

*W3C Workshop on Annotations
San Francisco, California USA
2 April 2014*

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



search



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

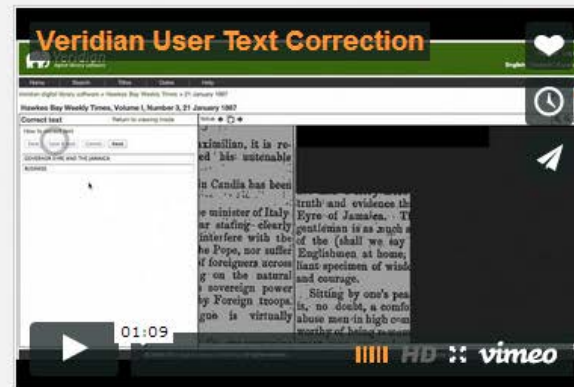
Engaging with users and building virtual communities is just as important to the users as providing the data itself. They want to be part of a community.

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.



Veridian Digital Library Software

The screenshot shows a web browser window displaying the Illinois Digital Newspaper Collections website. The browser's address bar shows the URL `idnc.library.illinois.edu`. The website header features the logo for Illinois Digital Newspaper Collections and navigation links for HOME, SEARCH, BROWSE, TITLES, COLLECTIONS, TAGS, FAQ, and HELP. A search bar is located in the top right corner of the header.

The main content area is divided into three columns:

- ON THIS DAY:** Features a thumbnail of a newspaper page from March 26, 1910, titled "The Broad Ax".
- ABOUT THIS COLLECTION:** Provides a welcome message and details about the collection's size and content. It also includes an acknowledgements section.
- SEARCH THE COLLECTION:** Includes a search input field and buttons for "Browse by title" and "Browse by date". It also lists "TOP TEXT CORRECTORS" with a table of names and counts.

At the bottom of the page, there is a prominent "donate" button.

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 6,648,636 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](#)

SEARCH THE COLLECTION



[Browse by title](#)



[Browse by date](#)

TOP TEXT CORRECTORS

1. Wes Keat	4,088
2. magsjoy	1,238
3. Erica P.	1,147
4. mlt	817
5. Joe Sciacca	358

[More information...](#)

To learn more about how to help correct newspaper text, click [here](#)

Correct text

How to correct text

Save Save & exit Cancel Next

CELEBRATE AS THEY NEVER
HAVE BEFORE)

jr _ _ j _ _ æpæe'Àjæe' . . ja Åæ fc _Æ li _Åæ :

Streets Thronged Within Few Min .
utes After Signing) of _Armistice Is
Proclaimedæ€Jollification Con-
tinues Thruout Day ,

ltb . iim _cilobrted _todiiy an _ncevi

_brfor and ah s > 1 i g probably never will

_icntn _ , unless il lh when _Sniuiy cunidB

marching _lioiiip .

The _jubileo In Hie Twin Ulios bo

Ran with Hie press _finsli received n

_i l o'clock this morning Hint Germanj

had signed the terms of tlio armstllce

ending the wai y and continued thru-

out the rest of iho day .

A great jollification at (lie Unlver

_sity or Illinois auditorium ill 10 o _clock

this , m on) l Ufa

_ following parades ot

(Jrbuna and Champaign _delcgatjof

thru their respective _to-w-ns and moot

Jng at the university , was only one

feature of the day . The _proposition

was entirely too big to handle thp ,

re

Mayor Richards presided , introducing

Doah Kinley who _ln turn _Intioduced

...ces in East Africa must be surrendered. Un-
nally—Must Recompense United States for
Merchantmen and Other Unarmed Vessels
Destroyed by U-Boats.

United Press Bulletin

Washington, Nov. 11—The armistice was signed at 5 o'clock this morning, French time. Hostilities will cease at 11 o'clock this morning.

A message was given out by the state department in Washington at 2:46 a. m. The terms of the armistice were announced, but will be made public after the state department offices are opened at 9:30 this morning.

Washington, Nov. 11.—President Wilson this afternoon addressed congress and the world terms Germany accepted and signed the armistice. The terms pictured Germany's surrendering abjectly to Foch on the field, her armies beaten, her government overturned and her master in flight. The entire congress and a small crowd heard the president's stirring words, but enthusiasm ran riot. The terms follow: Germany's surrender terms include cessation of hostilities in invaded territory, including Alsace and Luxembourg. Surrender of vast amounts of guns and equipments. Surrender of vast amounts of rolling stocks in occupation of left bank of Rhine. Surrender of vast amounts of rolling stocks in occupation. Abandonment of Bucharest and Brest treaties.

CELEBRATE AS THEY NEVER HAVE BEFORE.

Streets Thronged Within Few Minutes After Signing of Armistice Is Proclaimed—Jollification Continues Thruout Day.

Urbana celebrated today as never before and as she probably never will again, unless it is when Sammy comes marching home.

The jubilee in the Twin Cities began with the press flash received at 2 o'clock this morning that Germany had signed the terms of the armistice ending the war, and continued throughout the rest of the day.

A great jollification at the University of Illinois auditorium at 10 o'clock this morning, following parades of Urbana and Champaign delegations through 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 presided, introducing Dean Kinley, who, in turn, introduced Representative W. H. 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, screeching of whistles and 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

in the ad-
painter res-
street, Cha-
the one wh-
who insulte-
Mrs. Hall
home by S-
junk dealer
who came t-
ago from T-
man who he-
it accordin-
the latter is
jail to awa-
coroner's in-
at 10 o'loc-
clues to the
cepting the
his wife.
being the a-
the sheriff,
which is ec-
part, a Cl-
deeps with
Church stre-
Woman
A woman
hood, whose
being with-
ing from be-
"Death Co-
Walnut - st-
before 11 o'
According
man who fir-
ble Hackett,
was attract-
man she su-
proved to b-
"Keep away
want any tr-
other kept

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

Correct · Review · About · Log out tim

Text: Heywood, T.: The Rape of Lucrece

Spelling: Collatin•

Lemma:

POS:

Citation:

Match: starting with ▾

Combine: and ▾

Sort: Text Order ▾

Filter: All ▾

Search

Edit word 4-b-2890 from *The Rape of Lucrece*

View EEBO Image

New values

Spelling (Collatin•): Collatine

Lemma (collate): Collatine

POS (vvg): np1

Annotation:

Update ▾ Save

Spelling in Context	Spelling	Lemma	POS	Edit
Oh Collatin• ! I am a true Cittizen and in this I will	Collatin•	collate	vvg	Edit
matter where if frō the court , I'lle home to Collatin• , And to my daughter Lucrece ; home breeds	Collatin•	collate	vvg	Edit
Enter Collatin• .	Collatin•	collate	vvg	Edit
then weepe with our heads off , I nere tooke Collatin• for a polititian till now . Come Valerius	Collatin•	collate	vvg	Edit

Page 1 of 1

Heeres modell, yea, and matter too to breed
Strange meditations in the prouident braines
Of our graue Fathers: some strange proekt liues
This day in Cradle thats but newly borne.
Valer. No doubt *Collatine* no doubt heetes 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-
olent and suddaine affrightment.
Colla. The wife of *Tarquinius* 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 scuffling to day in the Capitoll.
Colla. If there be any difference among the Princes and
Senate, whose faction will *Valerius* follow?
Valer. Oh *Collatine* ! I am a true Cittizen, and in this I will
best shew my selfe to be one, to take part with the stronger.
If *Seruius* ore-come, I am Liegeman to *Seruius*, & if *Tarquinius*
subdue, I am for *Vine Tarquinius*.
Colla. *Valerius*, no more, this talke does but keepe vs from
the fight of this tolemnitie : by this the Princes are entering
the Capitoll: come, we must attend.
Exeunt.
Senate

Text: Heywood, T.: The Rape of Lucrece

Spelling:

Lemma:

POS:

Citation:

Match:

Combine:

Sort:

Filter:

Edit word

New values

Spelling ():

Lemma ():

POS ():

Annotation:

Spelling in Context	Spelling	Lemma	POS	Edit
Enter Sextus , Arnus , Lucretius , Val•rius , Colatine and Senators .	Colatine	Collatine	np1	<input type="button" value="Edit"/>
Exeunt : manent Colatine and Val•rius	Colatine	Collatine	np1	<input type="button" value="Edit"/>
No doubt Colatine no doubt hee•es a giddie world , it Reeles	Colatine	Colatine	np1	<input type="button" value="Edit"/>
your preperation , but one thing more , goes Colatine along ?	Colatine	Collatine	np1	<input type="button" value="Edit"/>
, the incurdveng•nce Of my wrongd kinsman Colatine , the Treason A Gainst diuin'st Lucrece	Colatine	Collatine	np1	<input type="button" value="Edit"/>
My Husband Colatine	Colatine	Collatine	np1	<input type="button" value="Edit"/>
body Denide all funerall rites , and louing Colatine Shall hate thee euen in death : then saue	Colatine	Colatine	np1	<input type="button" value="Edit"/>
them Orators , To pleade the cause of absent Colatine , your friend and kinsman .	Colatine	Collatine	np1	<input type="button" value="Edit"/>
displeasure : but foft , some of my Lord Colatines menlye in the next chamber , I care not	Colatines	Collatine	npg1	<input type="button" value="Edit"/>
B•utus , Valerius Horatius . Arnus , Sc•uola , Colatine .	Colatine	Collatine	np1	<input type="button" value="Edit"/>
Did nobody see my Lord Colatine ? oh , my Lady commends her to you , heer's	Colatine	Collatine	np1	<input type="button" value="Edit"/>
Cosen Colatine the news at Rome ?	Colatine	Collatine	np1	<input type="button" value="Edit"/>
, that with such sad presage , Distu•bed Colatine , and noble Brutus Are hurried from the	Colatine	Collatine	np1	<input type="button" value="Edit"/>
Horatius , Valerius , Sc•uola , Lucretius , Colatine .	Colatine	Collatine	np1	<input type="button" value="Edit"/>

Correction	Corrector	Approver	Applier	Status
Update A04206-6-b-1460: To myfe (myfe , zz) my (my , po11) Rebecca .	martin Nov. 10, 2013, 10:07 a.m.	None None	None None	Unapproved
Delete A11909_03-44-a-0070: THE THYRD TRAGEDY OF L. ANN AEVS (AEVS , np1) Seneca : entitled Thebais , translated	martin Oct. 10, 2013, 4:40 p.m.	None None	None None	Unapproved
Update A11909_03-44-a-0060: THE THYRD TRAGEDY OF L. ANN (and , ee) ANNAEVS (Annaeus , np1) AEVS Seneca : entitled Thebais , translated	martin Oct. 10, 2013, 4:40 p.m.	None None	None None	Unapproved
Delete A14875-44-b-3320: ENG. and (and , ee) E.	nayoon Sept. 15, 2013, 12:08 a.m.	None None	None None	Unapproved
Update A14875-44-b-2140: that belong to Great men remember th' ould wides (wides , ng1) wives (wife , n2) tradition , to be like the Lyons ith Tower	nayoon Sept. 15, 2013, 12:06 a.m.	None None	None None	Unapproved
Delete A14875-44-a-0450: Thou hast to good a face to be a hang-man , (,,) If thou be doe thy office in right forme	nayoon Sept. 15, 2013, 12:03 a.m.	None None	None None	Unapproved
Update A14875-43-b-2160: Sirha (sirrah , n1) Sirah (sirrah , n1) you once did strike mee , Ile strike you	nayoon Sept. 15, 2013, 12:01 a.m.	None None	None None	Unapproved
Update A14875-43-b-1080: , take this president : Man may his Fate foresce (foresee , vvi) foresee (foresee , vvi) , but not preuent . And of all Axiomes this	nayoon Sept. 15, 2013, 12:01 a.m.	None None	None None	Unapproved
Delete A14875-43-a-4060: A matachine it By your drawne swords . (seemes (seem , vvz) , Chuch-men turn'd reuellers .	nayoon Sept. 15, 2013, midnight	None None	None None	Unapproved
Delete A14875-43-a-4050: A matachine it By your drawne swords . ((,)) seemes , Chuch-men turn'd reuellers .	nayoon Sept. 15, 2013, midnight	None None	None None	Unapproved
Insert A14875-43-a-3990: A matachine it (it , pn31) seemes (seem , pn31) By your drawne swords . (seemes , Chuch-men	nayoon Sept. 15, 2013, midnight	None None	None None	Unapproved
Update A14875-43-a-3990: A matachine it (it , pn31) it (it , pn31) By your drawne swords . (seemes , Chuch-men	nayoon Sept. 15, 2013, midnight	None None	None None	Unapproved
Update A14875-43-a-3990: A matachine it (it , pn31) seemes (seem , pn31) By your drawne swords . (seemes , Chuch-men	nayoon Sept. 15, 2013, midnight	None None	None None	Unapproved
Delete A14875-43-a-0120: You see the Fox comes many times short ho me (i , pn11) , 'Tis here prou'd true .	nayoon Sept. 14, 2013, 11:58 p.m.	None None	None None	Unapproved
Update A14875-43-a-0110: You see the Fox comes many times short ho (ho , uh) home (home , uh) me , 'Tis here prou'd true .	nayoon Sept. 14, 2013, 11:58 p.m.	None None	None None	Unapproved
Update A14875-43-b-2310: They shoot and run to him &	nayoon	None	None	

Filter Corrections

Corrector: (All) ▾
Status: Unapproved ▾
Applied: (All) ▾
 Unapproved
 Approved
 Rejected
 Held



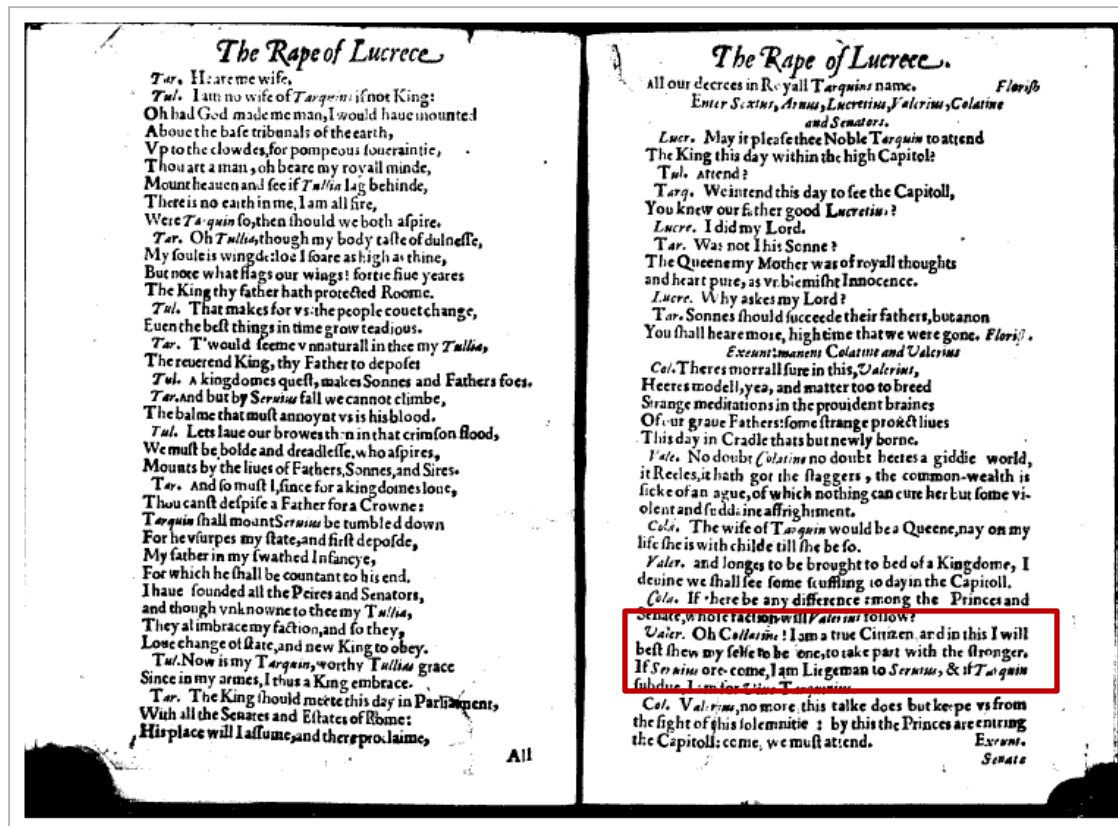
Title: The rape of Lucrece a true Roman tragedie. With the seuerall songs in their apt places, by Valerius, the merrie lord amongst the Roman peeres. Acted by her Majesties Seruants at the Red-Bull, neere Clarken-well. Written by Thomas Heywood.

Author: Heywood, Thomas, d. 1641.

Collection: Early English Books Online

Table of contents | Add to bookbag

◀ prev next ▶



Exeunt: manent Colatine and Val•...rius

Col.

Theres morrall sure in this, *Valerius*,
Heeres modell, yea, and matter too to breed
Strange meditations in the prouident braines
Of our graue Fathers: some strange proiect liues
This day in Cradle thats but newly borne.

Vale.

No doubt *Colatin*•... no doubt hee•...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|olent and •...ddaine affrightment.

Cola.

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.

Cola.

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

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 *Se*•...ius ore-come, I am Liegeman to *Serutus*, & if *Ta*•...quin subdue, I am for *Viue Tarquinius*.

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.

//*[@id="doccontent"]/div/div[39]

```
<div class="sp">
  <div class="speaker">Valer.</div>
  <p>Oh Collatin<span class="gap">•...</span>! I
    am a true Cittizen and in this I will best shew my selfe to
    be one, to take part with the stronger. If
    <span class="rend-italic">Se<span class="gap">•...</span>ius</span>
    ore-come, I am Liegeman to <span class="rend-italic">Serutus,</span>
    &amp; if <span class="rend-italic">Ta<span class="gap">•...</span>quin</span>
    subdue, I am for <span class="rend-italic">Viue Tarquinius.</span>
  </p>
</div>
```

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)

www.pov.bc.ca x Internet Arch x W Lucius Tarqui x The Shakespe x cts canonical x Canonical Te x A Brief Guide x de commentariis.net

decommentariis.net

Apps Gcalendar Gmail Plug-ins Timesheets Different Flash VoyagerFundRpts WhiteHouse OA HireTouch Other bookmarks


de commentariis Text editions About Contact

de commentariis β (on commentaries)

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.

Some sample texts: [Caesar, *Gallic Wars*](#).
[Sallust, *Catiline's Conspiracy*](#). [Cicero, *On The Republic*](#).
[Apollonius Rhodius, *Argonautica*](#). [Herodotus, *The Histories*](#).
[Pausanias, *Description of Greece*](#).

All available works >>

 The text contents of this work are licensed under a [Creative Commons Attribution-ShareAlike 4.0 International](#)

Canonical Text Services

Center for Hellenic Studies: Canonical Text Services

[CTS: home](#) | [catalog of texts](#) | [credits](#)

Shakespeare, *Sonnet 35*: 35

Edition of 1609

(= urn: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.

Canonical Text Services

Center for Hellenic Studies: Canonical Text Services

CTS: [home](#) | [catalog of texts](#) | [credits](#)

Shakespeare, *Sonnet 35*: 35.10

Edition of 1609

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

10 Thy adverse party is thy advocate,

[prev](#) | [next](#)

XML

Go directly to passage

reference

Lines of context: ▼

Built with the cts3 library, version cts3-beta-06, packaged on August 23, 2010

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