Key

Q= Quarter *University = Texas A&M

L.M. = Laura Mandell (IDHMC/ARC)*

C.L. = Cushing Rare Books Library*

P.S. = Performant Software

P.R. = PRImA (University of Salford)

R.G. = Ricardo Gutierrez-Osuna*

R.F. = Rick Furuta*

S.Z. = SEASR (University of Illinois)

R.M. = R. Manmatha (University of Massachusetts)

I.M. = IMPACT (National Library of the Netherlands)

Appendix: Detailed Schedule of Tasks

Unit	Goal	Tasks	**Q1** 10/12–12/12	**Q2** 1/13–3/13	**Q3** 4/13–6/13	**Q4** 7/13–9/13	**Q5** 10/13-12/13	**Q6** 1/14-3/14	**Q7** 4/14-6/14	**Q8** 7/14-9/14
A. OCR Engine Develop-ment	1. Engines	Optimize what goes in: find optimal image settings using ImageMagick	**L.M.**							
		Optimize what happens inside: put Tesseract's line segmentation procedure into the Gamera Toolkit	**P.S.**							
		Optimize what comes out: create and tweak XSLT transforms that a) put xml outputs (hOCR, Gamera's xml) into the xml form required by Gale, ProQuest; b) create TEI-A; c) use whitespace to mark up paragraphs	**L.M.**	▨	▨	▨				
	2. Fonts	Select documents containing representative fonts & run them to see results, creating typed versions to test them against	▨							
		Create a font importation database	**C.L.**	▨	▨	▨				
		Scan samples of fonts from Cushing, Antwerp, St. Bride's	▨	▨	▨	▨				
		Train engines in fonts from EEBO/ECCO; train engines in and transcribe samples of font images from Cushing, Antwerp, St. Bride's	**L.M.**	▨	▨	▨	▨	▨	▨	▨
	CHECKPOINT 1: make sure font database needed		**Nov. 2012**							
	3. Testing	Add x-y coordinates for each line of the test data set, indicating place on the page image, making font documents usable to calibrate	▨							
		Calibrate the algorithm that compares OCR outputs with hand-typed text;	**R.M.**							
		Modify algorithm to compare OCR outputs with hand-typed text	▨	▨	▨	▨				
		Create API for sending us (and making available to all) early modern test set & comparison algorithm, and then use it to test all OCR engine tweaks				**R.M.**				
	4. OCR'ing EEBO and ECCO page images	Set up Taverna workflow to run OCR process			**I.M.**	▨				
		After getting best results, 93% accuracy or higher, run 260,000 documents through engines on HPC at 10 seconds per page			**P.S.**	▨	**L.M.**	▨	▨	▨
	CHECPOINT 2: make sure test set can be made automatically			**Jan. 2013**						
Milestone 1: we now know that 23.7 million pages can and will be 93% correct and are running through the engines – Sept. 2013										
B. Human-machine interaction	5. Crowd-sourcing a) Cobre	Launch Django server with instance of Cobre backed by D-Space allowing all 18thConnect and REKn members to create and save Frankenbooks	**C.L.**							
		Add features to Cobre that allow automated creation of structure that allows for filmstrip presentation, metadata-editing, font identification, and transcription	**C.L.**							
		Load page images of "unreadable" documents into Cobre along with other editions of the same title		**L.M.**						
		Conduct usability studies by consultants who are book history and early modern experts (Raven, Hume, and Mosley)			**R.F.**					
		Re-design tool and					**C.L.**			
		Re-work the interface based upon usability studies						**P.S.**		

	Goal	Tasks	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
	b) Aletheia Web	Create web version of Aletheia	P.R.	▨	▨	▨				
		Design interface and stand up Aletheia in 18thConnect and REKn, Ruby on Rails				P.S.				
		Conduct usability studies on graduate and undergraduate students, adjusting the design and interface				R.F. P.R.				
	c) Type-Wright	Add capacity for adjusting lines	P.S.	▨	▨					
		Add other features to tool, including red squiggly underline feature to thumbnail for indicating probable errors as indicated in post-processing output	P.S.	▨	▨					
	d) all tools	Conduct wide-ranging usability studies and measure effectiveness of all tools					R.F.	▨	▨	▨
		Re-work both the triage system based upon document evaluation (see last item of goal 6, immediately below) and the interfaces based upon usability and effectiveness studies							▨ P.S.	▨ ▨
	Milestone 2: Release Tools – Sept. 2013									
	6. Document Evaluation a) Check coordinates produced by OCR engine	Run clustering algorithm on word coordinates to isolate documents with too many letter sizes per letter		R.G.	▨	▨				
		Run clustering algorithm on line coordinates to isolate pages with inconsistently ordered lines		▨	▨	▨				
	b) Check N-grams and words	Count number of words that are unique and that contain internal punctuation other than hyphen	S.Z.	▨	▨	▨				
		Count number of impossible n-grams in three or four languages	▨	▨	▨	▨				
		Count number of unique words in the dictionary with 0, 1, 2, and 3 editing distances	▨	▨	▨	▨				
		Count number of replacement rules that apply	▨	▨	▨	▨				
	c) Find Document Signature	Select among 47,000 keyed texts the documents with OCR results that fail because of font id, line segmentation, and page-image inadequacy		R.G.	▨	▨				
		Measure these known failures using clustering and counting 6a. and b., immediately above, and correlate ranges of measures obtained into document signatures corresponding to specific engine failures (font, lines, bad images) if possible		▨	▨	▨				
	d) Use signals	Correlate typical n-gram errors in three languages with need for font training		▨	▨	▨				
		Count number of single-and-double character words in document		▨	▨	▨				
	e) Draw conclusions	Determine document signatures and signals that indicate what went wrong in OCR process, whether it was font misidentification, unknown layout, or unknown problems		▨	▨	▨				
	7. Optimize OCR Output with Human Assistance (Optimize HMI)	Set up automated triage system: font mis-id and unknown go to Cobre; layout indeterminacy goes to Aletheia Web			P.S.	▨				
		Select subset of documents in each tool to monitor			L.M.	▨				
		Based on usability studies and human-made improvements in document subset, determine how to optimize human / machine intervention (i.e., tool tweaking; adding automated processes for tasks that are too repetitive; not sending specific problems to the tools, or allowing agents to forward problems to 18thConnect / REKn directors					R.F.	▨	▨	▨
		Adjust measures that indicate where document needs to be sent based on degrees to which crowd is able and willing to help (first item in 7, immediately above)					▨	▨	▨	▨
	CHECKPOINT 3: confirm time/correctness correlation				**Apr. 2013**					
	8. Launch Crowd Tools	TypeWright and Cobre demo at pre-conference workshop at MLA, January 2013 (dhCommons, already scheduled)		L.M.						
		TypeWright/Cobre paper, REKN announcement, on Restoration and 18thC division panel, with James Raven, MLA, January 2013 (chair Catherine Ingrassia; already scheduled)		▨						

	Goals	Tasks	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
		Day-long pre-conference workshop on how to use TypeWright, Aletheia Desktop, and Cobre workshop, REKN announcement, at ASECS national meeting, 2013 (already scheduled)			L.M.					
		Set up editing groups to work on Cobre documents (Defoe Society, History of Science etc.)			L.M.	▨	▨	▨	▨	▨
		REKn Launch (Ray Siemens, Richard Cunningham): will apply by 1 March 2013/2014 for paper session to be held at SAA (Shakespeare Association of America) meeting in St. Louis and Vancouver								
Milestone 3: Document Evaluation Working – December 2013										
C. OCR Correction	9. Manually Correct the OCR Output	Set up and run "voting algorithm" to compare the outputs of the three engines and choose the reading that has the most votes				R.M.	L.M.	▨	▨	▨
		Create n-gram analysis and replacement rules		S.Z.	▨	▨				
		Create dictionary lookups by Levenshtein editing distance		▨	▨	▨				
		Develop parameters for replacement rules of name and place gazetteers		▨	▨	▨				
		Install Gazetteers from Underwood in Taverna					P.S.			
	10. Engage Humans in the Correction Process	Crowds work in Cobre, Aletheia Web, and TypeWright					L.M.	▨	▨	▨
		Re-run documents after people have identified fonts or diagramed the page layouts, then send documents to TypeWright					P.S.	▨	▨	▨
		Send all corrected documents to TypeWright as set up in 18thConnect and REKn					▨	▨	▨	▨
		Forward texts corrected in TypeWright by users (deemed reliable) to Gale, ProQuest, and the TCP, and index them in the ARC Catalog.								▨
	CHECKPOINT 4: mechanical correction improves by 60%							**March 2014**		
	11. Save the Data	Give corrected texts to the people who corrected them					L.M.	▨	▨	P.S.
		Help correctors create library-quality electronic editions					▨	▨	▨	▨
		Export metadata corrections to the English Short Title Catalog for review								▨
		Save correction histories to create a crowd-sourced correction data set in Institutional Repository (IR)					C.L.			P.S.
		Extract font identifications from Cobre Frankenbooks into Font History database, correlating ESTC number with typeface								
Milestone 4: 23.7 million pages now 97% correct, 99.9% once through TypeWright – Sept. 2014										
D. Dissemination	12. Release of Tools, OCR Workflow, and ESTC Databases to Improve knowledge	Release History of Font Importation Database				C.L.				L.M.
		Release database of documents needing rescanning by ESTC number								
		Submit for publication in REKn all revisions made in Cobre and saved by author (corrector) in the Texas A&M D-Space								L.M.
		Release the tools and Taverna workflow for download on Github and in IMPACT Competency Center								▨
	13. Strengthen and sustain Crowd intervenetion process	Create a plan for strengthening ARC support of NINES, 18thConnect, and REKn								L.M.
		Enlist Professors among special interest groups to lead (as editor/promoters) users of the tools for correcting and assisting OCR								L.M.
		Formulate a plan for how to record, monitor, and pool corrections made in tool instances worldwide								▨
	14. Publish Results	Publish History of Fonts Database and Rescanning Database in Institutional Repository (IR)								▨
		Publish Report on OCR'ing Early Modern Texts in IR and submit to CLIR								▨
		Submit paper on optimizing Human-Computer Interaction								R.F.
Milestone 5: Tools and Workflow are being used worldwide – Dec. 2014										