Scholarly Text Curation & Robust Anchoring Requirements

Timothy W. Cole (t-cole3@illinois.edu)
Thomas G. Habing (thabing@illinois.edu)
Anchoring Methods to Support Curatorial Annotation of Scholarly Text Resources

- Should be fine-grained – for text this means individual words and phrases
- Should ensure persistence, e.g., even as adjacent content is updated / corrected
- Can be aligned across derivative formats and serializations, even across repository boundaries
- Can support search & replace, e.g., the target is set of all instances found in a specific context
- Should help distinguish curatorial annotations of a specific digitization & its derivatives from annotations of intellectual substance
Use cases

• Correction of OCR & manual transcriptions
  – HathiTrust DL: 11 million digitized volumes automated OCR
  – Text Creation Partnership (TCP): 50,000 manually transcribed
  – Support corrections by curators, outside experts, the crowd

• Correction of automated annotation, e.g., part-of-speech tagging of TEI

• Distinct from proposed targeting schemes for other kinds of scholarly use cases, e.g., commentary
Why Annotations?

• So proposed corrections can be reviewed and themselves annotated as needed

• To share with other repositories

• To maintain portability of provenance
Crowdsourcing Features

Unique to Veridian is a set of crowdsourcing features that allow registered users to correct OCR text errors, add tags and comments to any type of content, and transcribe material that cannot be converted to text such as handwriting.

Benefits

- Patron engagement
- Improved text quality
- Better search results
- Creates an online community

Engaging with users and building virtual communities is just as important to the users as providing the data itself. They want to be part of a community.

Rose Holley - National Library of Australia

User Text Correction

OCR software is hardly a perfect way to extract text from the image of a digitized document and paying for manual correction is costly and time consuming.

Veridian’s User Text Correction tool enables users to correct OCR errors as they come across them, helping to improve the collection without the expense.
ON THIS DAY

The Broad Ax 26 March 1910

ABOUT THIS COLLECTION

Welcome to the Illinois Digital Newspaper Collections.
This collection contains 45 newspaper titles, 90,104 issues comprising 1,116,753 pages and 6,646,636 articles.

Since 2005, the History, Philosophy, and Newspaper Library at the University of Illinois has been developing unique digital newspaper content. The works collected here include digital facsimiles of newspapers and trade journals in a variety of fields. Users may search, browse, tag, and correct OCR text to improve searchability and access.

ACKNOWLEDGEMENTS

The Illinois Digital Newspaper Collections is supported in part by the donors and grant agencies listed here.

The University of Illinois at Urbana-Champaign newspaper library holds millions of pages in print and microfilm, but only a small fraction of the collection has been digitized. A $5, $15, or $25 contribution will help us fund the cost of digitizing more newspapers for free online access.

Please click the Donate button and specify that your donation is supporting IDNC, which is part of the History, Philosophy and Newspaper Library.

donate
CELEBRATE AS THEY NEVER HAVE BEFORE

United Press Bulletin

Washington, Nov. 11—The armistice, signed at 5 o'clock this morning, French time, hostilities will stop 11 o'clock this morning.

Message was given out by the state department in Paris at 2:46 a.m. The terms of the armistice were read, but will be made public after the state departments are opened at 9:30 this morning.

President Wilson this afternoon and the world terms Germany accepted. Signed the armistice.

The armistice pictured Germany's surrendering abjectly. Foch on the field, her armies beaten, her government her master in flight. All congress and a small crowd heard the presiding words, but enthusiasm ran riot.

Terms followed:

Any's surrender terms include cessation of hostilities in invaded territory, including Alsace, and Luxembourg.

Termination of vast amounts of guns and equipments, destroyment of left bank of Rhine.

Termination of vast amounts of rolling stocks in occupancy. Abandonment of Bucharest and Brest.

Urbana and Champaign delegations thru their respective towns and meeting at the university, were only one feature of the day. The proposition was entirely too big to handle there.

Mayor Richards president, introducing Dean Kinley, who, in turn, introduced Representative W. H. M. Miller. All made brief but stirring talks.

The people of the Twin Cities were awakened at 2 o'clock this morning by the ringing of bells, accompanying the firing of guns. The fire department was the first to proclaim the tidings in Urbana. The armistice was signed at 5 a.m., Paris time, and

The jubilee in the Twin Cities began with the press funeral at 3 o'clock this morning that Germany had signed the terms of the armistice ending the war, and continued throughout the day.

A great celebration at the University of Illinois auditorium at 10 o'clock this morning, following parades of Urbana and Champaign delegations thru their respective towns and meeting at the university, was only one feature of the day. The proposition was entirely too big to handle there.

Mayor Richards president, introducing Dean Kinley, who, in turn, introduced Representative W. H. M. Miller. All made brief but stirring talks.

The people of the Twin Cities were awakened at 2 o'clock this morning by the ringing of bells, accompanying the firing of guns. The fire department was the first to proclaim the tidings in Urbana. The armistice was signed at 5 a.m., Paris time, and...
Align across representations

• Treat the OCR as an annotation of a segment of PDF / JPEG page image
  – Annotating agent is OCR program

• Proposed OCR correction is then an annotation of the OCR annotation of the page image

• Complicating factor – OCR outputs at page level, correction is usually done at line level
Annotating repeated errors?

The string “Jrbana” appears 782 times in OCR texts of the *Urbana Daily Courier* (1903-1935)

Do we need to require that users find every instance of “Jrbana”....?

Can we have search-and-replace annotations?

* “Urbana” appears ~ 200,000 times
http://annolex.at.northwestern.edu/
| Text: Heywood. T.: The Rape of Lucrece |
| Spelling: Colatine |
| Lemma: |
| POS: |
| Citation: |
| Match: starting with |
| Combine: and |
| Sort: Text Order |
| Filter: All |

**Spelling in Context**

<table>
<thead>
<tr>
<th>Spelling</th>
<th>Lemma</th>
<th>POS</th>
<th>Edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colatine</td>
<td>Colatine</td>
<td>np1</td>
<td>Edit</td>
</tr>
<tr>
<td>Exeunt : manent Colatine and Valerius</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No doubt Colatine no doubt hewes a giddie world , it Reele</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>your preperation , but one thing more , goes Colatine along ?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the incourvengence Of my wrongd kinsman Colatine , the Treason AGainst diunst Lucrece</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>My Husband Colatine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>body Denide all funerall rites , and louing Colatine Shall hate thee euen in death : thei saue</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>them Orators , To pleada the cause of absent Colatine , your friend and kinsman .</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>displeasure : but soft , some of my Lord Colatines menlyve in the next chamber , I care not</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brutus , Valerius Horatius . Amus , Scæula , Colatine .</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did nobody see my Lord Colatine ? oh , my Lady commends her to you , her's</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosen Colatine the news at Rome ?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that with such sad presage , Disturbed Colatine , and noble Brutus Are hurryd from the</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horatius , Valerius , Scæula , Lucretius , Colatine .</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correction</td>
<td>Corrector</td>
<td>Approver</td>
<td>Applier</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>Update A04206-b-b-1460: To my b (my, z) my (my, pol1) Rebecca.</td>
<td>martin</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Delete A11909-03-44-a-0070: THE THYRD TRAGEDY OF L. ANN AEVS (AEVS, np1) Seneca: entituled Thebais, translated</td>
<td>martin</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Update A11909-03-44-b-0060: THE THYRD TRAGEDY OF L. ANN (and, ccc) ANNAEVS (Annaeus, np1) AEVS Seneca: entituled Thebais, translated</td>
<td>martin</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Delete A14875-44-b-2140: that belong to Great men remember th’ old widoes (widoes, np2) wives (wife, n2) tradition, to be lie the Lyons lth Tower</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Update A14875-44-b-2140: that belong to Great men remember th’ old widoes (widoes, np2) wives (wife, n2) tradition, to be lie the Lyons lth Tower</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Delete A14875-44-a-0450: Thou hast to good a face to be a hang-man, f (f) If thou be doe thy office in right forme</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Update A14875-43-b-2160: Sirah (sirah, n1) Sirah (sirah, n1) you once did strike mee, Ile strike you</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Update A14875-43-b-1080: , take this president : Man may his Fate foresee (foresee, vvi) foresee (foresee, vvi), but not prevent , And of all Axiomes this</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Delete A14875-43-a-4060: A matachine it By your drawne swords , (seemes, seem, vvi) Chuch-men turn’d reuellers.</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Delete A14875-43-a-4050: A matachine it By your drawne swords , (seemes, seem, vvi) Chuch-men turn’d reuellers.</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Insert A14875-43-a-3990: A matachine it (it, np31) seemes (seem, np31) By your drawne swords , (seemes, Chuch-men</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Update A14875-43-a-3990: A matachine it (it, np31) seemes (seem, np31) By your drawne swords , (seemes, Chuch-men</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Update A14875-43-a-3990: A matachine it (it, np31) seemes (seem, np31) By your drawne swords , (seemes, Chuch-men</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Delete A14875-43-a-0120: You see the Fox comes many times short home (home, uh) home (home, uh) me , Tis here proud true.</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Update A14875-43-a-0110: You see the Fox comes many times short ho-(ho,-uh) home (home, uh) me , Tis here proud true.</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Update A14875-43-b-2310: They shoot and go to him ?</td>
<td>nayoon</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>
Title: The rape of Lucrece a true Roman tragedie. With the seuerall songs in their apt places, by Valerius, the merrie lord amongst the Roman peeres. Acted by her Majesties Servants at the Red-Bull, neere Clarken-well. Written by Thomas Heywood.

Author: Heywood, Thomas, d. 1641.
Collection: Early English Books Online

Table of contents | Add to bookbag
Exeunt: manent Colatine and Val•...•ius

Col.
Theres morall sure in this, Valorius,
Heeres model, yea, and matter too to breed
Strange meditations in the provident braines
Of our graue Fathers: some strange proiect lues
This day in Cradle thats but newly borne.

Vale.

No doubt Colat•...• no doubt he•...• es a giddie world, it Reeles, it hath got the staggers, the common-wealth is
sicke of an ague, of which nothing can cure her but some vi•...•ent and •...•ddaine affrightment

Cela.
The wife of Tarquin would be a Queene, nay on my life she is with childe till she be so.

Valer.

and longes to be brought to bed of a Kingdome, I deuine we shall see some •...•uffling to day in the Capitoll.

Cela.

If •...•here be any difference among the Princes and Senate, whose faction will Vale•...•ius follow?

Valer.

Oh Collain•...•! I am a true Citizen and in this I will best shew my selfe to be one, to take part with the stronger.
If Se•...•ius ore-come, I am Liegenman to Serius, & if Ta•...•quin subdue, I am for Vlue Tarquintus.

Col.

Val•...•ius, no more, this talke does but keepe vs from the sight of this solemnitie: by this the Princes are entring the Capitoll: come, we must attend.

Exeunt.
Valer.

Oh Collatin! I am a true Cittizen and in this I will best shew my selfe to be one, to take part with the stronger. If Seius orecome, I am Liegeman to Serutus, & if Taquin subdue, I am for Viue Tarquinius.
Relation between anchoring method & what’s being targeted

• 1st challenge – understanding the annotator’s intention.

• 2nd challenge – using a targeting approach that is consistent with annotator’s intent

• Some schemes limit possible range of interpretations
  – Chapter, verse & line approaches (e.g., CTS)
Welcome to De Commentariis, an open network for the crowd-sourcing of ancient text commentaries. Construct your own commentaries on classical texts, and view the commentaries others have made. Use it to get your students to critically engage with their learning texts.

Shakespeare, **Sonnet 35: 35**

**Edition of 1609**

( = um:cts:demo:shakespeare.sonnets.1609:35 )

1. No more be grieved at that which thou hast done:
2. Roses have thorns, and silver fountains mud:
3. Clouds and eclipses stain both moon and sun,
4. And loathsome canker lives in sweetest bud.
5. All men make faults, and even I in this,
6. Authorizing thy trespass with compare,
7. Myself corrupting, salving thy amiss,
8. Excusing thy sins more than thy sins are;
9. For to thy sensual fault I bring in sense,
10. Thy adverse party is thy advocate,
11. And 'gainst myself a lawful plea commence:
12. Such civil war is in my love and hate,
13. That I an accessory needs must be,
14. To that sweet thief which sourly robs from me.
Shakespeare, *Sonnet 35: 35.10*

Edition of 1609

( = urn:cts:demo:shakespeare.sonnets.1609:35.10 )

10 Thy adverse party is thy advocate,
Anchoring Methods to Support Curatorial Annotation of Scholarly Text Resources

• Should be fine-grained – for text this means individual words and phrases
• Should ensure persistence, e.g., even as adjacent content is updated / corrected
• Can be aligned across derivative formats and serializations, even across repository boundaries
• Can support search & replace, e.g., the target is set of all instances found in a specific context
• Should help distinguish curatorial annotations of a specific digitization & its derivatives from annotations of intellectual substance