Position Statement

Submitted for W3C Workshop on Annotations

2 April 2014, San Francisco, California

Timothy W. Cole & Thomas G. Habing

University of Illinois at Urbana-Champaign

In his 2000 presentation at King's College London, John Unsworth identified annotating as a "scholarly primitive" core to scholarship and scholarly communication.[1] The advent of digital publishing, large-scale retrospective digitization, various forms of online discourse seemed to offer abundant opportunities to migrate the scholarly activity of annotating to the Web and to extend its reach and usefulness; yet so far, at least for most scholars, this promise has yet to be achieved. Additional work is needed, especially with regard to sophisticated use cases involving a mix of automatically generated and manually created annotations and with regard to lowering barriers for interoperating across collections of content and annotations. This position statement argues that the collaborative scholarly curation of retrospectively digitized resources is an important use case integrating workflows that could be greatly facilitated by a robust, fulsome annotation data model accompanied by practical guidance addressing the design of annotation-related APIs and the implementation of tools to create and consume annotations.

Our Perspective:

The authors participated in the Open Annotation Collaboration project (2009-2013) and are co-founders of the Open Annotation Community Group.[2] We approach annotation from the perspective of digital library developers and university scholars who use digital content in our own research and provide services over digital content to other scholars in academic settings. Professor Timothy Cole (prospective participant in the workshop) is Mathematics and Digital Content Access Librarian at Illinois and has been a member of the University of Illinois faculty since 1990. His research focus since 1994 has been on digital library design and interoperability and metadata workflows. Thomas Habing is Manager of the University Library's Software Development Group and has been a member of academic staff at Illinois since 1997, working primarily on the design, implementation and development of digital libraries and repositories. Definition and development of new services to support annotation is an active area of research for us, e.g., as part of planned work on the *HathiTrust Research Center*[3] and *Emblematica Online*.[4] Because scholarly annotations often have lasting value, annotation description and modeling are also of interest in connection with our work on *Medusa*, the Library's digital preservation repository.

---

[1] Unsworth, building on a theme from Aristotle, uses the term scholarly primitives "to refer to some basic functions common to scholarly activity across disciplines, over time, and independent of theoretical orientation." Unsworth, John. 2000. "*Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this*?" Presented at Humanities Computing: formal methods, experimental practice, King's College, London, May 13, 2000. Available: http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html

[2] http://www.w3.org/community/openannotation/

[3] http://www.hathitrust.org/htrc

[4] http://emblematica.grainger.illinois.edu/

Our View:

Annotations are useful to a wide range of users in a number of contexts and for a variety of reasons; in the context of this Workshop we are especially interested in annotation applications that facilitate the use and collaborative scholarly curation of retrospectively digitized resources.

Consider as examples projects like the Text Creation Partnership (TCP)[5] and the HathiTrust Digital Library[6] / Google Book scanning initiative. These projects provide scholars with unprecedented access to retrospectively digitized content:

- Working from digitized scans of microfilmed primary sources, TCP has so far manually transcribed and encoded in SGML TEI (Text Encoding Initiative) some 50,000 English language works published between 1463 and 1700. Another 20,000 are scheduled to be transcribed by the end of 2015, and the resulting collection will include at least one version of every English book published before 1700.

- HathiTrust includes retrospectively digitized surrogates for more than 11 million volumes digitized from print sources held by 90+ libraries worldwide. The vast majority of these surrogates consist of Google-scanned page images and corresponding plain text transcriptions generated by automated OCR (Optical Character Recognition).

However, to be truly useful for many scholars, both collections need additional human and machine-mediated curation. Some of the overlapping curation needs of these disparate collections illustrate our annotation use case. The text transcripts in these two repositories differ significantly in terms of mark-up (plain-text versus TEI) and how made (keyboarded versus OCR'd), but the transcripts found in both collections contain errors. For certain kinds of scholarly analysis, e.g., literary studies, these errors are numerous enough to inhibit the use of the digital surrogates by many scholars. By the standards of many scholarly disciplines, even the human-transcribed TCP texts contain a large number of errors per work (see below).

A way to address this problem for sub-collections of retrospectively digitized content is through distributed, collaborative (i.e., expert-sourced), machine-aided correction and curation. To conserve the time of human copy-editors, compute resources can be used at the outset to identify potential errors, e.g., incompletely transcribed words, non-dictionary words, etc. Annotations can be generated to record the identification of suspect words. Human agents can then examine the suspect text highlighted through this computer-based annotation. They in turn add annotations recommending updates, deletions, and insertions as required to correct the error, or they can annotate the suspect word as definitively not an error. A human reviewer adds another layer of annotation, approving or rejecting recommended edits and determinations regarding suspect words. Curatorial annotation chains are then used to preserve provenance (if underlying texts are subsequently modified), applied real-time when rendering texts, and/or used to ensure scholarly attribution (i.e., credit) for edits recommended. Using these chains also, variant versions of a text can be maintained and made available side-by-side should edits be contested.

[5] http://www.textcreationpartnership.org/tcp-eebo/
[6] http://www.hathitrust.org/

AnnoLex (Northwestern University): A Concrete Example of Curatorial Workflow

An ongoing digital curation effort being organized and implemented by Martin Mueller, Phil Burns and Craig Berry of Northwestern University has created a tool called AnnoLex,[7] and has used this tool for curating a subset of TCP texts -- specifically 600+ Early Modern English plays by authors other than Shakespeare. These 600+ plays contain about 15 million words. As considered against total word count, the original TCP transcriptions are quite good; however, there are enough errors to be troublesome for certain kinds of scholarly use and analysis. Just in terms of incompletely transcribed words there are estimated to be over 60,000 instances across these 600+ plays (i.e., an average of almost 100 incompletely transcribed words per play). There are additional classes of errors as well, e.g., words incorrectly broken apart, words incorrectly concatenated, etc. Sampling suggests that the average error rate for Google/HathiTrust OCR is substantially higher, especially for texts printed before 1850.

AnnoLex illustrates the kind of collaborative curation workflow that could be captured, shared and preserved as annotations. Figure 1 (end of this brief), a partial screen shot from AnnoLex, shows some of the incompletely transcribed words (in context) found in one of the TCP plays. Recorded locally within the AnnoLex tool backend, these snippets could be modeled as annotations for export (e.g., for export back to the shared repository holding the text). In essence AnnoLex, like other tools and prototypes for this sort of work, then stores expert-sourced recommendations for text correction and curation as annotations (illustrated in Figure 2), pending review by an authorized editor. These recommended edits are reviewed and (if approved) applied to local copies of the texts. In AnnoLex these corrections are made on top of an adorned version of the text -- i.e., a copy of the text in which each word has been machine tagged (conceptually another kind of annotation) as to lemma and part of speech. AnnoLex was piloted last year by undergraduate students at Northwestern in an experiment that proved quite successful.[8]

Modeling considerations & our suggestions

Obviously, no one group or community is likely to curate all of TCP, let alone even a small fraction of HathiTrust. But there is considerable interest in curating sub-collections and worksets of these collections to support communities of scholars. We are interested in generalizing curatorial tools and workflows like AnnoLex such that corrections can be maintained or at a minimum exported as annotations against copies of text stored in repositories. Such stored annotations can result in changes to canonical copies or can be applied on a just-in-time basis when rendering from the repository. Curatorial annotation supports maintenance of provenance and attribution. It supports a Web model of curation and storage (e.g., expert-sourcing) and allows for multiple corrected variants should there be disagreements among scholars about edits.

The kind of collaborative scholarly workflow suggested above has implications for how we need to think about annotation modeling, the management and storage of annotations, client and server side annotation-related APIs, annotation system implementation, etc. Many of the issues in curatorial annotation workflows are common across other use cases; we feel especially that these three issues are worth highlighting.

[7] http://annolex.at.northwestern.edu/
[8] http://cscdc.northwestern.edu/blog/?p=867

- *A need for robust anchor schemes*: Current text curatorial tools that are closely coupled to the local storage of texts tend to use idiosyncratic annotation target anchoring -- e.g., a local, sequence-based scheme for word identifiers with gaps in numbering included to allow for insertions. This is especially likely if the intent is make corrections in the base text as soon as reviewed. In a closed system this is perfectly fine because the annotations are not meant to persist and be shared; however, for provenance purposes or for exporting curatorial annotations, mappings to more generic anchor approaches are needed. This requires that adequate shared anchor (e.g., annotation target) schemes exist.

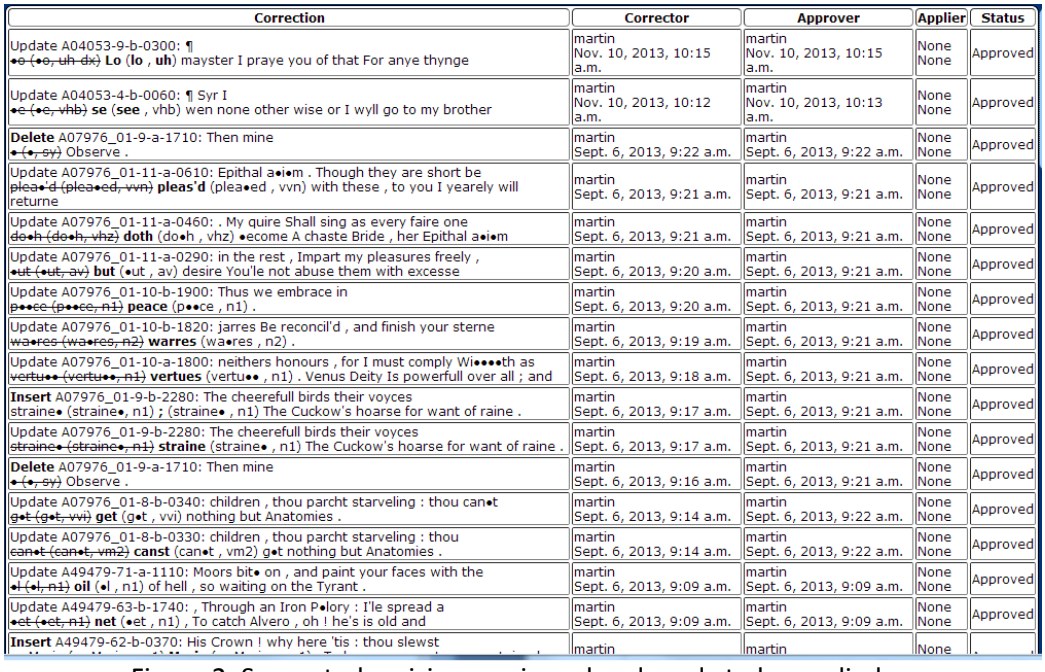
- *Extension of current model to support repeated segments*: Certain kinds of errors tend to recur throughout a given text's transcription. The transcriber, OCR, or part of speech tagger may consistently mishandle an uncommon proper name that is not found in any dictionary. Or a particular character in a particular font may be difficult to recognize, resulting in a recurrent incomplete transcription, e.g., the string *plea●'d* instead of the word *pleas'd* (the dot being a common placeholder for a character that could not be recognized). This suggests the need for a selector class having instances that describe multiple (return a list of) segments within a resource rather than only a single segment. Essentially oa:SpecificResource instances having oa:Selectors of this class would provide identity and an easy way to target unenumerated ad hoc aggregations in a resource. The enumeration of aggregation members could be unknown when the annotation is created, as long as the description provided by the selector is known to be determinate. Such selectors should be allowed and explicitly addressed in our model and in documented use cases.

- *Cross Resource Annotations*: Similarly, errors and other classes of curatorial annotation targets may repeat across volumes in a multi-volume work or even across multiple digitizations of a given manifestation or expression of a work. This suggests the need to express a specific target that exists simultaneously in multiple resources. This is similar to, but perhaps more general than the Cross Version Annotation and Cross Format Annotation use cases described elsewhere in the 24 February Editor's Draft of Annotation Use Cases. As in that discussion, the most straightforward solution (and perhaps the only viable solution) is to define a resource (identified with a single URI) that encompasses the multiple volumes or digitizations involved. Alternatively it may be feasible to consider specificResources (constrained as to Selectors) which could be allowed to have a multiplicity of Sources.

Our focus in this discussion has been annotation workflows that support curation of scholarly resources, but we think that the 3 requirements listed above support other scenarios as well. For example, in a less scholarly context, a seller may wish to annotate multiple entries in his or her product database to indicate that each of these items is on sale next week. Or a publisher may wish to annotate that an author name is misspelled in multiple articles appearing in multiple formats. An environmental scientist may wish to use a query to annotate multiple data rows in a database, and may wish for this annotation to persist and apply as new rows are added to the database. A mathematician may wish to comment on (i.e., annotate) a proof mentioned multiple times in an article and also in multiple articles. These illustrations suggest a need for a selector class that supports giving identity to repeated segments appearing in a resource. In one context or another we would like to see this potential possibility discussed at the upcoming Workshop.

Spelling in Context	Spelling	Lemma	POS	Edit
raise these cries , lodg'd in thy flaughtered **arm•** some base Groome dies , And Rome that hath	arm•	arm•	n1	Edit
Ar• .	Ar•	Ar•	np1	Edit
let mischiefe frown , With all his terror **ar•d** with ominous fates , To all their spleenes	ar•d	ar•d	vvn	Edit
Behis hand **ar•d** with an imperiall Scepter .	ar•d	ar•d	vvn	Edit
relish , a note , a tone , we must get an **a•re** betwixt vs.	a•re	a•re	n1	Edit
He **beare•** her out	beare•	beare•	n1	Edit
conuey , And in this gift thou dost thy bed **be•ray** . To morrow we shall meete , this night	be•ray	be•ray	vvb	Edit
may tast of louers blisse , Be merry and **ble•h** , imbrace and kisse , That Ladies may say	ble•h	blesh	j	Edit
cates , That straight dissolue to purity of **blo•d** . That keep the veines full , and enflame	blo•d	blo•d	n1	Edit
Oh this were a **brau•** controuersie for a Iury of weomen to arbitrate	brau•	brau•	n1	Edit
repentance dwell , It is perhaps the sansing **b•ll** , That rings all in to heauen or hell :	b•ll	b•ll	n1	Edit
B•u .	B•u	B•u	np1	Edit
Oh **Collatin•** ! I am a true Cittizen and in this I will	Collatin•	collate	vvg	Edit
matter where if frō the court , I'le home to **Collatin•** , And to my daughter Lucr•ce ; home breedes	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
Since **C•urt** and Country both grow proud , And safety	C•urt	c•urt	n1	Edit
corpes wele embrase , And when we sea ha **dea•** We ‹◇› will cry alasse . Fala la lero la	dea•	dea•	n1	Edit
corpes wele embrase , And when we sea ha **dea•** We ‹◇› will cry alasse . Fala la lero la	dea•	dea•	n1	Edit
in many ballads , Iohn for the king downe **din•** , Iohn for the king , has eaten many sall•ts	din•	din•	n1	Edit
What wil the madman **doe•**	doe•	doe•	n1	Edit
By a God you sweare to doe a de•ils **d•de** : sweete Lord forbea• By the same Ioue I	d•de	d•de	vvd	Edit
so much hurt as to desire her cōpany vpon **ear•h** agin yet vpō my honour , though she be	ear•h	ear•h	n1	Edit

Figure 1: Illustrations of incompletely transcribed words from a TCP play



Correction	Corrector	Approver	Applier	Status
Update A04053-9-b-0300: ¶ ~~•e (•e, uh dx)~~ **Lo** (**lo** , **uh**) mayster I praye you of that For anye thynge	martin Nov. 10, 2013, 10:15 a.m.	martin Nov. 10, 2013, 10:15 a.m.	None None	Approved
Update A04053-4-b-0060: ¶ Syr I ~~•e (•e, vhb)~~ **se** (**see** , vhb) wen none other wise or I wyll go to my brother	martin Nov. 10, 2013, 10:12 a.m.	martin Nov. 10, 2013, 10:13 a.m.	None None	Approved
Delete A07976_01-9-a-1710: Then mine ~~• (•, sy)~~ Observe .	martin Sept. 6, 2013, 9:22 a.m.	martin Sept. 6, 2013, 9:22 a.m.	None None	Approved
Update A07976_01-11-a-0610: Epithal a•i•m . Though they are short be ~~plea•'d (plea•ed, vvn)~~ **pleas'd** (plea•ed , vvn) with these , to you I yearely will returne	martin Sept. 6, 2013, 9:21 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Update A07976_01-11-a-0460: . My quire Shall sing as every faire one ~~do•h (do•h, vhz)~~ **doth** (do•h , vhz) •ecome A chaste Bride , her Epithal a•i•m	martin Sept. 6, 2013, 9:21 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Update A07976_01-11-a-0290: in the rest , Impart my pleasures freely , ~~•ut (•ut, av)~~ **but** (•ut , av) desire You'le not abuse them with excesse	martin Sept. 6, 2013, 9:20 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Update A07976_01-10-b-1900: Thus we embrace in ~~p••ce (p••ce, n1)~~ **peace** (p••ce , n1) .	martin Sept. 6, 2013, 9:20 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Update A07976_01-10-b-1820: jarres Be reconcil'd , and finish your sterne ~~wa•res (wa•res, n2)~~ **warres** (wa•res , n2) .	martin Sept. 6, 2013, 9:19 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Update A07976_01-10-a-1800: neithers honours , for I must comply Wi••••th as ~~vertu•• (vertu••, n1)~~ **vertues** (vertu•• , n1) . Venus Deity Is powerfull over all ; and	martin Sept. 6, 2013, 9:18 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Insert A07976_01-9-b-2280: The cheerefull birds their voyces straine• (straine•, n1) **;** (straine• , n1) The Cuckow's hoarse for want of raine .	martin Sept. 6, 2013, 9:17 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Update A07976_01-9-b-2280: The cheerefull birds their voyces ~~straine• (straine•, n1)~~ **straine** (straine• , n1) The Cuckow's hoarse for want of raine .	martin Sept. 6, 2013, 9:17 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Delete A07976_01-9-a-1710: Then mine ~~• (•, sy)~~ Observe .	martin Sept. 6, 2013, 9:16 a.m.	martin Sept. 6, 2013, 9:21 a.m.	None None	Approved
Update A07976_01-8-b-0340: children , thou parcht starveling : thou can•t ~~g•t (g•t, vvi)~~ **get** (g•t , vvi) nothing but Anatomies .	martin Sept. 6, 2013, 9:14 a.m.	martin Sept. 6, 2013, 9:22 a.m.	None None	Approved
Update A07976_01-8-b-0330: children , thou parcht starveling : thou ~~can•t (can•t, vm2)~~ **canst** (can•t , vm2) g•t nothing but Anatomies .	martin Sept. 6, 2013, 9:14 a.m.	martin Sept. 6, 2013, 9:22 a.m.	None None	Approved
Update A49479-71-a-1110: Moors bit• on , and paint your faces with the ~~•l (•l, n1)~~ **oil** (•l , n1) of hell , so waiting on the Tyrant .	martin Sept. 6, 2013, 9:09 a.m.	martin Sept. 6, 2013, 9:09 a.m.	None None	Approved
Update A49479-63-b-1740: , Through an Iron P•lory : I'le spread a ~~•et (•et, n1)~~ **net** (•et , n1) , To catch Alvero , oh ! he's is old and	martin Sept. 6, 2013, 9:09 a.m.	martin Sept. 6, 2013, 9:09 a.m.	None None	Approved
Insert A49479-62-b-0370: His Crown ! why here 'tis : thou slewst	martin	martin	None	

Figure 2: Suggested revisions reviewed and ready to be applied.