

# Intellectual life in the age of Google

Gregory Crane

Winnick Family Chair in  
Technology and Entrepreneurship

# Thanks

- Elli Mylonas
- Steve (Nick) Derosé
- David Durand
- Allen Renear
  - “What is text, really?”, ACM DL 1992

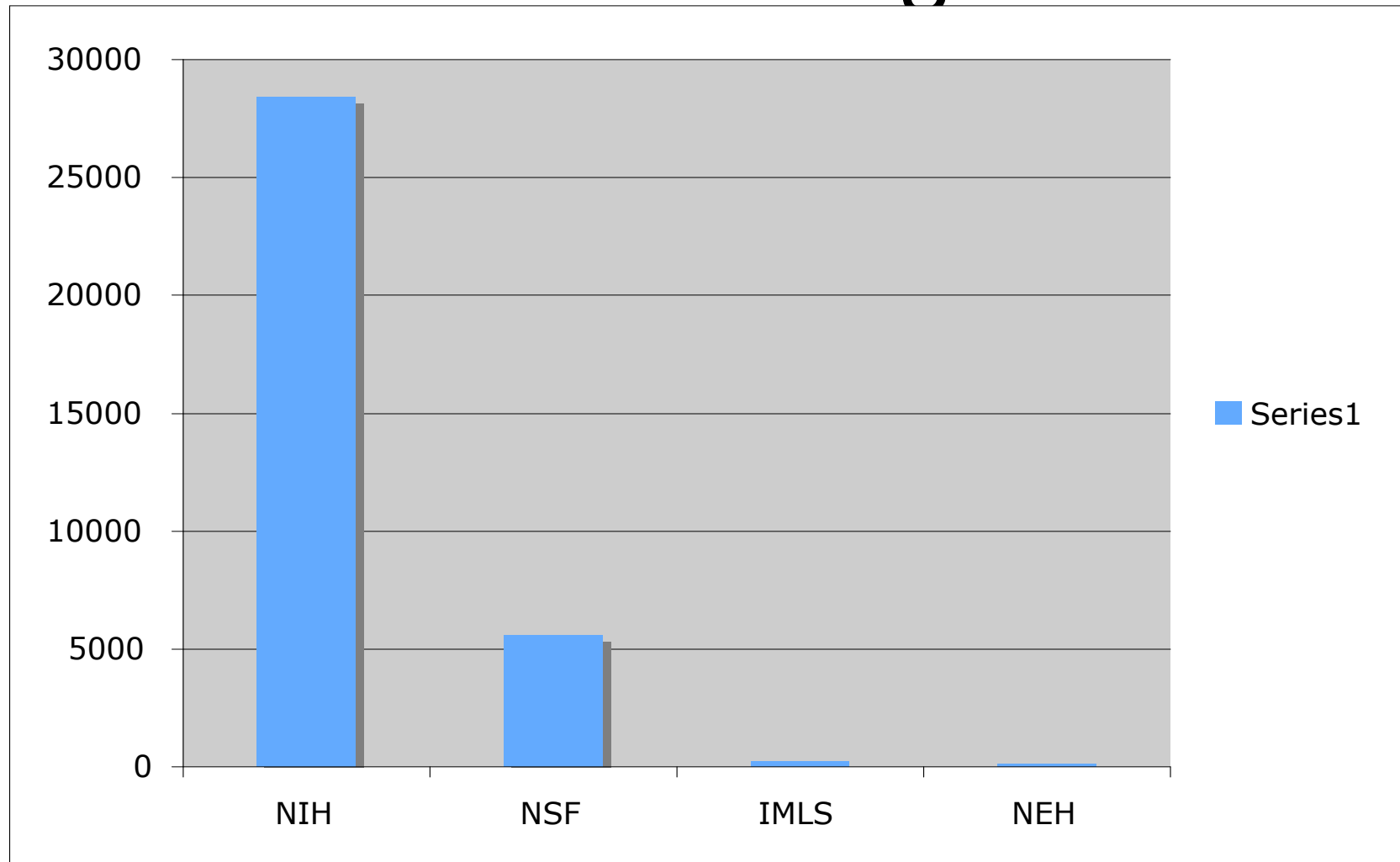
# Welcome to the New World

- 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?

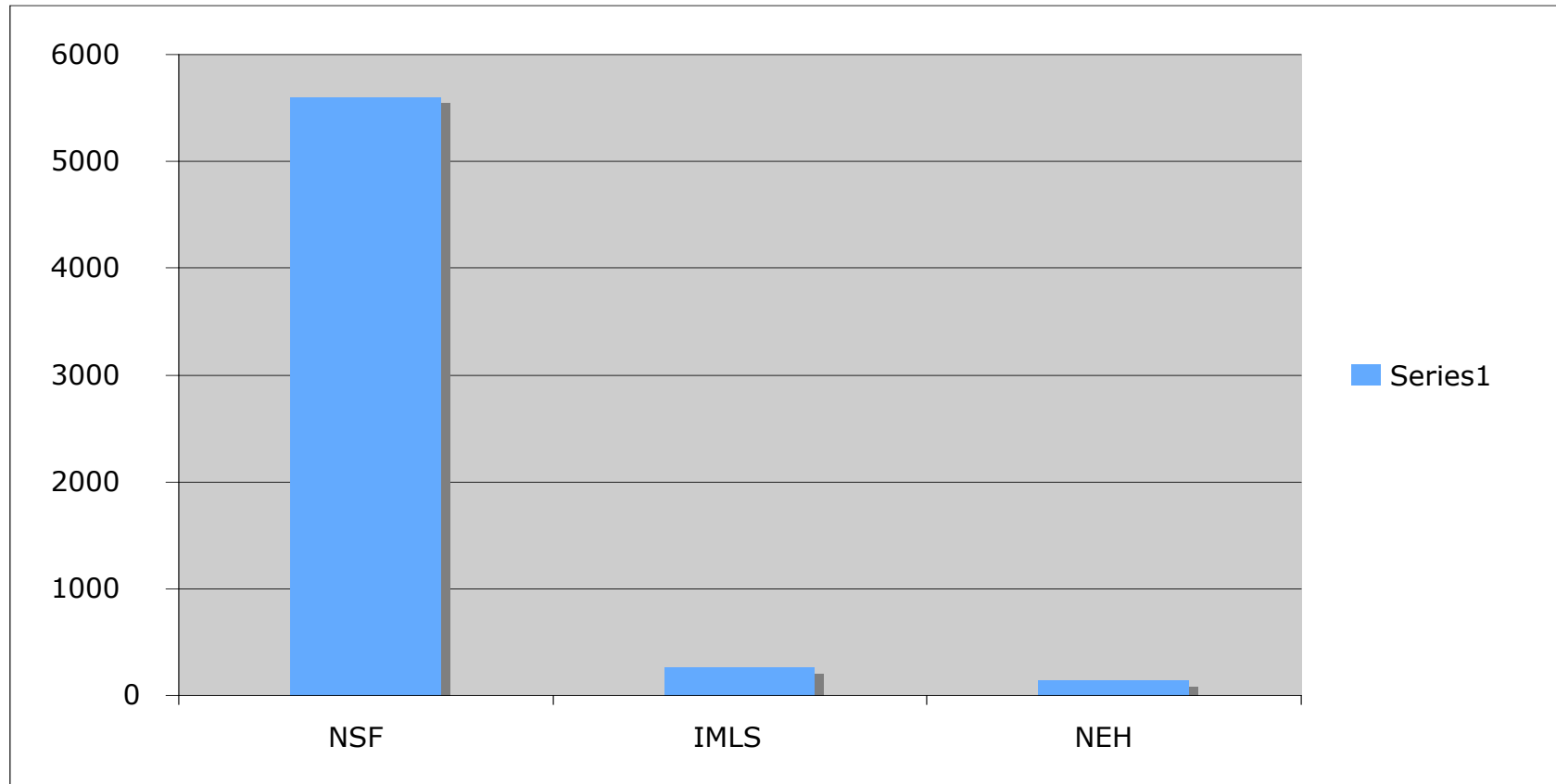
- Where have the finest minds in the humanities brought us?
- What could we accomplish if we were able to inspire the leaders in our fields?

# Some US Budgets



What's wrong with this picture?

# Some US Budgets



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 Proofr)
    - 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](mailto: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).

# Google Books

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



# OCR and Pattern-recognition

- First page of *Life of Solon*
- Scanned image left, recognized text right

## ΣΟΛΩΝ

<sup>1</sup> Ἰ. Δίδυμος ὁ γραμματικὸς ἐν τῇ περὶ τῶν ἀξόνων τῶν Σόλωνος ἀντιγραφῇ πρὸς Ἀσκληπιάδην Φιλοκλέους τινὸς τέθεικε λέξιν, ἐν ἣ τὸν Σόλωνα πατρὸς Εὐφορίωνος ἀποφαίνει παρὰ τὴν τῶν ἄλλων δόξαν, ὅσοι μέμνηται Σόλωνος. Ἐξηκестίδου γὰρ αὐτὸν ἅπαντες ὁμαλῶς γεγενῆναι λέγουσιν, ἀνδρὸς οὐσία μὲν, ὡς φασι, καὶ δυνάμει μέσου τῶν πολιτῶν, οἰκίας δὲ πρώτης <sup>2</sup> κατὰ γένος· ἦν γὰρ Κοδριίδης ἀνέκαθεν. τὴν δὲ μητέρα τοῦ Σόλωνος Ἡρακλείδης ὁ Ποντικὸς ἰστορεῖ τῆς Πεισιστράτου μητρὸς ἀνεψιᾶν γενέσθαι. καὶ φιλία τὸ πρῶτον ἦν αὐτοῖς πολλή μὲν διὰ τὴν συγγένειαν, πολλή δὲ διὰ τὴν εὐφυΐαν καὶ ὦραν, ὡς ἐνιοὶ φασιν, ἐρωτικῶς τὸν Πεισίστρατον ἀσπαζομένου τοῦ Σόλωνος. ὅθεν ὕστερον, ὡς ἔοικεν, εἰς διαφορὰν αὐτῶν ἐν τῇ πολιτείᾳ καταστάτων οὐδὲν ἠνεγκεν ἢ ἐχθρα σκληρὸν οὐδ' ἄγριον πάθος, ἀλλὰ παρέμεινεν ἐκεῖνα τὰ δίκαια ταῖς ψυχαῖς, καὶ παρεφύλαξε,   
Τυφόμενα Δίου πυρὸς ἔτι ζῶσαν φλόγα,   
<sup>3</sup> τὴν ἐρωτικὴν μνήμην καὶ χάριν. ὅτι δὲ πρὸς τοὺς καλοὺς οὐκ ἦν ἐχυρὸς ὁ Σόλων οὐδ' Ἐρωτι θαρραλέος “ ἀνταναστήναι πύκτης ὅπως ἐς χεῖρας,”<sup>1</sup>

<sup>1</sup> Ἐρωτι μὲν νυν ὅστις ἀντανίσταται πύκτης ὅπως ἐς χεῖρας, οὐ καλῶς φρονεῖ. (Sophocles, *Trachiniae*, 441 f.)

1. Δίδυμος ὁ γραμματικὸς ἐν τῇ περὶ τῶν ἀξόνων τῶν Σόλωνος ἀντιγραφῇ πρὸς Ἀσκληπιάδην Φιλοκλέους τινὸς τέθεικε λέξιν, ἐν ἣ τὸν Σόλωνα πατρὸς Εὐφορίωνος ἀποφαίνει παρὰ τὴν τῶν ἄλλων δόξαν, ὅσοι μέμνηται Σόλωνος. Ἐξηκестίδου γὰρ αὐτὸν ἅπαντες ὁμαλῶς γεγενῆναι λέγουσιν, ἀνδρὸς οὐσία μὲν, ὡς φασι, καὶ δυνάμει μέσου τῶν πολιτῶν, οἰκίας δὲ πρώτης κατὰ γένος· ἦν γὰρ Κοδριίδης ἀνέκαθεν. τὴν δὲ μητέρα τοῦ Σόλωνος Ἡρακλείδης ὁ Ποντικὸς ἰστορεῖ τῆς Πεισιστράτου μητρὸς ἀνεψιᾶν γενέσθαι. καὶ φιλία τὸ πρῶτον ἦν αὐτοῖς πολλή μὲν διὰ τὴν συγγένειαν, πολλή δὲ διὰ τὴν εὐφυΐαν καὶ ὦραν, ὡς ἐνιοὶ φασιν, ἐρωτικῶς τὸν Πεισίστρατον ἀσπαζομένου τοῦ Σόλωνος. ὅθεν ὕστερον, ὡς ἔοικεν, εἰς διαφορὰν αὐτῶν ἐν τῇ πολιτείᾳ καταστάτων οὐδὲν ἠνεγκεν ἢ ἐχθρα σκληρὸν οὐδ' ἄγριον πάθος, ἀλλὰ παρέμεινεν ἐκεῖνα τὰ δίκαια ταῖς ψυχαῖς, καὶ παρεφύλαξε,

Τυφόμενα Δίου πυρὸς ἐπιζῶσαν φλόγα,

τὴν ἐρωτικὴν μνήμην καὶ χάριν. ὅτι δὲ πρὸς τοὺς καλοὺς οὐκ ἦν ἐχυρὸς ὁ Σόλων οὐδ' Ἐρωτι θαρραλέος ἀνταναστήναι πύκτης ὅπως ἐς χεῖρας

# 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

```
9 <milestone unit="sentence" n="974"/></seg> </p>
0 <p> <milestone n="108" unit="chapter"/> <milestone unit="section" n="1"/>
0 <seg> <milestone unit="para" ed="P"/>About the same time <persName
0 n="Alcibiades,,,,," id="n-0001.0000.00000.01912"
0 reg="mostcommon:Alcibiades,nomatch:0"><surname>Alcibiades</surname></persName
0 > returned with his <num value="13">thirteen</num> ships from <placeName
0 key="perseus,Caunus">Caunus</placeName> and <placeName reg="Phaselis, Antalya
0 Ili, Akdeniz kiyisi" key="tgn,7002612">Phaselis</placeName> to <placeName
0 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

bisque iugo Rhenum, bis adactum legibus Histrum  
et coniurato delectos vertice Dacos 20  
aut defensa prius vix pubescentibus annis  
bella Iovis, tuque o, Latiae decus addite famae,

*1-389 desunt in K: totus lib. i et lib. ii vv. 1-66 desunt in N (sup-  
plevit manus saeculi xv) 11 cognota P 12 arcum D Cod.  
Coll. Magd. Oxon. 16 limen Bachrens 18 spirare Heinsius,  
Bentley 21 prius sup. lin. Q 22 teque (u suprascr.) o P:  
tuque o w: tuque ut Lachmann*

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

atque adeo iam nunc gemitus et prospera Cadmi 15  
**praeteriisse** sinam: limes mihi **carminis** esto  
Oedipodae confusa domus, quando Itala nondum  
signa nec **Arctos** ausim sperare triumphos  
bisque iugo Rhenum, bis adactum legibus Histrum  
et **coniurato deiectos** vertice **Dacos** 20  
aut defensa prius vix pubescentibus annis  
bella Iovis, tuque o, **Latiae** decus **addite** famae,

*1-389 desunt in K: totus lib. i et lib. ii vv. 1-66 desunt in N (supplevit manus saecti xv) Ir cognota P I2 arcuin D Cod. Coll. Magd. Oxon. I6 limen Baehrens I8 spirare Heinsits. Bentley 21 prius sup. lin. Q 22 teque (u suprascr.) o P: tuque o c: tuque ut Lachmann*

pagebreak=p.10002  
P. PAPINI STATI

quem nova mature subeuntem exorsa parentis  
aeternum sibi **Roma** cupit (licet **artior** omnis  
limes agat **stellas** et te plaga lucida caeli, 25  
Pleiadum **Boreaeque** et **hiulci** fulminis expers,

# 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</co  
rr> 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't 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!