can to do?
  - Depends on who wants to do what
Some possible questions?

How do I get my Phd?
How do I get an academic job?
How do I get tenure?
How do I become full professor?
How do I publish and look clever?
Why aren’t professors from super-elite institutions at the TEI meeting?
Better questions for the TEI

• How do we support an intellectual endeavor, e.g., increasing the role of Anglo-Saxon or classical Greek?
• How do I contribute to the intellectual life of humanity as a whole?
Conclusions

• High level automation may challenge the TEI more significantly than the indifference of the elite
• The TEI can and should be part of any positive change
Thank you!