Transcription for Improved Research, Teaching and Learning at Harvard: Library Lab proposal, 4/23/2012

Submitted by: Harvard University Archives (Robin McElheny), Houghton Library (Susan Pyzynski), Schlesinger Library (Amy Benson)

Please note: A new first paragraph (just below) and a new measurement goal (p. 4) has been added to address the review committee's comments on the draft. New text is in Italics.

Summary

The goal of this project is to identify a transcription tool for use at Harvard from the many candidates already available. After a first phase spent assessing user needs, existing tools will be reviewed with the object of meeting user needs and ensuring that the tool can successfully integrate with Harvard's Library and academic computing systems. Two promising approaches to participatory transcription are the MIT Edgerton Notebooks project and the National Archives and Records Administration (NARA) Citizen Archivist project. In addition to assessing the functionality and platforms of existing tools, the team will also evaluate a tool's success in terms of adoption by specific research communities and integration into classrooms.

Transcription Tools

With the increasing interest in text-mining, transcription tools are emerging as an important component of the digital research and scholarship toolkit. The 2012 AHA conference devoted a session to transcription projects and technology (see *Crowdsourcing History: Collaborative Online Transcription and Archives* at http://aha.confex.com/aha/2012/webprogram/Session6679.html). At Harvard, there is growing interest in the use of Harvard's collections in the classroom (see http://news.harvard.edu/gazette/story/2011/04/objects-of-instruction/) and in providing faculty, students, researchers, and librarians/archivists the ability to interact with original texts. The Collaborative Annotation Tool, created by the Academic Technology Group in partnership with Harvard faculty members (see http://atgportfolio.fas.harvard.edu/node/4) represents a first step in this direction, but there is no interactive tool in place at Harvard to create, store, index, search, and share transcriptions of digitized documents maintained by the Library. There are an estimated 26,500 linear feet of pre-1900 manuscripts in Harvard's special collections and archives, a good portion of which were created before the regular use of the typewriter. [1] Due to their format, they are not full-text searchable, require heavy metadata creation to make them accessible, and are, in many cases, out of reach to students, faculty, librarians, archivists and other scholars alike.

A multi-purpose transcription tool for Harvard

So far, most of the archival and manuscript materials that have been digitized at Harvard are in formats that are available as static page images to be flipped through. And because Optical Character Recognition (OCR) software cannot interpret handwritten, typescript, or early printed documents, these

[1] Based on calculations of pre-1900 manuscripts in Harvard's libraries and archives as captured in *Results of a Survey of Harvard's Manuscripts and Archives Collections (MASC),* February 2010.

materials are not available for full-text searching. *The Papers of Samuel Williams* (http://discovery.lib.harvard.edu/?itemid=|library/m/aleph|012324967), *Samuel Johnson letters, 1731-1784 (http://nrs.harvard.edu/urn-3:FHCL.Hough:hou00245),* and *Charlotte Perkins Gilman (http://nrs.harvard.edu/urn-3:RAD.SCHL:sch00019)* are cases in point.

Transcriptions of handwritten documents can serve multiple avenues of research. They can support full-text searching functionality that enables and optimizes discovery of informational content by researchers. Additionally, if transcribed texts are made openly available to scholars and students, in combination with descriptive metadata for the objects, research questions best addressed by the mining and analysis of large swaths of the historic record could more readily be undertaken. Two such examples are "Quantifying the evolutionary dynamics of language" (http://www.nature.com/nature/journal/v449/n7163/full/nature06137.html) and "300 years of list-making: Personal inventories spanning three centuries are helping researchers unlock the mysteries of how economies edge towards growth and prosperity" (http://www.cam.ac.uk/research/features/300-years-of-list-making/).

Beyond the content itself, the ability to interact with text through the transcription process is equally important in a teaching environment. We digitize collections and make them available online in order to make history more accessible and to encourage people (researchers, students, the general public) to learn about and make connections with the past. How a research community, such as a student and a class, interact with a text and participate in the development of creating a public, living record of a document is also a real benefit of the transcription process. Transcription and annotation give researchers, students, and others the opportunity to move beyond consuming information to an active dialogue with primary sources.

There are a number of open-source transcription and annotation tools that have been developed by the scholarly and library community. Nearly all of these were designed to meet the requirements of specific projects and range from a simple interface based on e-mail that supports crowd-sourced transcription (see the Civil War Diaries & Letters Transcription Project at http://digital.lib.uiowa.edu/cwd/transcripts.html) to a highly focused tool that supports line-by-line transcription and annotation of selected medieval manuscripts (see T-PEN at http://t-pen.org/TPEN/). To date, none of the projects have been designed to interact across a number of repositories, especially important at Harvard where the riches of hand-written materials span dozens of libraries and archives. A comparative list of existing transcription tools follows below. In addition, none of these tools fully address the dual functionality of transcription and annotation that would enable users to share and build on the analysis of primary sources in an interactive environment.

The goal of this project is to identify one transcription tool from the many candidates already available and to draw up specifications for further development to provide three levels of functionality.

1. An easy-to-use layer for capturing transcriptions and associating the transcribed text with corresponding digital images to support full-text searching. Transcriptions could be made private or public by the user. Support for multiple versions of transcriptions would allow for comparison among versions, such as in a classroom setting, and possible future developments such as machine comparisons that could identify the "best" matches.

2. A full-service layer, a transcription dashboard of sorts, for transcribing, annotating, and tagging documents, with the ability to share annotations and tags among designated communities such as classes or interest groups. This level of functionality could be used for collaborative, long-distance projects for groups of researchers or for a multi-repository project such as reuniting like materials at Harvard or with repositories outside the University. In addition, we envision a group of students working collaboratively on a documentary transcription and analysis project.

3. A third layer that would allow open access to the entire corpus of transcribed text. Subsets of the available texts could be searched for and identified based on the corresponding metadata about the objects. Scholars could analyze the resulting text sets based on their research interests. Farther out, APIs could be developed for additional creative uses of the available texts.

Additional questions to address:

- Are different levels of functionality necessary for different audiences/users (class, researcher, librarian/archivist, public)?
- Is controlled access (login) advisable for use of the tool? What level of quality control, if any, is needed? (see http://manuscripttranscription.blogspot.com/2012/03/quality-control-for-crowdsourced.html for an extensive discussion of the various options)
- Can control be addressed by supporting the existence of multiple versions of transcriptions?

Project Plan

Our approach will be a two-phase effort.

Phase 1: The initial focus of this proposal will be to survey three audiences for a transcription/annotation tool - a selected group of faculty, representatives of specific research communities, and librarians and archivists at Harvard - regarding potential uses and desirable attributes of a transcription/annotation tool. In addition, we will consult with Harvard Library specialists in Preservation, Conservation, and Digital Imaging, Library IT, and members of the Academic Technology Group in HUIT. These conversations will allow us to better evaluate and plan for any programming that may be needed to ensure that the transcription tool can successfully integrate with existing Library and academic computing systems.

Estimated time: 3 months

Phase 2: Based on this input, we will initiate a pilot project to review the existing transcription and annotation tools to determine which ones best meet our needs. With the assistance of a student transcriber and a developer, we would expect to test 2-4 of the transcription tools, using a diversity of material from our three repositories. The test will involve staff- and community-generated transcription. A preliminary survey of existing transcription tools (see list below) reflects a range of functionality and user interfaces. Many of these tools have been developed with open licenses and could be adapted for implementation at Harvard.

Estimated time: 9 months

Budget and resource requirements:

The first phase of our project 2-3 months, as described above, would be an internal, evaluation component handled by the staff at the three repositories (Harvard University Archives, Houghton Library, Schlesinger Library), with help from developers at the Berkman Center to set up a test site.

Cost: Cost share of staff time borne by the individual libraries and costs associated with Berkman Center services.

The second phase of our project would be to undertake actual transcriptions, test and evaluate the possible tools, and work with a developer who will adapt and mount a tool that would be useful to all three libraries with consideration given to linking transcriptions to HOLLIS records or OASIS finding aids and determining how to maintain this information together in the DRS.

Cost: One student for transcription work across all three libraries, estimated at $5200 (10 hours per week for 20 weeks at $20 per hour)

Developer assistance – We do not know how to estimate time and cost but we will need the assistance of a developer from Berkman.

In the end, we anticipate that the project may take one year to complete and will require participation by the following people:

- Public Services archivists and librarians at Houghton, Schlesinger, and Archives to reach out to selected faculty, help develop a list of required functionality, test a pilot system, recruit and oversee class/researcher participation in testing, and determine specifications for the user interface
- Archivists and librarians at Houghton, Schlesinger, and Archives who are responsible for digital object management to test the pilot system and determine functional requirements for back end transcription management
- Student assistant for transcription
- IT/developer specialists to develop a test platform and identify technical requirements for production implementation

Measuring benefits/success of the project:

At the end of the pilot project, archivists and librarians will assess the following aspects of the test system through user studies and statistical reports, possibly partnering with the Simmons GSLIS:

- *Platform on which the tool is built*
- Ease of use
- Classroom adoption and feedback
- Popularity of tool
- Reasonable accuracy and utility of resulting transcriptions
- Ease of system management
- Researcher interest in transcribed documents

A comparative list of existing transcription tools follows below.

	Name KB: first impressions from looking at a project implementing the tool	Maintainer	License	Platform	Text Type	Hosted?	TEI?	CMS Integration	Unique Features	Project URL	Code URL	Implementing Sites	Twitter Account	"Pages/Records Transcribed"
1	**Wikisource** KB: I like-- easy to navigate, understand, use. Page-turner. No login. **Best**???	Wikimedia	GPL 2.0	MediaWiki	Free-form	Yes	No	Archive.org	Workflow management	http://en.wikisource.org/wiki/Main_Page	http://www.mediawiki.org/wiki/MediaWiki	NARA Citizen Archivist Dashboard	none, but #wikisource hashtag gets response	
2	**FromThePage** KB:straightforward-looking. Page image on right, transcription on left. Not sure if interface issues are the transcription tool or the site. Not pretty. Login required.	Ben Brumfield	AGPL 3.0	Ruby on Rails	Free-form	Yes	No	Archive.org	Semantic mark-up for indexing/annotation	http://fromthepage.com/	http://github.com/benwbrum/fromthepage/wiki	San Diego Natural History Museum Laurence J. Klauber Field Notes	@benwbrum	
3	**Scripto** KB: well-supported (NEH). Requires login. Not sure how to see transcriptions--maybe can't without login? Images appear to be from microfilm (or at least the one's I landed on).	Center for History and New Media at George Mason University	GPL 3.0	PHP library, MediaWiki	Free-form, wikitext	No	No	Omeka, WordPress, Drupal	Can be integrated into potentially any CMS or personal archive	http://scripto.org	• https://github.com/chnm/Scripto • https://github.com/omeka/plugin-Scripto • https://github.com/chnm/scripto-wordpress-plugin • https://github.com/chnm/scripto-drupal-module	Papers of the War Department, 1784-1800	@scriptotool	
4	**Bentham Transcription Desk** KB: TEI is obvious-- lots of boxes to put metadata into, also relatively easy to see what the point of everything on the screen is. No page-turning (ugly).	University of London Computer Centre; UCL Bentham Project	GPL2.0	MediaWiki	Free-form	Yes	Yes		Full TEI mark-up support; customized toolbar to automatically apply TEI tags to transcript	http://www.ucl.ac.uk/transcribe-bentham	http://code.google.com/p/tb-transcription-desk/	Forthcoming!	@transcribentham	As of 24 Feb 2012: 2,845 manuscripts (c.1.5 million words, plus extensive TEI markup)
5	**Scribe**	Zooniverse	MIT	jQuery/Ruby on Rails	"Structured data"	Upon application.	No	none	Blind triple-keying, data linked to images		http://github.com/zooniverse/Scribe	What's the Score at the Bodleian (earlier versions at OldWeather.org)	@the_zooniverse	
6	**PyBOSSA**	Citizen Cyberscience Centre/OKFN	AGPL 3.0	Python/GDocs	Tabular		No			http://pybossa.com/	https://github.com/PyBossa/pybossa		@pybossa	
7	**OpenScribe**	?	Perl	Drupal	Free-form	?	?	Drupal			http://code.google.com/p/openscribe/		none	

	Name KB: first impressions from looking at a project implementing the tool	Maintainer	License	Platform	Text Type	Hosted?	TEI?	CMS Integration	Unique Features	Project URL	Code URL	Implementing Sites	Twitter Account	"Pages/Records Transcribed"
8	**TextLab**	?	?	?	Free-form	No	Yes	?	Direct annotation of TEI add/del tags onto images.	http://mel.hofstra.edu/textlab.html		Melville Electronic Library		
9	**T-PEN**	St. Louis U Center for Digital Theology	EPL 2.0	Java/Javascript	Line-based **medieval**	?	Yes	users can create export pipelines that can export transcriptions directly into a CMS database (such as Drupal)	Direct linking of transcription to lines of text in image	http://digital-editor.blogspot.com/		http://t-pen.org/TPEN/	@DH_editor	
10	**Ancestry World Archives Project**	Ancestry.com	Proprietary	Installed .exe client	Structured data (Genealogy)	?	?	?	difficulty rating, context based help, multiple archive sources	http://community.ancestry.co.uk/awap		World Archive project and http://www.worldmemoryproject.org/ (essentially another 'way in')		
11	**Islandora TEI Editor**	UPEI (?)	GPL 3.0	Drupal/Fedora	Free-form	No	Yes	Fedora	TEI mark-up of documents hosted in Fedora	http://wiki.tei-c.org/index.php/IslandoraTEIEditor	https://github.com/Islandora/islandora_tei_editor	Public Records Office, Victoria http://prov.versi.edu.au/		
12	**FieldData**	Atlas of Living Australia/Gaia Resources	Mozilla Public License 1.1	Java			No			http://www.ala.org.au/get-involved/citizen-science/fieldata-software/	http://code.google.com/p/ala-citizenscience/	http://volunteer.ala.org.au/project/index/42780		

	Name KB: first impressions from looking at a project implementing the tool	Maintainer	License	Platform	Text Type	Hosted?	TEI?	CMS Integration	Unique Features	Project URL	Code URL	Implementing Sites	Twitter Account	"Pages/Records Transcribed"
13	**National Archives Transcription Pilot Project**	U.S. National Archives		Drupal	Free-Form			Drupal	Difficulty rating (Beginner, Intermediate, Advanced), lock out feature, commenting, links to online catalog	http://transcribe.archives.gov/			@USNatArchives	1,000+ pages (300+ records)
14	**Old Weather**									http://www.oldweather.org/				
15	**North American Bird Phenology Program**	USGS			Structured data					http://www.pwrc.usgs.gov/bpp/				560,271 cards transcribed; 1,104,494 cards scanned
16	**What's On the Menu?**	New York Public Library			Structured data					http://menus.nypl.org/			@nypl_menus	796,136 dishes from 12,541 menus
17	**Family Search Indexing** KB: not prose transcription, but fill-in-the box indexing by transcribing bits of data (hmm... Tolman??!!)	Family Search			Structured data (Genealogy)					https://indexing.familysearch.org/newuser/nuhome.jsf?3.9.6				
18	**Harold "Doc" Edgerton Project** KB: page-turned, a bit slow, nice lightbox display of notebook covers, good level of digitization for transcription, easy to understand pages. **Best**??? Best navigation.	MIT?			Free-Form					http://edgerton-digital-collections.org/notebooks				
19	**Civil War Diaries & Letters Transcription Project**	The University of Iowa Libraries			Free-Form					http://digital.lib.uiowa.edu/cwd/transcripts.html			@UIL_transcripts	As of 2/24/12: 9,043 pages
20	**Unbindery**	Ben Crowder		PHP/Javascript	Free-Form	Yes	?	Is a CMS		http://bencrowder.net/blog/category/unbindery/	https://github.com/bencrowder/unbindery	http://bencrowder.net/books/mtp/	@mormontexts	