Initiative for Digital Humanities, Media, and Culture

Serving the Colleges of Architecture, Education,
Engineering, and Liberal Arts

TEXAS A&M UNIVERSITY

Laura Mandell
TAMU 4227
mandell@tamu.edu
Ph. 979.845.8345

June 30, 2012

Dr. Donald J. Waters, Program Officer
Dr. Helen Cullyer, Associate Program Officer
Scholarly Communications
The Andrew W. Mellon Foundation
140 East 62nd Street
New York, NY 10065

Dear Dr. Waters and Dr. Cullyer:

I write to request a grant of $733,794 in order to fund the project called "OCR'ing Early Modern Text," explained in detail in the attached text. Thank you.

Sincerely,

Laura Mandell
Professor, English
Director, Initiative for Digital Humanities, Media, and Culture (IDHMC)

227 John R. Blocker Building
4227 TAMU
College Station, TX 778434227

Tel. 979.845.3451 Fax 979.862.2292
www-english.tamu.edu

LAURA MANDELL
DIRECTOR, INITIATIVE FOR DIGITAL HUMANITIES, MEDIA, AND CULTURE
TEXAS A&M UNIVERSITY

OCR'ING EARLY MODERN TEXTS

Table of Contents

LAURA MANDELL
DIRECTOR, INITIATIVE FOR DIGITAL HUMANITIES, MEDIA, AND CULTURE
TEXAS A&M UNIVERSITY

OCR'ING EARLY MODERN TEXTS

I. Summary

Collectively, the US, the UK, and scholars around the globe face a problem: rare books and pamphlets from the early modern era which have not yet been made available digitally threaten to become invisible to future scholars. The mode of finding materials in special collections has not happened via metadata alone, and thus, insofar as finding aids and collection catalogs are supplanted by digital databases, much that has a value not reflected in its metadata—books bound with other books, authorship attributions made by readers and librarians—much of value could be lost. With the mountain of digital research materials growing ever larger, to use Vannevar Bush's metaphor,[1] early modern documents—everything from pamphlets to ballads to multi-volume poetry collections—preserved only by metadata records and page images could fall beneath notice, becoming very difficult for even the most devoted researchers to locate. But we can give early modern texts a higher digital profile. Optical Character Recognition software (OCR) could be used to create machine readable versions of these texts, making them more findable through being made fully searchable—increasing, as it were, their digital trail. OCR technology is now excellent, but when dealing with the vagaries of early modern printing technology and practices, as well as page images that have been digitized from microfilm, automated transcription can only go so far. We propose that adequately transcribing early modern texts from the image-resources already at hand can be accomplished via carefully orchestrated human-machine interaction. The overall effort will be a series of linked activities that build different kinds of tools and user communities which will collectively meet the challenges in the transcribing these page images for digital preservation. Each of these activities will be taken up by experts in the field and collectively managed with a focus on improving the accuracy of early modern texts, the page images of which have been mechanically transcribed (OCR'd).

Past and current efforts give us a clear sense of how to move forward and delineate clear benchmarks concerning what we will achieve. Our overarching goals are 1) to optimize the performance of three open-access OCR programs as they work on early modern texts by training them to "read" specific fonts; 2) to align the use of specific font training libraries with specific sets of documents via mapping the importation of typefaces from Europe into the London book trade; 3) to deploy, empirically test, and refine error-evaluation mechanisms that will allow us to determine what went wrong when OCR output is inadequate (Has the font been misidentified? Are the engines unable to find the base of each printed line, their focal point from which to identify printed

[1] Vannevar Bush, "As We May Think," *The Atlantic Monthly*, July, 1945: http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/ accessed June 1, 2012.

characters? Is the image inadequate?); 4) to use crowd-sourced assistance and correction applications (TypeWright, Aletheia, and Cobre) that are specific to the different types of errors (minor misreadings, line-segmentation, misidentification of fonts, bad page images) that have been determined to be mechanically uncorrectable; 5) to re-run documents after human assistance, save them after human correction, and weed out digital documents that are too flawed to be readable; and 6) to mirror and export this workflow procedure, our book history database, our database of unreadable digital page images, and our crowd-sourced correction tools so that others may use them in digitizing and transcribing early modern documents.

The objective of the work is to create a work-flow that libraries and collectors can use on early modern texts as well as information about what can and cannot be done mechanically, what works need to be re-imaged, what must be hand-typed in order to be digitized. Of the 260,000 texts that we will be working on, we estimate based upon past successes that we can produce: **162,730** (approximately **23.7 million pages**) at 97% correct, and **10,000** (**1.46 million pages**) approaching 100% (hand-corrected).
We will establish a sustainable crowd-sourced correction system that will continue to correct documents and assist in the OCR process beyond the two years of grant tenure.

II. Overview

We will create keyed versions of early modern texts that are far more correct than is now possible with the current set of tools —which is to say, we will be "**OCR'ing Early Modern Texts**." The corpus we are concerned with is organized according to "document" title. A document, be it a pamphlets, book, journal, broadside, encyclopedia, or poetry collection, for example, is comprised of all the material that falls under one title, ranging in size from one page to many volumes. But whereas a book historian creating a digital replica would focus on the document itself, capturing bindings and slip covers, for instance, we are primarily concerned with the typed texts in those documents, the words and their meanings. We speak of documents in this proposal insofar as the dataset we use is organized and counted by document: thus, Gale Cengage Learning's Eighteenth-Century Collections Online, one of our datasets, is made up of 182,000 documents but over 200,000 volumes—a document can be made up of multiple volumes.

By the end of the two years we will have created and disseminated a set of tools for further text correction and also OCR'ed a total of

- 151,200 documents printed 1721-1800, with improved correctness;
- 3,000 documents printed 1700-1720, improving correctness by 17%, making them correctable by humans to 100%;
- 8,000 documents printed between 1473-1700 that have never been OCR'd before, at 97%--again, available for correction to 100%;
- 100,000 documents, printed 1473-1800, for which the results are uncertain: if these are not 97% correct, we will know why and what needs to be done to make them that way.

We will create, publish, and disseminate worldwide the tools, worklow, and databases that we needed to create in order to accomplish this task, making available to libraries and museums, state of the art research in the fields of **cultural and book history**. We will make it available in the most systematic way it has yet been presented—as workflows, dictionaries, and databases—so that it can be used in the digital preservation processes.

We will **orchestrate human-machine work**, creating the means for interaction in these efforts that will inspire and demand the best work possible from both.

Most important, we will forge a pathway making possible contributions to the effort of preserving early modern texts by the scholarly community—both professors and the "citizen" scholars who care about preserving our cultural heritage. We will **foster and harness intellectual curiosity** by providing the means to work with early modern text, and for academics, professional rewards that will make doing this work possible.

III. Background

A drawing in the Appendix schematizes the early modern corpus with which we are concerned (p. 79). For 80,000 early modern documents ranging in page numbers from 1 to 1500, page images have been made but there are still too few transcriptions: 45,000 documents have been manually typed in order to allow searching a portion of the Early English Books Collection of 125,000 texts. For 182,000 more texts published between 1700 and 1800, documents comprising Eighteenth-Century Collections Online, only 2,000 texts have been manually typed, and the transcription of the digital page images that has been produced mechanically via OCR is not yet adequate for scholarly uses. While re-scanning these rare books and pamphlets for better images and manually transcribing the page images could solve this problem, both options are too expensive and too time-consuming to accomplish. Is there a way to improve the OCR results? We know that the performance of OCR engines is greatly increased when the engine is not trained to read everything in general but to look for specific letter shapes. We know that they are improved in this way from the development of proprietary engines such as ABBYY Finereader by the IMPACT group (IMProving ACcess to Texts): this group has been developing "font libraries" for the last ten years on Continental European texts from all eras. However, it is very expensive to use proprietary engines and processes such as those that were developed by the company called "Prime Recognition" which compares the outputs of several proprietary OCR engines. Therefore we propose to solve this problem by combining scholarship into the history of typefaces, new crowdsourcing technology, and newer, better OCR engines that are now open access, freely available to everyone. Additionally, we will augment the efficacy of OCR technology through orchestrating human intervention at key points in the OCR process by developing automated methods for determining what sets of documents need what specific kind of care.

For the last fifteen years or so, librarians and their vendors have known that we need to transcribe into typed text the 45 million page images of early modern texts because they will otherwise not be machine readable, nor indeed preservable in library-quality form. One group set to work on using OCR—Optical Character Recognition—technology in order to mechanically type the texts: Gale Cengage Learning, vendor to the

British Library, hired the OCR experts from the company called "Prime Recognition" to mechanically transcribe ECCO (Eighteenth Century Collections Online), a digitized set of documents spanning 1700-1800. Another group, a consortium of libraries spearheaded by the University of Michigan called the Text Creation Partnership (TCP), began to hand-type (or "key," as they say) the EEBO Collection, Early English Books Online, owned by ProQuest. The only typed, machine-readable versions of texts in that collection of documents, spanning 1473-1700, were hand-typed, and there will be approximately 45,500 of them by summer 2012. Currently, Google is OCR'ing the billions of page images comprising "Google Books" to make them searchable and usable for data-mining. Other initiatives are working on improving both OCR and typed text: The Visualizing English Print Group and the Dariah group at King's College London are developing the open-access OCR engine called "OCRopus"[2]; Martin Mueller and Phil Burns at Northwestern University are working on a central piece of Bamboo's Data Curation suite, Mophadorner 2.0, that can correct mis-typed text whether by human or machine.

It is through working based on what they have learned, as well as in tandem with these groups, that we see a clear way forward:

First, we wish to create our own version of the procedure used by Prime Recognition for ECCO and extend and improve it for EEBO.[3] The items in the ECCO catalog published in 1721 or after are currently approximately 95% correct (see Appendix, "Document Error Evaluation," pp. 73). Prime Recognition uses a "voting technology": they run documents through the top six commercial OCR engines and then take each engine's reading of a word as one vote for that particular word. Thus, if engines 1, 2, and 3 see "that," but 4 sees "clat," 5 "hat," and 6 "tat," "that" wins the vote. For this grant, we will adopt a similar process by using three open-access OCR engines: Tesseract, Gamera, and OCRopus. Prime Recognition's voting algorithm is their own, and so we will use our own, a different algorithm. We explain below our own particular method for voting among the three engines, but suffice it to say here that we will set up three engines to run early modern documents on the Brazos High Performance Computing Cluster (HPC) at Texas A&M University, supervised by Mandell's Initiative for Digital Humanities, Media, and Culture (IDHMC).

While we wish to adapt the widely known voting-strategy that Prime Recognition has used successfully, we wish also to surpass the quality of Prime Recognition's output of the ECCO texts: we will re-run those ECCO texts published between 1721 and 1800 in order to improve their correctness beyond what Prime Recognition has already accomplished. We have the benefit of knowing in general what they accomplished, and we add to that the knowledge we have gained from other successful enterprises, including Google Books.

Second, we will imitate Google Books. In articulating how to elicit the best performance from Optical Character Recognition engines on early modern texts, Jon

---

[2] On the latter: Michael Bryant, Tobias Blanke, Mark Hedges, Richard Palmer, "Open Source Historical OCR: The OCRopodium Project," *Lecture Notes in Computer Science* 6273 (2010): 522-25, DOI: 10.1007/978-3-642-15464-5_72.

[3] We have no access to and do not wish to have access to, Prime Recognition's proprietary voting algorithm since we are developing our own, as discussed below. See also Intellectual Property Section, X.1.

Orwant of Google has said, "training, training, training." The user trains the OCR software using that software's interface by indicating it when it has correctly recognized a letter and when not, giving the engine detailed information about letter shapes. Eventually the software is trained to recognize that letter, and the training data resides in what are called "font libraries." Training is one thing when one confronts a collection of texts that are roughly similar, as they most often are: one trains the OCR engine using sample documents from the set, and it can read the rest. It is another thing when the data set is gigantic and as diverse as possible, as in the case of Google books. In that situation, one typically trains an OCR engine to recognize as many fonts as possible. Such training makes the engine flexible, but it can also blunt its powers of discrimination: if an "I" can be many different shapes, many splotches on a page will look like "I."

We know from work done during tenure of the Mellon Officer's grant we received last year that by training Gamera specifically to read two very similar fonts, Caslon and Baskerville, rather than training it to read as many fonts as possible, we have honed its capacities: it can distinguish between the long-s and f by discriminating the length of their crossing lines:

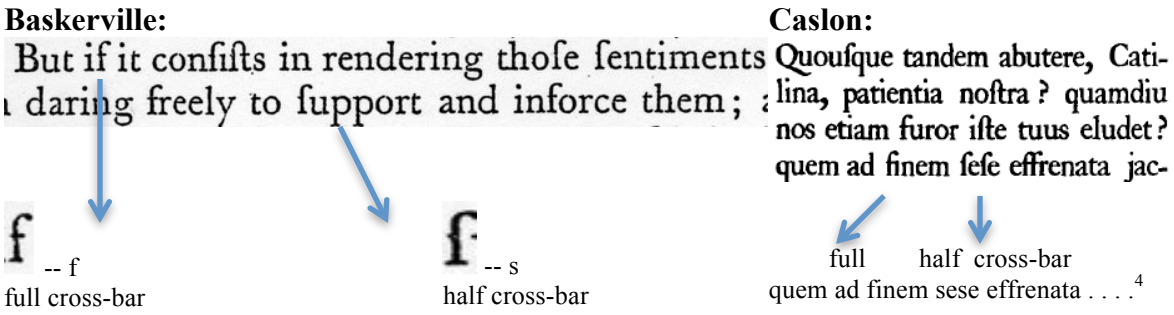


Figure 1: OCR engines capable of Fine Discriminations

As our work in training our OCR engine to "see" this difference shows, instead of training for generalities, it is best to train OCR engines for specific fonts and font families, dividing documents up into subsets according to which font-training sets will best recognize the fine discriminations among "noisy"—that is, blotchy—images.

People often ask: can't we simply re-scan the page images? Yes, of course. However, the cost, time, and effort would be exorbitant. Though some new page scans have been and are currently being made of early modern texts, many rare books rooms are reluctant to put their early books through that ordeal. It is for this reason as well that one of our deliverables, a "need to re-scan" database of unreadable digital images, is so crucial. We know that microfilm lasts only 100 years, and we know that we should have back-up copies of very rare cultural materials: digital copies can preserve these texts, but not if they are only page images; typed text is needed to make those materials findable and usable by future scholars. Special Collections librarians may be feeling unjustifiably

[4] Quo usque tandem abutere, Catilina, patientia nostra? Quam diu etiam furor iste tuus nos eludet? Quem ad finem sese effrenata iactabit audacia? How long, O Catiline, will you abuse our patience? And for how long will that madness of yours mock us? –Cicero, *Catline Orations*.

secure that ProQuest and Gale have preserved copies of their rare books through digitization when in fact some of these images, digitized from microfilm, are inadequate to the task. After we have tried to use those low-quality images to produce machine-typed text, throwing at the problem all the fire-power of new OCR technology and cultural and book history, we—all curators of our Western Cultural Heritage—will know and/or be able to determine which digital surrogates are inadequate to the task and where we need to concentrate our scarce financial resources in the future.

IV. Workplan

Unit A. OCR engine development

The overall aim of Unit A is to optimize three OCR engines in order to improve their ability to accurately read modern texts. The unit will have three goals: first to optimize the engines, next to further enhance their performance through font training, and finally to develop algorithms for assessing their performance.

Goal 1. Optimize OCR engines (IDHMC and Performant)

Description: The goal of this effort will be to optimize OCR engines to customize them for the advanced processing that will be required for these modern texts. The three OCR engines we will use are Tesseract, OCRopus, and Gamera because they are open access and very powerful. Tesseract was released into the Open Source community, and its development has been spearheaded by Ranjith Unnikrishnan and Ray Smith of Google.[5] Gamera is an open-access OCR Toolkit published by Johns Hopkins University. It was originally developed by Ichiro Fujinaga of McGill University for Musical Character Recognition, and thus is specifically designed to make fine discriminations of shape against noisy backgrounds.[6] OCRopus was developed by T. M. Breuel. It is a modular system: early versions of it can use the Tesseract engine as a plugin, but, from the .4 release onward, OCRopus has its own engine which produces results that differ from Tesseract's. It too is freely available.[7]

Initial input for optimal image settings will be modified using ImageMagick. Next Tesseract's line segmentation procedure will be integrated into the Gamera toolkit. The eXtensible Stylesheet Language Transformation (XSLT) will be modified to create eXtensible Markup Language (XML) outputs that are subsequently required by Gale and ProQuest as well as create the Text Encoding Initiative's most basic standard TEI

---

[5] Originally developed by Ray Smith, Tesseract is now being managed at Google by Ranjith Unnikrishnan with whom we are in close contact—he tested Tesseract's line segmentation routines on our documents for us (see Appendix, pp. 85). Tesseract is available on Google Code: http://code.google.com/p/tesseract-ocr/. See Ray Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition, 23-26 September 2007, IDCAR 2007 (p. 630).

[6] http://gamera.sourceforge.net/doc/html/writing_toolkits.html. Gamera is a toolkit, and you can see how we have configured that toolkit in the Appendix, pp. 85).

[7] T. M. Breuel, "The OCRopus Open Source OCR System," Proceedings SPIE / IS&T 6815 (2008), available http://pubs.iupr.com/; the engine itself is available on Google Code (http://code.google.com/p/ocropus/).

Analytics (TEI-A).[8] Finally we will use whitespace XSLT, xslt scripts that determine document features such as paragraph and page breaks based upon line coordinates, repeated words (catch-phrases), signatures, and page numbers, to mark paragraphs and pages in the document results.

Work already done: OCRopus has been improved for 18th-century texts by the Visualizing English Print Project, Tesseract by Google, and Gamera by Mandell. We have tested Tesseract's line segmentation algorithm on page images that Gamera could not read, and it was able to find the lines (see Appendix, pp. 86). Installing Tesseract's line segmentation algorithm will make the engine function as well as it does in the example given below (see the next Goal, Fonts). During that earlier Officer's grant, XSLTs were created for producing TEI-A out of Gamera's xml output and they can be tweaked to work on the xml output of Tesseract and OCRopus, which are in the same hOCR format. Matt Christy of the IDHMC has begun development of the paragraph-identifying XSLTs based upon the work performed by Katrina Fenelon.[9]

Technical Details: The Gamera toolkit is written in Python. It has been installed on the Brazos HPC, plugins installed, and routines have been created to run batches of documents through it in C++ by David Woods's team at Miami University (see Appendix pp. 98-102) whose job will be taken over by Performant Software. Tesseract is written in C that has mostly been converted to C++, though that conversion is not complete, with Python training libraries. OCRopus is written in Python and C++. OCRopus and Tesseract can be run command line on Linux, which is what we have installed on the HPC. ImageMagick is an open-access image modification tool that can be installed on Linux.[10] The outpus of Tesseract and OCRopus are in hOCR, an xhtml format, so that XSLTs can be used to adjust their format (the IDHMC is an XSLT shop).

Goal 2. Fonts (IDHMC and Cushing)

Description: The goal here is to train OCR engines to recognize letters that were printed in very specific fonts. At first, during the month of October 2012, training libraries will be created using images from Texas A&M's Digital Donne: The Online Variorum.[11] We will have more images from documents in the Donne Collection[12] as well as from book history collections as the grant work progresses, but at first, we will

---

[8] TEI-A was developed for the Monk project and has specific capacities for marking up words: http://monkproject.org/MONK.wiki/Abbot%20and%20TEI-Analytics%20texts.html. In an email exchange dated May 3, 2012, Mandell had a lengthy exchange with James Cummings of the Oxford Text Archive about using TEI-A. That discussion will continue with James and others during the TEI Members meeting to be held at Texas A&M University, November 10-12, 2012. The theme of the conference is how to use TEI in crowd-sourced tools and at scale, as we intend to do during grant tenure and subsequently.

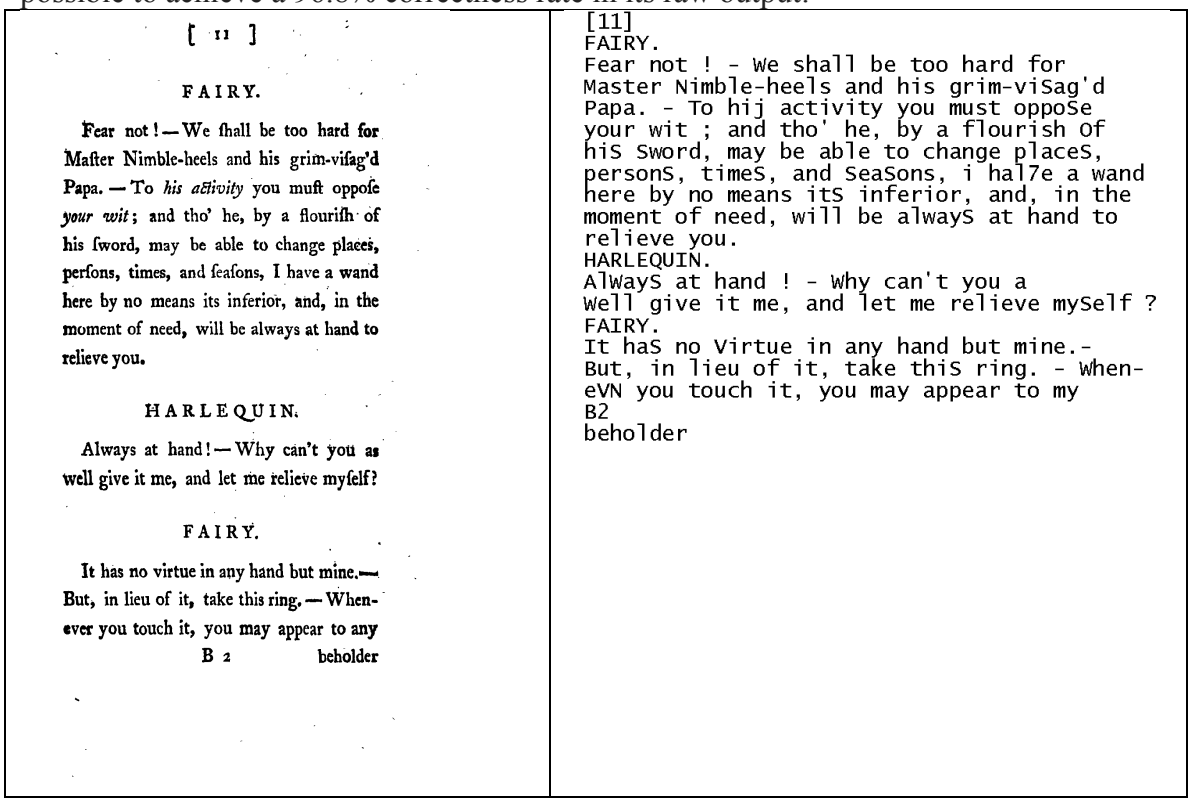
[9] Katrina Fenelon, "Exploring the Viability of Semi-Automated Document Markup," final practicum paper, sent to Mandell 4/19/10.

[10] http://www.imagemagick.org/script/index.php.

[11] http://digitaldonne.tamu.edu/.

[12] Smith, Steven Escar, and Gary Stringer, *The Texas A&M John Donne Collection, Cushing Memorial Library and Archives*, Texas A&M Libraries, 2006 (Exhibition Catalog, reproduced in the Appendix, pp. 133).

only focus on texts using three different fonts: Todd Samuelson will decide which three fonts to train the engines in and to test after looking closely at the texts for which we have transcriptions and determining which would be the best to try. The trained OCR engine Gamera will then be run on then images in EEBO for which we have transcriptions that were printed in these three fonts. This is a test, a way to make sure that font training works as well on EEBO page images as it has on ECCO. Described above in relation to the f/s difference above, Gamera's training in Caslon and Baskerville has made it possible to achieve a 96.8% correctness rate in its raw output:



[11]

FAIRY.

Fear not!—We ſhall be too hard for Maſter Nimble-heels and his grim-viſag'd Papa.—To *his activity* you muſt oppoſe *your wit*; and tho' he, by a flouriſh of his ſword, may be able to change places, perſons, times, and ſeaſons, I have a wand here by no means its inferior, and, in the moment of need, will be always at hand to relieve you.

HARLEQUIN.

Always at hand!—Why can't you as well give it me, and let me relieve myſelf?

FAIRY.

It has no virtue in any hand but mine.—But, in lieu of it, take this ring.—Whenever you touch it, you may appear to any

B 2 beholder

[11]
FAIRY.
Fear not ! - We shall be too hard for
Master Nimble-heels and his grim-viSag'd
Papa. - To hij activity you must oppoSe
your wit ; and tho' he, by a flourish of
hiS Sword, may be able to change placeS,
personS, timeS, and SeaSons, i hal7e a wand
here by no means itS inferior, and, in the
moment of need, will be alwayS at hand to
relieve you.
HARLEQUIN.
AlWayS at hand ! - Why can't you a
Well give it me, and let me relieve mySelf ?
FAIRY.
It haS no Virtue in any hand but mine.-
But, in lieu of it, take thiS ring. - When-
eVN you touch it, you may appear to my
B2
beholder

Figure 2: Gamera's OCR Results after Font Training

CHECKPOINT 1: We know that engines run better with training: that has been proven by all OCR providers. In this checkpoint, we need to test an additional hypothesis. Our hypothesis is that, when specific rather than general letter shapes are sought, the engine will be able to distinguish letters from noise more easily. Checkpoint 1 does not require running a huge amount of documents: we need only run 10 to 20 EEBO documents for which we have transcriptions through Gamera after training it to read the typefaces in which each was printed. Why? Because we will be testing the trained set against itself: we will run the same set again through the engine, not pairing document with font library and batching the run that way, but instead all the font training that we have (Baskerville, Caslon, and the three new fonts) all operating in the engine at the same time. Comparing the two sets (specialized vs. general training) is all we need to do in order to determine unequivocally that training OCR engines to read specific fonts improves the quality of its textual output. Ten to twenty documents or 1500 to 3000 pages is only a small selection of the EEBO corpus, but we will be hand selecting them to

make sure that we test images with a range of different problems (bleed-through, lightness, blotchiness, and font variety).

Passing Checkpoint 1 insures that the money required to create the Font Importation History database as well as font-identification tools is well worth spending. If it is not worth spending, that is, if the engine runs as well with all font training libraries in it at once as it does with the documents separated by font and run using specific libraries, we will no longer employ the post-doc Jacob Heil in the project but will have him work on other IDHMC projects (his salary will be paid by the IDHMC in either case). Todd Samuelson will still help us collect font types and train libraries, but we won't need the database that Dr. Heil was going to build because we will not need to sort documents by the fonts they used. **Date: November 1, 2013**

After passing that test constituting Checkpoint 1, we will continuously train the three OCR engines to read a diverse set of typefaces using images of printed texts in the Texas A&M Donne Collection that have been and will be scanned. Research will be conducted on the importation of fonts into England before 1720, tracking what printing houses owned which typefaces during what particular times and fed into a database containing the English Short Title Catalog (ESTC) metadata for the EEBO and ECCO collections, connecting texts from these collections to specific fonts whenever possible.

A History of Font Importation Database will be created. After consulting with James Moseley here in College Station about how to best use research time, Dr. Samuelson will travel with Postdoctoral Candidate Jacob Heil to two major font repositories:

- The special collections of St. Bride Library and Institute in London, UK, contain remarkable holdings of physical material and archives related to England's typographical past, including the types (with punches and matrices) from such notable institutions as the Caslon and Figgins Foundries, the Chiswick Press, and the Oxford University Press (containing typographical material beginning in the 17th century, including the famous "Fell" types). Another great resource is its collection of type specimens; given its 10,000 examples, the Library remains the world's largest repository of British specimens.
- The Plantin-Moretus Museum in Antwerp, Belgium, http://museum.antwerpen.be/plantin_moretus/index_eng.html, was founded in the 16th century by a member of the notable Plantin family of printers and has remained intact since. It contains two of the world's oldest known printing presses (pre-1610), but more importantly, it is a hoard of typographical material, still holding the original punches and matrices produced for the press (sometimes as proprietary designs, and often by the greatest type designers and punchcutters of the day, such as Claude Garamond). This trove has only recently been evaluated and catalogued, with Hendrik Vervliet's *French Renaissance Printing Types: A Conspectus* appearing only in 2010. This reference provides specimen facsimiles of over four hundred faces, which, it argues, provide the basis for the majority of today's Western text types, whether Roman, Italic, Greek, or Hebrew.

The font database that will be built will enable automatically separating documents according to which font libraries ought to be installed in the engines to run those documents, for best results, and we will use information in it as it becomes available in document sorting and batching, the pre-processing that will go on along with ImageMagick rectifications. As the Cushing Library team performs this research very clear images of fonts will be collected both for creating training libraries, a process that will go on continuously during grant tenure, and for creating images to be used for font-identification in the Cobre tool (see Tools, B.5.a, below).

Work Already Done: As mentioned, digital images of some of Donne's work are currently available via the Digital Donne. The Donne Collection of Cushing Library includes not only 21 editions of Donne's works published during the 17th century, but also 71 books that were in Donne's library, most of them published late-16th and early 17th century, in Venice, Rome, Cologne, Paris, Lyons, Leyden, Antwerp, and Frankfurt, and more (see Appendix p. 130). That collection provides a good starting set for finding font images and training data.

Technical Details: Though we trained Gamera via the toolkit plugin, Aletheia Desktop—not the tool we are developing, but its current, desktop version—was used to create another Caslon and Baskerville font training set that we will plug into Tesseract: XSLT's were written that transformed the PAGE output of Aletheia Desktop into Tesseract box files. In doing so, a routine for staff-training of font libraries for our OCR engines was established, a routine that involves using Aletheia Desktop in exactly the same way as the IMPACT group has used it—the tool was originally built for that purpose as well as for creating a test set to evaluate OCR engines.[13]

Goal 3. Testing (Manmatha)

Description: We will create a way of checking OCR outputs by comparing each result to hand-typed transcripts made from the same page images. The Text Creation Partnership (TCP) has double-keyed 47,000 texts and is allowing us to use their data (Appendix, pp. 7). R. Manmatha will create tools and algorithms that allow comparing OCR outputs from all three OCR engines (Tesseract, Gamera, and OCRopus) to their typed counterparts. He will adapt his current algorithm in order to process early modern texts known to have high OCR error rates, which thus far Manmatha has not tested (see Challenges, just below). Manmatha will calibrate the algorithm using a small set of marked up page images to be provided by Mandell in the first month of grant tenure, October 2012. Manmatha's testing process will be made available via an API so that other can use it as well. Creating a test set will allow us—or indeed anyone who wishes to OCR early modern texts—to test the effects of adjustments to OCR engines (in our case, adding line segmentation routines to Gamera, for instance) as well as any training they might do (in our case, font training).

---

[13] Apostolos Antonacopoulos, "IMPACT Work Package leader from PRImA, University of Salford, Introduces the work on image enhancement, evaluation, and evaluation datasets," IMPACT youTube Channel, http://www.youtube.com/watch?v=Uzp1Rt0oH4k&list=UUrxMNhuEqIn_tzkNX-J3cpg&index=9&feature=plcp, accessed 28 May 2012.

Work Already Done: To find out how correctly your OCR engine transcribes a document, you need to compare it character-by-character to a version of the document that has been correctly typed. If you have a line of text that looks like this

"tda cloq valdcd acnoff dhc moom "

-- and what it means is

"the dog walked across the room,"

you can correct it if you know what the second sentence is what it is supposed to say. What if the document is comprised of 200,000 sentences, all of them looking like that--how would you ever know which word any given clump of letters is *supposed* to be transcribing? Working on the Proteus Project, R. Manmatha has written an algorithm that performs much more quickly than algorithms that compare two texts line-by-line because Manmatha's algorithm takes less time and less computing power than it would to go through page images and find each line on a page. This algorithm does NOT identify the lines, nor count words from the beginning, in order to figure out that "valdcd" in the OCR results is a (bad) transcription of "walked" in the correctly typed text.

Technical Details: The algorithm that Manmatha has developed relies on the principle that by Zipf's law 50% of the vocabulary words in any book appear only once. By using a recursive segmentation technique based on this principle, the words in a book's OCR results can be compared to its typed counterpart in just one second. Instead of dividing the documents into corresponding lines, the process breaks the text into segments divided between unique words. A "leaf" segment is about 200 words long. The leaf segments can be rapidly compared using dynamic programming; they are much more versatile than lines created by page images.

Challenges: Manmatha's algorithm works by finding key points in each document so that each version can be compared with the other by matching using unique words, such as the statistically improbably phrase "implacable family" in *Clarissa*.[14] The challenge with early modern texts is that the original OCR could be too flawed to make that possible. As can be seen in Figure 8 below, the number of unique occurrences of words is very high in texts that have been poorly OCR'd. (Figure 8 shows 430 + 4,146 = 4,576 in the first 100 pages, beginning with the alleged word "!hot.") Such OCR noise will severely test the functionality of Manmatha's algorithm, a test he is very excited to perform. We have contracted the IMPACT team as an agency to consult and oversee our progress. IMPACT offers a contingency plan. Should Manmatha's work fail to produce a test set, we can create a test-set in the same way that IMPACT has created one: they too have a faster method, viz., using Aletheia's semi-automatic identification of regions. The Texas A&M Co-PIs believe strongly that Manmatha's algorithm will work.

CHECKPOINT 2: in January 2013, we will evaluate Manmatha's progress (see "Reporting," below) and decide whether semi-automated ground truth is needed. Manmatha's test set will not be needed until we have run all 151,200 ECCO texts

[14] This is very like Amazon's "SIPs," or Statistically Improbably Phrases (http://www.amazon.com/gp/search-inside/sipshelp.html). The example from Clarissa comes from http://www.amazon.com/Clarissa-History-Young-Broadview-Editions/dp/1551114755/ref=sr_1_2?s=books&ie=UTF8&qid=1338437985&sr=1-2.

published after 1721 (see Overview, above), primarily in Caslon and Baskerville, which we will begin to do in April 2013. Since running those texts will take up to six months' time, we have some leeway as to when the test set can be delivered. **Date: January 31, 2013**

Goal 4. OCR'ing EEBO and ECCO page images (Performant, IMPACT, IDHMC)

Description: After testing the OCR engines to make certain that lines of text are being found by Gamera and that font training helps the engines produce >95% corrected raw OCR output, we will run 80,000 EEBO documents and 180,000 ECCO documents (all the documents that have not been keyed) through the three OCR engines. We will begin 4/13/12 and continue through the end of grant tenure, 10/1/14. To make this process easy and continual, even while the engines are being trained for new fonts, the Taverna Workflow system will be set up. We will use this system ourselves, embedding our routines, tools, and libraries in it, and will make it available to others. The engines will be set up to time out and move onto other page images when the images in a document take longer than 10 seconds to transcribe. These page images will be flagged, "not yet readable," and the documents containing more than one or two of them per 100 pages placed into a folder for error evaluation (see Unit B below, goal 6, below).

Work Already Done: the IMPACT group has collaborated with the MyGrid Project[15] in order to develop a workflow system for OCR'ing documents, Taverna. They have refined this system over the last 2-3 years.[16]

Technical Details: We will batch documents accordingly: all ECCO texts after 1721; all texts for which we have font training and know to be in specific typefaces based on the Font Importation History Database, at first manually, and later automatically; according to error evaluation metrics (see below). The process of running documents through the OCR engines will be parallelized, limiting the amount of time that each engine is allowed to spend on a page to approximately 10 seconds per page.

"The Taverna suite is written in Java and includes the Taverna Engine (used for enacting workflows) that powers both the Taverna Workbench (the desktop client application) and the Taverna Server (which allows remote execution of workflows). Taverna is also available as a Command Line Tool for a quick execution of workflows from a terminal."[17] The Taverna Hackaday contains a paper about an interface for the Taverna Server using a Ruby Gem which we may use to run our Taverna server.[18] Principles can be developed for substituting misspellings in "n-grams," word or word parts of various ("n") letter lengths: often "s" can be substituted for "f," for instance. The machine can

---

[15] http://www.mygrid.org.uk/ ; http://www.taverna.org.uk/

[16] Clemens Neudecker, "IMPACT Interoperability and Evaluation Framework," 12 July 2011, http://impactocr.wordpress.com/2011/07/12/evaluation-framework-and-taverna-with-clemens-neudecker/.

[17] Clemens Neudecker, "The IMPACT Interoperability Framework," 24 October 2011, http://impactocr.wordpress.com/2011/10/24/the-impact-interoperability-framework/, accessed 28 May 2012.

[18] http://impactocr.wordpress.com/2011/11/14/impactmygrid-taverna-hackathon-taverna-server-as-a-portal/

also calculate "Levenshtein" distance, that is, how many "edits" or changes to various letters it takes before a string of characters become a recognizable word. Having developed principles about spelling changes and edit-distance on our corpus in the process of document evaluation, SEASR will give us those n-gram replacement and Levenshtein edit rules in XML form so that we can create dictionary look-ups in Taverna, along with variant spelling lists and dictionaries produced by Martin Mueller and Ted Underwood. The former will be available in RESTful services; the latter are Python-based Gazetteers and wordlists.

MILESTONE 1: We will know that we have successfully achieved the goal of optimizing our OCR engines if, for each 1,000 documents from Gale, 840 are readable, and if, for each 1000 documents from ProQuest, approximately 100 are readable.[19] We hypothesize that the readable documents pass through the OCR engines at the same error rate as the test set that was compared to typed documents, given that the primary faults, unreadable fonts and faulty line segmentation, will cause the OCR engines to take too long at the recognition task, thereby relegating documents to the not-yet-readable list. So we will presume these texts to be 93% correct or higher, the figure we get from our tests (in B.4), before either post-processing or correction in TypeWright. The total number of documents that will be 93% correct, without any other project work besides fine-tuning these engines, is 162,730—approximately 23.7 million pages. Milestone 1 involves confirming our estimates made in Checkpoint 3 (see below) based on the number of pages that have run through the OCR engines: not all 23.7 million pages will have been run, but a substantial number of them will have been: by this milestone, the rates will be confirmed and we will know how much longer until those 23.7 million pages are finished. That we have achieved Milestone 1 will be confirmed in our midterm report because we will be able to give these figures with certainty based on our tests (see "Reporting" below). **Date: September 15, 2013**.

Transition:
While we are optimizing the OCR engines via research in book history and font training, primarily, and running the EEBO and ECCO texts through the OCR engines using the Taverna Workflow, Cushing Library and PRImA Labs will be developing tools that allow for the "crowd" to help the OCR engines do their work.

Unit B. Human machine interaction

In 2010, Martin Mueller held a Mellon-sponsored workshop on crowd-sourcing early modern texts, inviting representatives from the ESTC (Brian Geiger), the TCP (Maria Bonn), and 18thConnect (Laura Mandell), among others. We discussed the Australian Newspaper Digitisation Program: without any advertising, "the crowd" (9000+ people) has collectively corrected 12.5 million lines of text since the launch of the program in March 2007.[20] There were many reasons for people's willingness to work on the newspaper articles, among them the simple interface design for crowd-sourced correction and the user community that was reached, primarily elderly people interested

[19] See the Error Evaluation Document in the Appendix.
[20] http://www.nla.gov.au/ndp/, accessed 1 May 2012.

in genealogy.[21] We discussed creating user communities among scholars. Dr. Mueller's final report, "Scholarly Crowdsourcing of Early Modern Texts," recognizes "*a need for scholars in various sub-disciplines to work with each other and take charge of 'their' data as a continuing professional responsibility*" (item 1.9, p. 4). Dr. Mueller defines "data curation" as maintenance of digital objects that "supports [their] 'discovery and reuse.'" A process of "continuous enrichment" (4) occurs through a feedback loop, he argues, which optimally includes crowd-sourced data correction.[22]

18thConnect responded to Professor Mueller's hypothesis. Modeled on the Networked Infrastructure for Nineteenth-century Electronic Scholarship (NINES), created by Jerome McGann and funded by the Mellon Foundation, 18thConnect aggregates metadata in order to make available in one place the best digital scholarship about eighteenth-century literature and culture. 18thConnect is also a community of scholars, working at the ground level to engage traditional scholars in emerging digital technologies, and to make the voice of scholars heard by librarians and data providers as they create and disseminate digital work. 18thConnect provides a place to put crowd-sourcing tools and a means, through its annual workshops and its collaboration with the American Society for Eighteenth-Century Studies (ASECS), to reach experts in the field for correcting eighteenth-century texts. Similarly, the emerging Renaissance scholarly community called REKn (the Renaissance English Knowledge Project) provides a locale for earlier texts that need human attention.

Both 18thConnect and REKn will host tools, the use of which assists OCR engines in some of their tasks as well as offering a place to correct mis-typed OCR transcriptions.

Goal 5. Crowd-sourcing (Cushing, PRImA, Performant)

Description: We will adapt three web tools in order to get assistance for the OCR engines from "the crowd," both expert and untrained. One will ask experts to identify fonts and editions (Cobre); one will invite ordinary web surfers to draw page layout (Aletheia Web); and the third will ask both experts and citizen scholars (ordinary people with historical and literary interests) to correct text that has been mistyped by the OCR engine when it could not accurately recognize a character (TypeWright).

a) Cobre (pronounced Cobré) is a robust image comparison environment, presenting versions of texts (editions or witnesses) in filmstrip view along side each other and collating these images of different texts while allowing users to adjust the collation:

[21] Rose Holley, "Many Hands Make Light Work," March 2009 http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf

[22] Martin Mueller, "Scholarly Crowdsourcing of Early Modern Texts," Report on a workshop sponsored by the Andrew W. Mellon Foundation and held at Northwestern University, May 11, 2010.

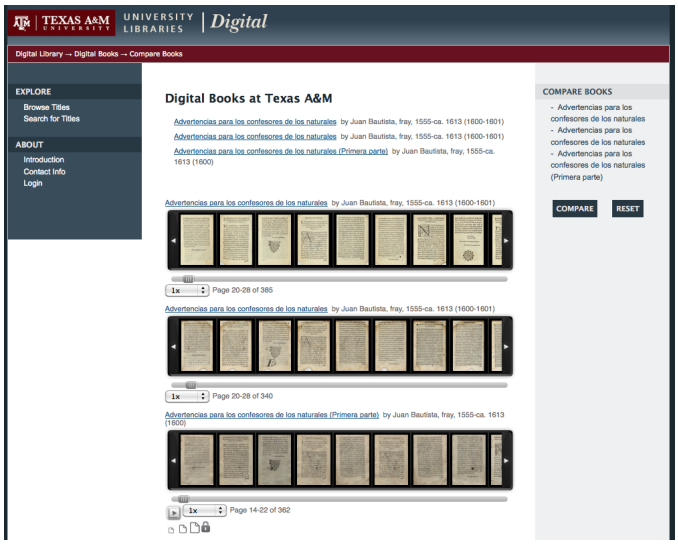


Figure 3: Cobre Filmstrip View

(For more images, please see the Appendix pp. 103-10.) Beginning in October 2012, we will be revising this open-access tool to make it usable for texts in the EEBO collection. After setting up the tool to use it on ECCO and EEBO texts, we will bring book-history expert, James Raven, and distinguished Restoration and 18th-century scholar, Robert Hume, to College Station (February 2013). We will train them to use the tool. Dr. Furuta's team will observe their interactions as part of their usability studies (see C.10 below). The Cobre development team at TAMU libraries will take two five-week periods to work early on in grant tenure and then midway to make revisions based upon feedback from Dr. Furuta's other usability studies (C.10).

Features and capacities to be added during the development cycles:

- Add the possibility for transcription of pages on the Annotations window;
- Make editable specific items of the metadata format on the Book Overview page, adding Dublin Core font categories (see Appendix pp. 111-14).[23]
- Make it easier for people to manually transcribe pages for particular ("local") editions and then share those transcriptions with other editions.

Using the Cobre tool, experts can:

- Determine which font(s) the document was originally printed in edit the metadata;
- Look at multiple copies of editions of the text to replace particular page images that are readable with those that are not, making what Anton duPlessis calls "Frankenbook"—a book that didn't really exist but that resembles what did exist more closely than what we have got. This book is saved separately from all editions and carefully distinguished from real-live documents, its output only used for data-mining and *not* archival purposes, and saved in D-

[23] As we discussed in the development meeting, pictures of typefaces with their names will be made available for consultation; people will be able to access that image/name list from every Book Overview page.

Space under the scholar's name so that others can check this revision and OCR history, visible in "The Repository View":

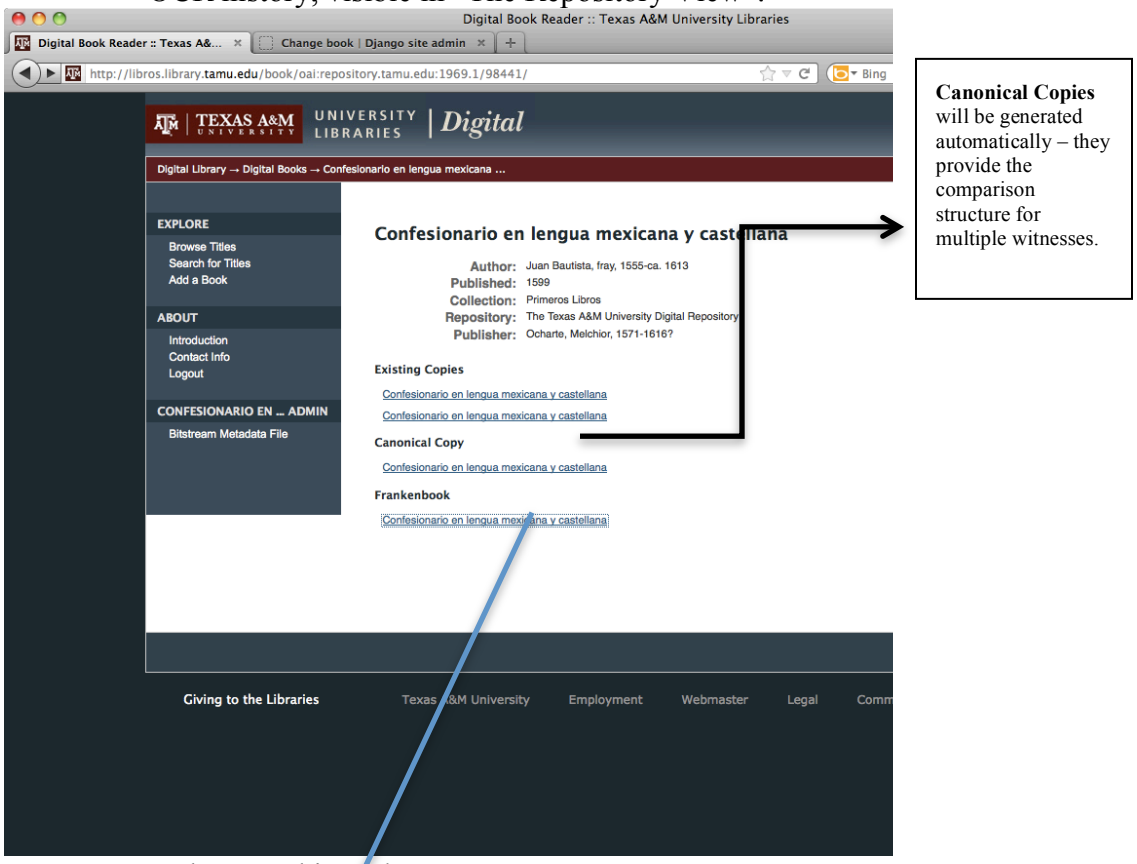


Figure 4: Book Record in Cobre

The Frankenbook, last item in Cobre's book record, will list the expert's name, and all users will be able to view his or her work.

- Type unreadable pages into a Transcription box that will appear in each page Annotation window;
- Correct the metadata. EEBO contains 125,000 texts. 45,000 have been keyed, but they are 45,000 single titles: in other words, according to the metadata, the remaining 80,000 texts are "the same" as what has been keyed. Given what we know about the deceptive uses of title pages described in detail by David Foxon, we are adding a metadata form to the Cobre tool so that scholars looking at multiple "editions" can submit metadata additions and corrections expressing precisely how the texts differ. After conversations with Brian Geiger at the ESTC catalogue, we have decided to submit Marc records for review by the ESTC, and so our metadata forms will output in Marc.

b) **Aletheia**: Beginning in October 2012, PRImA Lab, led by Apostolos Antonacopoulos, has developed Aletheia as a desktop tool that semi-automatically finds layout regions (blocks, lines, words, characters). A C++ program finds them automatically as well as or better than the line segmentation algorithm in Tesseract:

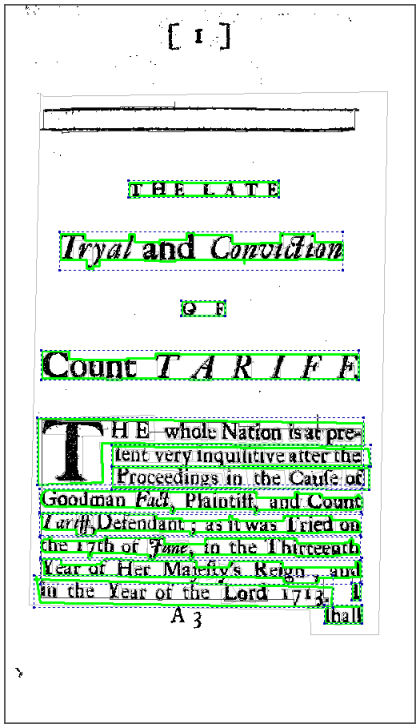


Figure 5: Aletheia Desktop's Automated Line Segmentation

Using the Aletheia desktop tool, a person can correct those segments and then associate typed text with them—hence it is "semi" automatic. We have used desktop Aletheia to train Tesseract on the Caslon font family, and the graduate students on the IDHMC team will use it a) to train the OCR engines in various fonts corresponding to the most used fonts in the Font Flow Database and b) to create a small subset of ground truth needed for each font.

Web Aletheia will invite manual line segmentation. Documents in this workflow will be exported to it when document evaluation techniques reveal that an OCR engine did not effectively find the lines on the page, the basis upon which all recognition is made possible (see Appendix, "Document Evaluation, pp. 72).

c) **TypeWright**: Currently up and running at the 18thConnect web site (http://www.18thConnect.org), TypeWright is a crowd-sourced correction tool that allows users to correct typed text while looking at page images of the text:

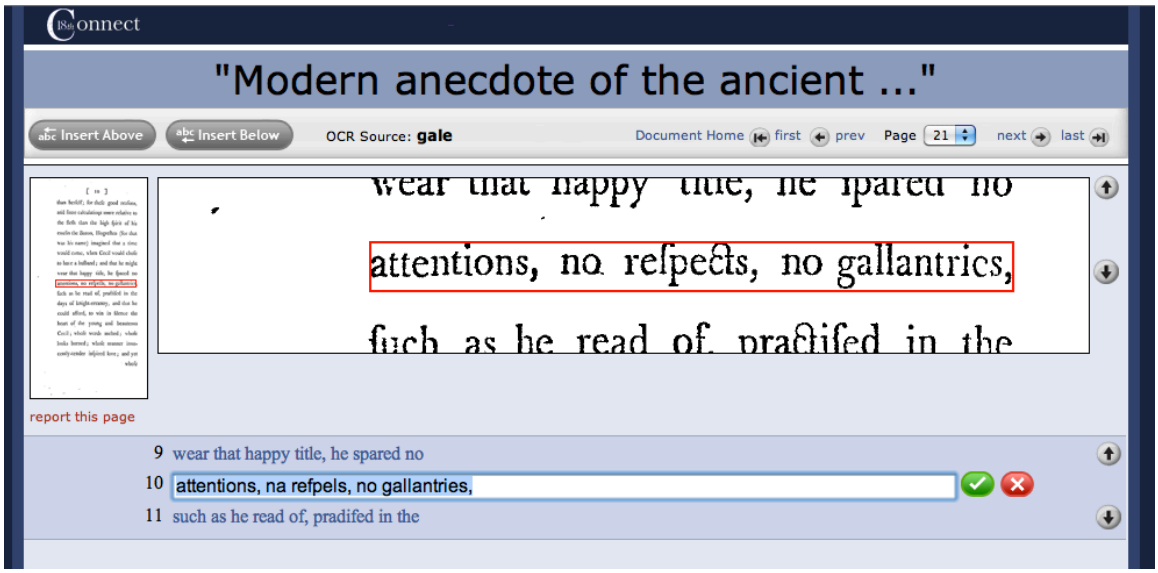


Figure 6: TypeWright

The development and first set of usability studies for TypeWright has been described in the Final Report to the Mellon Foundation for the Officer's Grant "18thConnect and Open Access Full Text."[24] These initial usability studies have given us a development agenda:

- add capacity for adjusting lines;
- add red squiggly underlines (as in Word) to thumbnail for indicating probable errors as indicated by post-processing output;
- add the capacity to italicize and underline.

For all tools, we will conduct wide-ranging usability studies and, as a result, will re-work the interfaces as well as tweak our criteria for what we call "automatic triage," or sending documents to specific tools once they have exited the OCR engine workflow.

Work Already Done: Because we are adapting from existing tools, the work already done has been explained in the "Descriptions" above. We do have some user statistics for TypeWright, as of May 8, 2012:

Number of users who have opened a TypeWright document: 98
Number of TypeWright documents that have been opened for editing: 494[25]
Number of lines with corrections: 53,041
Total number of pages touched: 1863

While TypeWright is currently running, it will be running with good OCR results in it, and running in both 18thConnect and REKn, beginning 1/1/14.

Technical Details:

[24] http://idhmc.tamu.edu/commentpress/final-report/

[25] The number of documents rose by almost 400 after the workshop held at the American Society for Eighteenth-Century Studies Conference in San Antonio, Texas, March 2012.

a) Cobre is built in Python. For the duration of this grant, Cobre will be set up on a Djakota image server using the Django framework. This system will be set up to permit logging in for members of 18thConnect and REKn, organizations which anyone can join. The Early Modern Cobre installation will have its own instance of D-Space running behind it to save the Frankenbooks created by users. After the grant is over, these corrections will either be ported to the Texas A&M repository and/or to REKn (please see the sustainability section below). Though assests stored in D-Space can use a MySql or PostgreSQL database, for our installation at TAMU Libraries, PostgreSQL is used. As users correct the metadata, **Ma**chine-**R**eadable **C**ataloguing (MARC) Records will be generated via the "BMB": the Bitstream Metadata Bitstream alters the D-Space METS records, and those can be crosswalked to and exported in MARC.

b) Aletheia Desktop is built in C++ (http://tools.primaresearch.org:8080/tools/primaweb/tool.php), but PRImA Labs has agreed to port it into whatever code base we wish for creating a web tool. We will ask them to integrate it into the Ruby on Rails Collex interface and to leave out the automated page-segmentation algorithm which we do not want to implement in our web version. Why? Aletheia Desktop is used to create ground truth, to create page segmentation from scratch. But web Aletheia will not be used to do anything from scratch at all. Its users will be helping the OCR engines, leaving the lines marked as they are where the engines succeeded in identifying them and then adding in boxes around lines, columns, pictures, and printers marks where the engines failed to identify those things properly. Aletheia Web will be used to CORRECT the line segmentation that has already been made by the OCR engines—we are installing Tesseract's line segmentation algorithms in all of them. In addition, Aletheia Desktop has a function to export to PAGE in XML format, which appears to be becoming the standard for OCR output. We will retain that functionality in Aletheia Web as well. The IDHMC will rewrite that exporting function possibly in XSLT, as we have already done with XSLTs, in order to transform PAGE into training inputs for Tesseract box files and will do the same for Gamera inputs. Aletheia will save data in a MySQL database that will be kept in the Institutional Repository after it is used to assist the OCR engines.

c) TypeWright is built in Ruby on Rails, as part of the Collex Interface. The TypeWright project is a web service that keeps the information about all the typewright-enabled documents.[26] Information is saved locally, in a MySQL database, until exported (see "Sustainability" below).

MILESTONE 2: We will have adapted and re-released three crowd-sourced correction tools, making them available open access via GitHub. (Please note that we will be releasing them into the open-source coding community AFTER launching them among users because some of the tweaks to the code will depend upon power-user feedback. This is the same schedule we adopted for TypeWright.) **Date: September 30, 2013**

Goal 6: Document Evaluation (Ricardo Gutierrez-Osuna, SEASR)

Description: Beginning in October 2012 through September 2013, we will work on discerning whether things went well or what specifically went wrong after any given document has run through the OCR engines. Thus, in this grant, by "document

[26] https://github.com/collex/typewright#readme

evaluation," we do not mean checking OCR output against text that has already been typed (see Testing, A.3 above). Instead, we mean evaluating the cause of transcription errors in texts for which there is no ground truth, no typed text that it can be measured against. Obviously humans can look at the output and, comparing it to the page image, determine how well or poorly an OCR engine has performed. But because the goal of this grant is to process massive amounts of textual data in the best way possible before asking humans to check it, we must determine how well an engine has performed on any given document automatically—that is, without human beings having to check each output.

So far in this proposal, "OCR Results" has referred to the textual transcriptions of page images. However, in reality, OCR engines give three kinds of output that can be used to diagnose errors. As can be seen in Figure 7, center, "Diagnostics," the OCR process is broken down into its three stages:

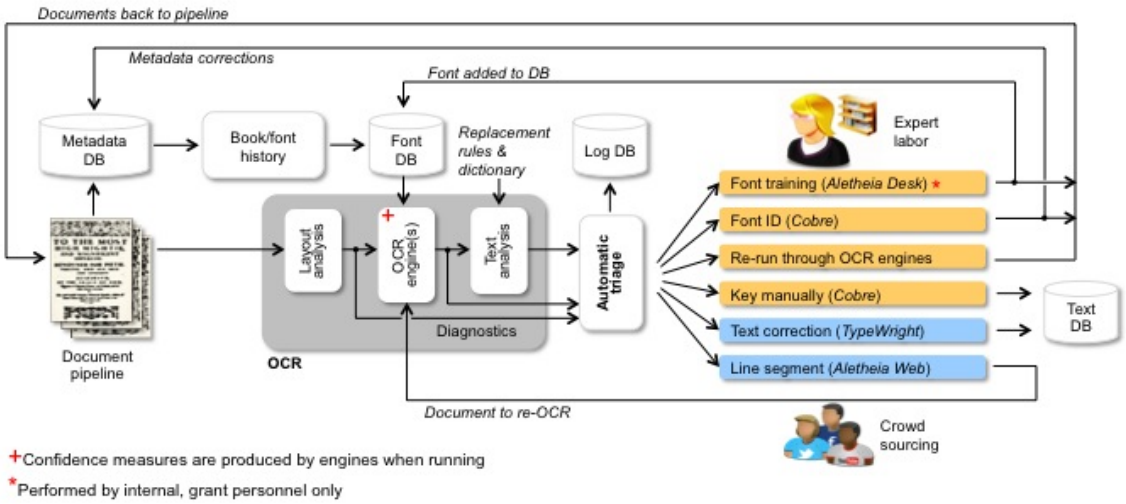


Figure 7: Workflow

Engines produce layout coordinates for all images on a page; confidence measures representing the engine's own certainty as to the correctness of its transcription; and finally textual data. All three of these outputs can be used in order to determine whether an OCR engine has been able to decipher a page image. Most important, they can be used to indicate what has gone wrong when the engine took too long to produce textual data and so timed-out, which is to say, it was forced to move onto the next image.

Document diagnostics involve examining 1) coordinates for line, word, and character; 2) confidence measures;[27] and 3) the text itself. Each one of these aspects as well as comparisons made between the three engines' results will be used to create a document signature—that is, a set of patterns characterizing document sets. Though we can say with certainty that specific results mean specific kinds of OCR transcription failures, ultimately we will build categories of documents based upon these measures based on what features are found empirically to correspond to what kinds of engine failures. Each category will be handled differently (see B.7 Automatic Triage, below):

[27] Each engine as it comes to us out of the box, so to speak, produces degrees of certainty for each character that it identifies.

1) Coordinates. Establishing a base-line upon which the bottom of letters rest, lower -case p's, g's, and y's dropping below the line, is crucial to character recognition.[28] Engines use that base line as a the point of reference around which to coordinate bits of letter shapes that it discerns, and so, in noisy documents with many miscellaneous black dots, finding a base line is the only way to accurately piece together letter bits. Research has found ways to detect when an OCR engine has fallen down in deciphering a page layout, starting with texts for which there is corrected text to compare it to, and moving on to those for which there is none.[29]

When a document fails to find the line, it pieces characters together from all over the place.[30] Therefore, unusual patterns in line and character coordinates usually indicate an engine's failure to find the base line, though it can also indicate that the engine has been presented with an unusual page layout or an unknown and unreadable font. If coordinates indicate that lines and characters have not been reasonably placed together on a page, the document will be evaluated as needing manual line segmentation. If the coordinates for individual characters describe three or fewer sizes of letter for each character (as determined by clustering character sizes in a graph), then line segmentation, we hypothesize, is not the problem: it is either font identification or bad page images.

2) **Confidence Measures**. Rose Holley writes, "OCR contractors often talk about OCR confidence levels and OCR accuracy as if they were the same thing, and in practice confidence levels are often used as a substitute for accuracy because determining true accuracy is not feasible for large volumes of text. Only one contractor to whom we spoke suggested a good solution for gaining an accuracy figure (rather than a character

---

[28] [To correct Mandell's confusion early in the project-planning stage, a confusion manifest in the original Milestones proposed, we need to point out here that Manmatha's work has nothing to do with helping OCR engines learn how to find lines on a page image. Only Ray Smith's work is relevant here.] Ray Smith, "A Simple and Effective Skew Detection Algorithm via Text Row Accumulation," *IEEE* 1995 (p. 1145) [accessed April 25, 2012, via IEEE Xplore, http://ieeexplore.ieee.org.lib-ezproxy.tamu.edu:2048/stamp/stamp.jsp?tp=&arnumber=602124]. The article describes Smith's work on line segmentation, but it has already been implemented. In other words, no one needs to be hired to work on this grant in order to develop line-segmentation techniques. Routines for segmenting lines have been made for Tesseract, which work really well on early modern texts (see Appendix, pp. 85), and those routines can be installed in Gamera by Performant Software.

[29] Dheeraj Mundhra, Anand Mishra, C. V. Jawahar, "Automatic Localization of Page Segmentation Errors" (J-MOCR-AND '11, Beijing, China, 2011); C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments," ICDAR '11, and "Scenario Driven In-Depth-Performance Evaluation of Document Layout Analysis and Methods," ICDAR '11 (2011).

[30] Mandell, Final Report to the Mellon Foundation for the Officer's Grant "18thConnect and Open Access Full Text," 16 December 2011, http://idhmc.tamu.edu/commentpress/final-report/

confidence figure) for libraries with large scale text projects."[31] The solution is to run the OCR engines on page images for which one has ground truth, feed the OCR outputs into a diff engine in order to compare it with the corrected text, and then correlate built-in confidence measures with the actual correctness or incorrectness. We only hypothesize that this figure could predict certain errors or contribute toward creating a document signature; we won't know whether it helps until it is tested.

3) **Text Itself**. Looking at early modern *documents* involves looking at the physical features of these works: their layout, their fonts. Looking at the "text itself" means trying to read the actual words on the page and how well they are typed ("spell-check," which involves with early modern texts, many variant spellings). Work done in completing the "18thConnect and Open Access Full Text" Mellon Officer's Grant shows that improperly segmented page images tend to produce unreadable text of a specific sort. When document layout analysis fails, the text returned consists of long strings of single- and double-character tokens—alleged by the OCR engine to be words—that are in fact not words in the English language, "k" for instance. If that is not the case, but the engines nonetheless timed out in returning textual data, other tests can be used to detect problems. In order to determine whether fonts have been mis-identified or the page images are too flawed, one applies principles of correction based upon known OCR mistakes, taking impossible N-grams (1, 2, or 3 characters, up to n characters, in a row) and transforming them into what they usually represent, a process that has to be carefully adjusted so that the number of errors introduced (there will be some) are maximally smaller than the number of errors corrected.[32] As long as the process is also meticulously documented to allow rolling back to previous OCR versions, nothing is lost from the original recognition work. Afterwards, a moving window of bigrams (testing to find words with impossible character combinations, e.g. "**th**ought," "t**ho**ught," "th**ou**ght," "tho**ug**ht," etc.) will give a sense of whether characters grouped together as words could be possible in a few languages. Next, one performs dictionary lookups and then counts the number of unique words and number of words containing internal punctuation marks (excluding **-** and **'**) per 100 words. One can see these estimates in the debugging or OCR developer's view of TypeWright:

[31] Rose Holley, "How Good Can It Get: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs," *D-Lib Magazine* 1.3/4 (2009): see the section titled "Measuring Accuracy Rates," included in the Appendix, pp. 73.
[32] Rose Holley calls this set of principles a "confusion matrix" ("How Good Can It Get?," note 2 above).

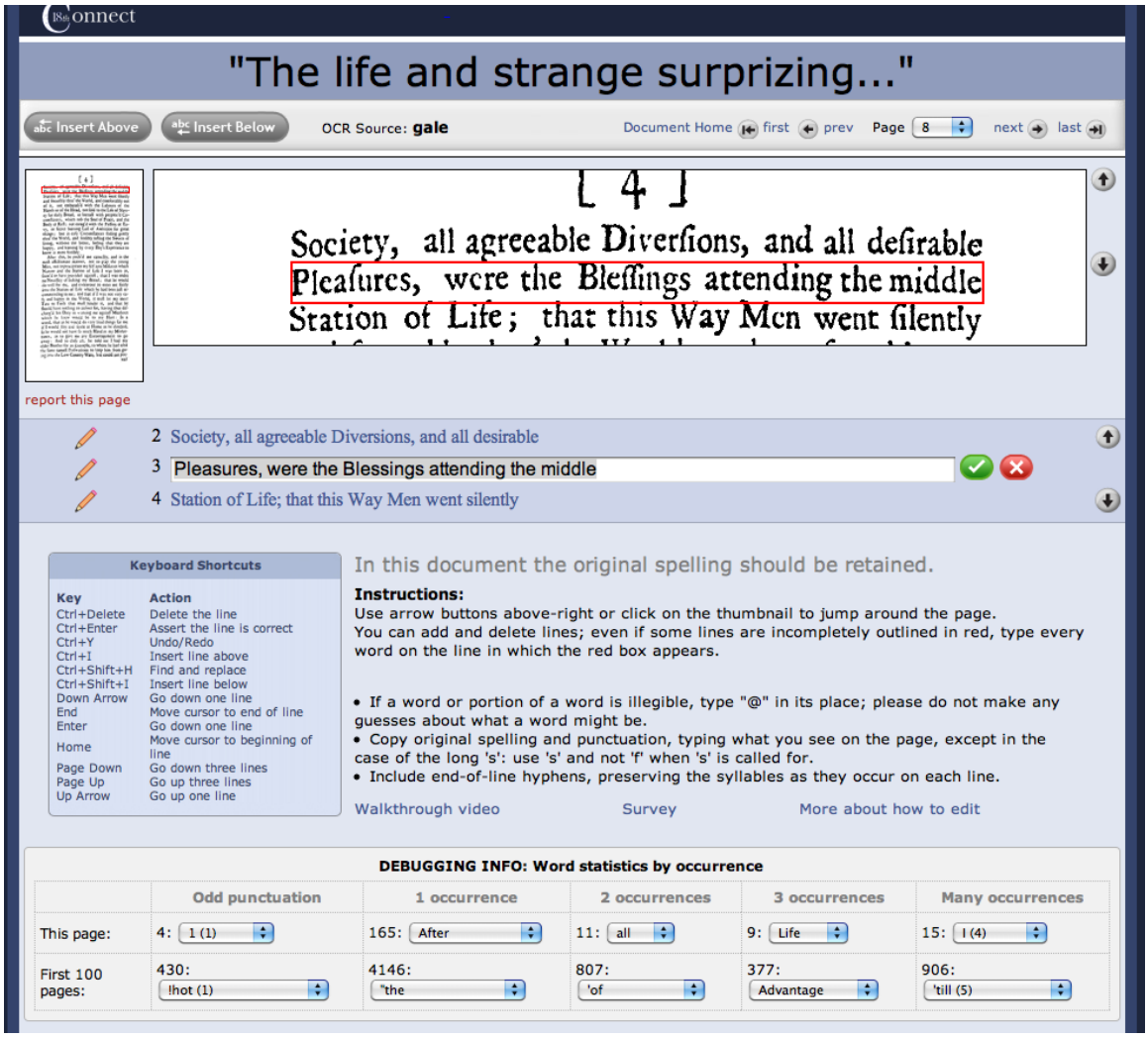


Figure 8: Debugging Information, Developer's View, TypeWright.

Words that recur many times are either correct or consistently misread. There are many legitimate reasons for 1, 2, or 3 occurrences of a word, including Zipf's law (discussed in A.3 above), capitalization, or proximate punctuation, but "odd punctuation" gives us a really good sense of errors. Again, through empirical tests, we will determine the ideal rate per 100 words that indicates a document to be correctable (the dictionary is finding words in the document) or not.

Additionally, automatic dictionary look-ups can be used to find words and word-candidates, that is, character groupings with the lowest Levenshtein distance or number of edits necessary to make them into words findable in a dictionary.[33] The Levenshtein distance measures can also be used to indicate words that we are unsure of and to offer suggestions for those who are correcting in TypeWright when users mouse over a word signaled as uncertain (underlined with red squiggly lines) in the thumbnail. In the case of unique words not found in a dictionary, as in the other cases of error estimation, when the

[33] Mehdi Achour, et. al., "Levenshtein," *PHP Manual*, ed. Philip Olson, 1997-2012, http://php.net/manual/en/function.levenshtein.php, accessed 28 May 2012.

level is over 50% (at some percentile range to be empirically determined) we may not know what has happened, only that something has gone wrong.

And finally, the textual outputs of Tesseract, OCRopus, and Gamera will be compared. Manmatha will adapt his algorithm and process to compare pairs of OCR outputs. Then the two pairs will be combined to create a 3-way alignment. The comparisons can then accept as truth those characters (those letters in the output) concerning which the majority of engines agree. When there is dissonance among the three comparisons, the character in question can be flagged for manual markup, and correction.

Goal 7: Automatic Triage (Ricardo Gutierrez-Osuna)

Based on what went wrong as determined in Goal 6, we will feed the documents to specific types of crowds and the tools needed for accomplishing specific tasks. Some of the documents in our early modern corpus will be worked on by people during grant tenure, the bulk of the documents afterwards. Therefore we will be able to run some documents back through the OCR engines. Though we cannot know how many people will participate, we will be able to tell how valuable is the specific way that we have orchestrated human assistance for the OCR engine—we'll be able to see whether people really can help these machines perform their task more efficiently and effectively, and vice versa (see Figure 7, above). The "Automatic Triage" box symbolizes the sorting process described above. More specifically, layout coordinates, OCR engine confidence measures, and specific kinds of misspellings will constitute diagnostic measures, as explained in B.6 above. These "Diagnostics" will be grouped together into specifiable patterns, and each of these specific document signatures will be handled in different ways. "Automatic" here refers to the fact that the diagnostics were performed mechanically as opposed to having human eyes look at each document's OCR results (layout, confidence measures, and text itself). The actual manual forwarding of the documents to the places they need to be put will be a semiautomatic process, involving personnel running scripts and checking that they worked (see Technical Details below).

- If a document does not "time out" or take longer to read then 10 seconds per page image for most of its pages, we hypothesize that it is correct at the rate which the OCR engines have performed in tests, 93% or more. These documents will be passed to TypeWright. We will know quickly if this assumption is false (see Checkpoint 3, below).
- If a document does time out, the next thing that the automatic triage system must determine is why. The diagnostic measures listed above in Goal 6 will be used to determine the answers:
    - When the OCR engine could not successfully perform because of the font, we will send the document to Cobre and experts: that is, the documents so identified will be sent per a system set up by Performant Software to be loaded into the Django server (on how we will conscript experts to work on these documents, beyond the consultants who have been hired for the grant, please see C.9 and D.11 below);
    - When the problem might be a bad page image, again, the document will be sent to Cobre for expert examination (that this tool is designed for both kinds of work is a direct result of the difficulty of distinguishing between font and image problems automatically);

    - When the OCR engine could not find the lines, we will send the document to Aletheia Web.
- As experts identify fonts in Cobre, we will train the engines in those fonts and re-run the documents through the engines, and afterwards re-submitted to the diagnostics and triage.
- As users mark layout in Aletheia Web, the documents will be sent through the OCR'ing process again with layout information provided, and then re-submitted to diagnostics and triage.

Technical Details: Document flags will be associated with results pertinent to document signatures via the OCR procedures. Using the triage scripts written by Performant, the flagged documents will be loaded into server directories housing "Texts needing Correction" (the current TypeWright server), "Texts needing Expert Attention," and "Texts needing Layout Analysis." (The actual running of the scripts to transfer documents and checking that they are loaded will be done manually by Performant personnel, as they are now in the case of TypeWright Documents.) These documents will be loaded into an Aletheia-Web service, just as TypeWright currently has a service, that will feed the respective tools via URI. The documents marked "Expert Attention" will be transferred to the instance of D-Space running the ARC installation of the Cobre tool, and a directory built for search access using PostgreSQL. This transfer will be overseen by Cushing Library personnel in conjunction with TAMU Libraries Digital Initiatives group.

CHECKPOINT 3: We will examine whether our assumption that documents taking fewer than 10 seconds to run through the engines is correct at the base 93% rate. We should be able to tell

- By the inability of post-processing routines to produce words that can be found in the dictionary;
- By the "report this page" link in TypeWright—if it is used too often, we are not getting relatively correct texts to the crowd-sourced correction tool.

Checkpoint three is the test preparatory to achieving Milestone 1, the topic of our midterm report, to be written in the fourth quarter of grant tenure and submitted September 15, 2013. If we do not pass this checkpoint, we will have to add further evaluation measures to that of time as a preliminary determinant of OCR performance (see Reporting). **Date: April 15, 2013**

Goal 8. Launch Crowd Tools (IDHMC, REKn)

Description: At the Modern Language Association (MLA) in January 2013, the American Society for Eighteenth-Century Studies (ASECS) in March 2013, and the Shakespeare Association of America (SAA) in March 2014, workshops will be held and/or talks given about Cobre, Aletheia Web, and TypeWright. These workshops and papers will advertise as well as teach people how to use the tools and enlist experts in the task.

Work Already Done: Mandell held a two-hour TypeWright Workshop at the annual ASECS Conference in San Antonio, Texas, on March 15, 2012. From that meeting, two groups have decided to join 18thConnect and a rise of approximately 400 was logged in the number of uses of TypeWright. We have scheduled a full-day

workshop for the ASECS 2013 meeting in Cleveland, Ohio, paid for by the IDHMC. Mandell is giving a talk about OCR'ing EEBO and ECCO at the panel of the MLA Division for Restoration and Eighteenth-Century Studies, chaired by Catherine Ingrassia, scheduled for January 2013 in Boston, MA, and will also be involved in the DHCommons pre-conference workshop where she will teach a session on using TypeWright and will advertise Cobre. Meetings with the directors of REKn will take place in June 2012; Ray Siemens has been in touch with many of the Renaissance societies who have offered support for REKn (he discussed it with them at MLA in Los Angeles, 2011). Per an email from Professor Richard Cunningham at Dalhousie University, the Canadian INKE (Implementing New Knowledge Environments) will be financially supporting REKn (see Appendix, p. 8).

MILESTONE 3: We will have achieved accurate document error evaluation and proper triage if the number of page-images labeled "too error-prone to correct" in TypeWright will be 2% or less of the total number of page-image views in the TypeWright Tool. The same will be true for Aletheia Web: those page images for which users indicate that the pages are too damaged to diagram must be 2% or less. This will mean that document evaluation is working with some strays getting through to the wrong tools. **Date: December 31, 2013**

Unit C. OCR Correction

Goal 9. Mechanically Correct the OCR Output (IDHMC)

Description: Repetitive correction activities must be done by machines as much as possible. We will begin correcting texts by machine in October 2013, after a substantial number of texts have been run through the OCR engines (A.4) and the triage system has been put into place (B.7) so that we know which documents are ready to correct using dictionaries, N-gram principles, and gazetteers. Our process will include automated dictionary look-ups and N-gram substitutions based on principles and rules about spellings in a particular language during a specific historic period. "He" can be substituted for "hc," for instance, the latter combination of 2-grams most certainly never used in the English language.[34] These principles will have been developed by SEASR during year one and Document Evaluation (see B.6 above): now these rules will become part of the post-processing correction routines that we will be installing in Taverna. Ted Underwood's Gazetteer along with early modern variant spelling lists that are available (both Underwood and Martin Mueller have created one) will also be installed for post-processing dictionary look-up. Again, from the experimental work with SEASR during year one, we will know what limits to set on edit distances except in specific instances. (For instance, "iii" may always be an "m," for instance, so one would in the case of three consecutive Is set the Levenshtein edit distance at 3, but in other cases, reject the results of 3 edits to a word that make into a dictionary word. One can see the ludicrous results of 3 edits when using a text-messaging or email system, for instance, with "auto-correct"

[34] Xian Tong, David A. Evans, "A Statistical Approach to Automatic Error Correction in Context," Proceedings of the Fourth Workshop on Very Large Corpora WVLC-4 (1996); Karen Kukich, "Techniques for Automatically Correcting Words in Texts," ACM Computing Surveys (1992).

built in.) Manmatha's techniques can also auto-correct when the majority vote favors a specific word. We will tag every word corrected via n-gram substitution principles as well as dictionary and gazetteer look-ups so that the documents sent to TypeWright will highlight them for humans to check. All post-processed documents will go to TypeWright as set up in 18thConnect and REKn.

Technical Description: The rules for n-grams and edit distances ultimately settled upon during SEASR's work at evaluating documents (B.6 above) will be output as xml documents; they can then be incorporated in routines established in Taverna. Ted Underwood's Gazetteer and variant spelling list will be written in Python, so work will have to be done to put them into Taverna which is in Java. Additionally, we will use whatever other resources are available. By the time we are working on this part of our grant, Martin Mueller's early modern word list will most likely be available in the form of Morphadorner 2.0 which will be a RESTful Web service that Performant can implement in Taverna; Morphadorner 2.0 will make use of "the Restlet library to implement the web services. Restlet eases the development of Java-based RESTful services."[35]

CHECKPOINT 4: We will know that we have succeeded at mechanically correcting these early modern texts if the error rates on the test documents can be reduced by 60%. Though that 50% figure has been proposed for modern spell-checking,[36] we will be using early modern dictionaries and gazetteers, which should allow the equivalent. The 50% figure should be improved upon via the voting technology and replacement rules, and we hypothesize that it can be improved by 10%. **Date: March 15, 2014**

Goal 10: Engage Humans in the Correction Process (Furuta, IDHMC)

Setting up crowd assistance and crowd correction makes specifying grant deliverables difficult. To make that aspect more clear: by the end of this two-year grant, 10,000 documents will be hand-corrected by the end of grant tenure, and the crowd-sourced assistance and correction processes will be set up.

Description: Human attention wanes when people are asked to make the same corrections over and over again. Mechanical tasks eliminated, people can do things that the machines cannot. They can validate the N-gram substitutions made by machine – occasionally, "hc" should be "ho" or "ha"—and confirm whether a dictionary or name substitution was successful. People can better select than a computer once it has identified multiple possibilities for one word by understanding the context, and they can decipher words machines cannot, barring damage to the page or image that prohibits seeing the word. But if tasks are too onerous, or too complicated, they will not continue. For instance, it may even be that experts lose their desire to help when confronted with Blackletter Typeface, in which case such documents must be keyed by people who are hired to do it. TypeWright currently frustrates users because they want to sometimes be able to adjust boxes rather than simply delete or insert them, but making TypeWright into a tool where layout can be analyzed would make its interface too complex and

[35] Per an email interchange among Martin Mueller, Nick Laiacona, and Laura Mandell dated May 30, 2012.
[36] OCR Accuracy Rates:
http://primerecognition.com/augprime/clean_data.htm

forbidding—Aletheia will take on that task, offering a simple interface with the number of possible commands limited to a comprehensible number. Beginning in October 2013, Furuta's team will conduct usability studies to determine the most motivating number—the most motivating balance between complexity and simplicity in interface design—as well as how human capacities for attention and interaction ought to affect document triage.

Furuta's team will need to know in each instance "who" comprises "the crowd," information that will be provided to him based upon Mandell's work in recruiting (described in detail in a, b, and c of this unit, below) as well as Google Analytics for those who come to work on various tools without having been recruited. Experts and professors are able to work on correcting these texts to the extent that this work can be published, and we have been able to work out such a possibility. Additionally, experts will be recruited to use the new Cobre tool from among the various special interest groups who care about particular sets of rare books, called "editing groups."

a) **Expert Users**: How will Cobre command expert attention? A queue will be set up in REKn and 18thConnect listing what documents need expert attention, and those documents will be loadable into Cobre via links. (Users will actually be passing out of the Collex environment into Django when they click on a text to load, but they won't know that because page styling will suggest continuity.) Initially we will ask our consultants to work with the documents they find there, pending usability studies, when we will begin to build editing groups. Once special interest groups have been contacted and engaged to become editing groups in 18thConnect and REKn, editors from among them will be appointed who will commission correctors, and we will set up a process of "publishing" this work along with discussions of it by fellow experts in what might be called a "Notes & Queries" section of 18thConnect and REKn. Again, usability studies will suggest the most meaningful terms and icons for signaling the spaces at these sites where one can find and participate in group-editing and discussion work.

Work Already Done: Mandell has been in contact with special interest groups such as the History of Science group[37] and the Defoe Society (please see the letter from Benjamin Pauley, Appendix pp. 3-4) in order to create editing groups of expert users.

b) **Everyman Users**: Because page layout analysis—drawing boxes around columns, lines, and images—is relatively easy, our goal will be to enlist anyone willing to work on this task. Aletheia Web will be set up in REKn and NINES, and it will be usable on mobile devices as well. We will advertise widely to get students working on layout. In the short term, our students at the IDHMC will work on layout analysis, for usability and effectiveness testing.

Work Already Done: Mandell has been in contact with the group at IBM's World Community Grid. They offered to support us in some way, and part of the work for this grant will be figuring out how to leverage their access to users (small tasks on screensavers, of the Captcha sort, is one possibility we discussed).

c) **TypeWright Correctors**: Mandell and co-director of 18thConnect Professor Brian Pasanek presented TypeWright at an ASECS executive council meeting in 2011.

[37] Mandell has been in touch with Wallace Hooper, Science Historian at Wells Library of the Indiana University at Bloomington, since DH2011 concerning possible relationships with ARC (whooper@indiana.edu).

The result has been that 18thConnect has been invited to run two pre-conference workshops at the annual ASECS meeting (see B.8 above). We will be working with ASECS President Laura Brown offer automatic membership in 18thConnect with ASECS membership and then to advertise TypeWright as providing search access to ECCO and EEBO for those who work at non-subscribing institutions (those with lower library budgets) and for independent scholars. The pre-conference workshops emphasize that TypeWright offers scholars the opportunity to create and publish peer-reviewed versions of electronic scholarly editions.

Mandell on behalf of 18thConnect and REKn has worked out contracts with Gale, ProQuest, and the TCP (Appendix, pp. 19, 27, 8) whereby scholars are given any texts that they correct. This contract benefits all member of the Advanced Research Consortium (ARC), an umbrella organization for NINES (http://www.nines.org), 18thConnect (http://www.18thConnect.org), and the forthcoming MESA (http://mesa.performantsoftware.com), and REKn, and sustainer of the catalog that is a web service funneled to these child organizations (http://catalog.performantsoftware.com/). Because ARC offers workshops in how to create electronic scholarly editions while 18thConnect and REKn offer peer-review of those editions, to insure that they are library-quality in technological ways as well as the highest quality scholarship, professors can get tenure and promotion as well as merit raises for correcting texts in TypeWright. Because of the unique status of NINES, 18thConnect, MESA, and REKn, and because of the data aggregation model[38] devised by Jerome McGann and Bethany Nowviskie, it has been possible to broker a give-and-take relationship with Gale and ProQuest through ARC—these companies have been unwilling and unable to work with any other kind of organization thus far. This social as well as technological work of to be done for this grant represents the first attempt to get people working together to preserve our cultural heritage, and it is a major one. No one has been able to get access to these images to improve quality until now.

Goal 11. Saving the Data (IDHMC)

Description: All the web services (Aletheia, TypeWright) will be hosted in the cloud, but, beginning in December 2012, all the previous, current, and future work done by 18thConnect and REKn users will be ingested into the ARC data-storage and Solr Indexer unit at Texas A&M. All texts corrected in TypeWright by users deemed reliable will be forwarded to Gale, ProQuest, and the Text Creation Partnership for integration into the instances of the data set that is preserved and shared by libraries. These texts will also be given to the people who corrected them, and those people will be encouraged by ARC to create library-quality electronic editions. All metadata corrections will be exported to the English Short Title Catalog for review. All correction histories will be saved as a data set in the Texas A&M Institutional Repository for use by computer scientists studying human machine interaction and crowd-sourced correction histories. All revisions made in Cobre will be saved by author (corrector) in the Texas A&M D-Space.

---

[38] Jerome McGann, Bethany Nowviskie, "NINES: A Federated Model for Integrating Digital Scholarship," whitepaper, September 2005, http://www.nines.org/about/wp-content/uploads/2011/12/9swhitepaper.pdf, accessed 24 May 2012.

When the page images and/or original printed pages are too flawed to allow for reading, we will put the metadata about those documents into a database called "Inventory of Documents Requiring Rescanning / Keying," which will ideally serve as a reference tool for rare-books librarians, proprietary companies, and the TCP in order to determine which texts in their collections are most important to scan and which need to be keyed. This database will be accessible through the Texas A&M Institutional Repository along with the other databases (see D.12 below).

Technical Description:

a) the ARC data-storage and Solr Indexer Unit that will be established during grant tenure (see Budget Equipment below) will store the OCR process and crowd-sourced correction results in various states and will house ARC's Solr installation. Solr will index plain text files that are the result of this grant (at 97-99.9% correct) by page image of all corrected documents, a change to the way the ARC catalog works that will be installed by Performant Software (see Statement of Work, Appendix p. 35, item 3).

b) The Texas A&M Digital Repository (http://repository.tamu.edu) collects, records, provides access to, and archives the research and scholarship of Texas A&M University. It contains digital works that reflect the intellectual and service environment of the campus. The Digital Repository provides increased access to the products of the University's research and scholarship endeavors, fosters the preservation of these digital works for future generations, promotes increasingly rapid advances in scholarly communication, and helps deepen community understanding of the value of higher education. The Texas A&M Digital Repository is a service of the University Libraries and is managed by the Office of Digital Services & Scholarly Communication. The Digital Repository currently runs on DSpace, an established, open source repository software package with more than 1000 installations worldwide.

MILESTONE 4: After two years of grant tenure, 162,730 documents will move from an average accuracy rate of 93% to 97% correct before being sent to TypeWright. TypeWright corrected texts, 10,000 by the end of grant tenure, and continuing, will be as close to 100% as humans can get, or 99.9% correct. These documents will be indexed in the ARC catalog and so fully searchable by anyone who comes to NINES, 18thConnect, REKn, or MESA. **Date: September 30, 2014**

Unit D. Dissemination

Goal 12. Release of Tools + OCR Workflow and Databases (IDHMC)

We will release the tools and the Taverna workflow (represented in Figure 7) on Github and put plans into place with other organizations (Bamboo Corpora Space, HathiTrust Research Center, JISC, and IMPACT) for setting up installations of the Crowd-Assistance and -Correction tools: Mandell has been and remains in contact with Bamboo about putting TypeWright into TextShop and with JISC about their desire to set up instances of TypeWright. Mandell has been in contact with John Wilkin (email 2/7/12) and Marshall Scott Poole, co-director of the HathiTrust Research Center (8/6/11),

and will discuss how we might make use of the process established for ECCO / ProQuest for HathiTrust. Finally, we will release the workflow tools in the IMPACT Centre of Competence (http://www.digitisation.nl).

The databases that we create, the Font History Database, and the Re-Imaging of Early Modern Texts Database, will be put up in the IDHMC web space on our Virtual Machines hosted by Guy Almes's Brazos HPC group.

Goal 13. Future Planning (IDHMC)

We will formulate a plan for supporting ARC so that it can continuously support REKn and 18thConnect (a plan already in the works), thereby releasing them to focus on text correction, including and especially recruiting scholars to use the new tools. We will also formulate a plan for recording, monitoring, and pooling corrections that are made continuously in the US, UK, and Canada, as well as anywhere else that we can install the tools.

Goal 14: Publish Results (IDHMC, Gutierrez-Osuna, Furuta)

To make the databases we create permanently available and findable via web and library search engines, we will follow the practice established by John Kunze of the California Digital Libraries of posting "data papers" in the Texas A&M Institutional Repository. The IDHMC has a beta site up now ready to take these papers that will explain and point to the data that others may use: https://dspacepre1.library.tamu.edu/handle/1969.1Labs/167153. Drs. Mueller and Mandell will submit to CLIR (the Council on Library and Information Resources) a report upon OCR'ing Early Modern Texts, if they are interested in reviewing it for publication. That report can reference the DOIs assigned to these data papers, the papers themselves serving as "wrappers" for the databases. Finally, we will submit for publication in Computer Science journals the results of our work on maximizing human-machine interaction.

MILESTONE 5: We will have had an impact worldwide if early modern repositories and collectors use our training libraries, our font-importation-history database, and our Taverna workflow, in order to OCR their collections, both large and small.[39] We will have had an impact if our font history database and our "must re-scan" database are used by libraries and museums in the planning stages of their digitization programs. And finally, we will have had an impact if editing groups and all kinds of users are correcting texts at REKn and 18thConnect using our Crowd-sourced correction tools. We will use Google Analytics to record database usage, will record downloads, and will ask users to contact us so that we can delineate our impact in our final report to the Mellon Foundation. **Date: results tracked September 30, 2014 to December 1, 2014**

[39] Here we are thinking in particular about the "Hidden Collections" program sponsored by CLIR (http://www.clir.org/hiddencollections): they can refer libraries, museums, etc., to our resources.

V. Timeline

Each task will be undertaken by particular groups at Texas A&M as well as subcontractors, indicated according to this key:

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)

The numbers in the following Timeline correspond to the numbering of units and goals in IV. Workplan, above.

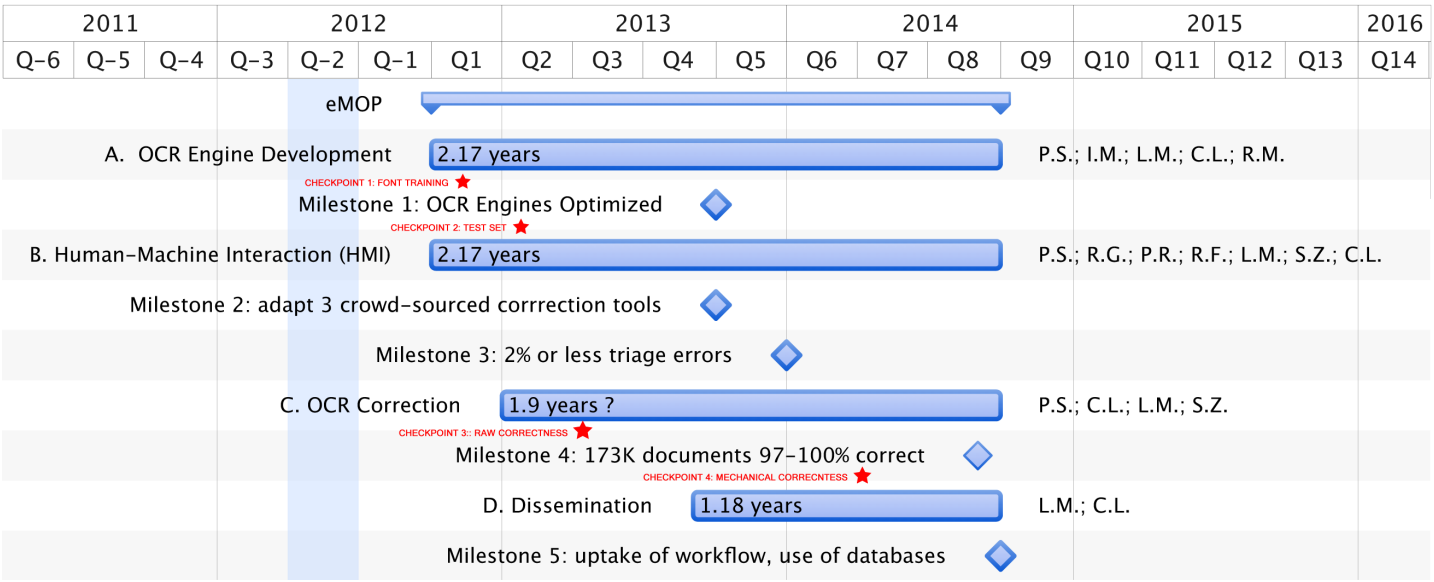
Unit	Goal	Tasks	Q1 10/12–12/12	Q2 1/13–3/13	Q3 4/13–6/13	Q4 7/13–9/13	Q5 10/13-12/13	Q6 1/14-3/14	Q7 4/14-6/14	Q8 7/14-9/14
A. OCR Engine Development	1. Engines	Optimize what goes in: find optimal image settings using ImageMagick	**L.M.**							
		Optimize what happens inside: put Tesseract's line segmentation 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	Set up automated triage system: font mis-id and unknown go to Cobre; layout indeterminacy goes to			**P.S.**					

	Output with Human Assistance (Optimize HMI)	Aletheia Web								
		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**					
	Goal	**Tasks**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**
	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)								
		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. Dis-semination	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 intervene-tion 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										

Here follows a summary of the milestones and checkpoints by unit:



VI. Anticipated outcomes and Benefits

With the two-years of support from the Mellon Foundation, OCR engines will be trained to read digital page images of documents printed before 1720. The outcome differs from what anyone has done so far: we will bring in book historians to help us isolate documents into subsets based upon the typeface used to print them and then train the OCR engines to read those fonts, specifically. Moreover, we will be able to use the typed EEBO texts in order to test all our strategies for improving the OCR (font training)

and to create dictionaries, variant spelling lists, automated correction principles, and gazetteers (drawing names of people and places out of the data set via named-entity extraction). Dictionaries, variant spelling lists, and gazetteers of place and person names, built both by us and by others specifically for the early modern era, can be used for correcting OCR'd text. To abet this process of bringing **Cultural and Book History** to bear on the OCR problem—an epiphenomena of which will be book-history, early modern name and word-variant databases—the TCP has generously offered to let us use their typed texts. Therefore, we should be able to produce machine-readable transcripts of a subset of EEBO texts for which there are currently no transcriptions at all, and these, like the ECCO texts that we produce, will be very correct as well.

But what do we mean exactly by "very correct"? By specifying typeface and creating a period-specific "spell-checker," we can achieve mechanically typed texts that are 97% correct. Our goal for how many and for the level of correctness has been carefully calculated based upon what has already been done, as is laid out in detail in the section of the Appendix called Document Error Evaluation (pp. 72). For texts published 1721 and after, we will improve the OCR by 2%—albeit a small amount, it makes a significant to people when confronted with the task of correcting the document to 100%. For texts published before 1721, we will improve the Gale OCR by 17%, and we will create OCR for EEBO documents that currently have none at the rate of 97% correct.

We have estimated conservatively the number of pre-1721 texts for which we will be able to perform this task but grounded that estimate in past successes. Thus can establish a clear and certain benchmark for OCR'ing EEBO and ECCO texts because of the work that has already been done on the ECCO texts and because we have the typed texts to work with. Based on these calculations: out of the EEBO and ECCO text collections, documents published in the US and UK between 1473 and 1800, we will produce during the two years of grant tenure

> Total Documents OCR'd and Post-Processed at 97% correct:
> **162,730** (approximately **23.7 million pages**).

However, we can and will achieve more than this by bringing human beings into the workflow, by **Orchestrating Human with Computer Work**. We will be creating tools that allow people to assist the OCR machines in reading correctly, performing some parts of their task that they cannot (specifically, font and line identification, as detailed above). And we will feed the 97% correct documents into a tool called TypeWright, built thanks to the generous funding of a Mellon Officer's grant. TypeWright allows people to correct words that were mis-typed by the OCR engine while looking at the page images. Based on current growth-rates in the usage of TypeWright—400 more documents either have been or are being corrected than were in March 2012—we estimate that, during grant tenure, we will produce:

> Total Documents corrected at 100% (corrected in TypeWright):
> **10,000** (**1.46 million pages**).

The grant work for this grant will enable us to speed up that process by a) giving people better texts to start; b) removing texts from the mix that are too flawed to work on; and c)

adding EEBO texts, which is to say, giving people more choices and opportunities to correct texts.

The ultimate goal of our work is to create a hybrid process of human computer work that persists until the work of correcting the early modern corpus constituted by EEBO and ECCO is done. This requires institutional investment in supporting the crowd-sourced correction process made available by ARC, which Mandell, along with her colleagues at NINES, MESA, and REKn, are all working toward procuring. Though we cannot obviously promise institutional support as a distinct grant deliverable, the Mellon Foundation's support of this grant will certainly help us to convince our universities of ARC's work to galvanize scholars around digitizing our cultural heritage.

Fostering and Harnessing Intellectual Curiosity isn't easy. But there is a way forward here as well. We will do outreach to citizen scholars, paying special attention to the successes of crowd-sourced correction projects such as the Australian Newspaper Digitisation Program and "Transcribe Bentham."[40] But we can also unleash academic interest, capacity, and concern through developing tools that let them engage deeply with the early modern documents and that rewards their curiosity as well as their desire to contribute to posterity.

Should we receive this grant, we will create an unprecedented opportunity for humanities and particularly literature faculty to engage with the proprietors who in fact provide the data preservation service that we need. Building correction tools, creating a correctable data set, and carefully saving and honoring scholarly contributions to its integrity allows scholars (both professorial and lay) to intervene in the quality of the archive that is being preserved by this model.

VII. Management Plan

The management of the Early Modern OCR Project (eMop) will be under the direct supervision of the PI, Laura Mandell. Day-to-day operations will be assisted by Mary Farrington. The PI and the co-PIs will form the Management Team which will have an advisory capacity to the PI and the project. The Management Team will in turn be informed by an assessment process based upon the various metrics established in the milestones and checkpoints (see above).

The Management Team along with Performant Software will meet on a weekly basis to assess progress against the project timeline, and twice per month with each of the collaborators during their heaviest works times (see "Tasks," Appendix pp. 71-8). Once a month, the Management Team will gather input from all collaborators and tests performed. Noting critical points, the team will project out one quarter to insure that attention is paid to these critical points and contingency plans articulated. Notes will be taken and posted along with a set of action items.

In order to share these postings, action items, and projected quarterly activities, a Wiki will be established. Here, the whole team can follow progress and contribute. The Wiki will be organized around the major tasks of the project and each task lead will be

[40] See Holley, "Many Hands" (note 19 above); Transcribe Bentham: http://www.ucl.ac.uk/transcribe-bentham/, Accessed 28 May 2012.

responsible for posting progress. Special attention will be paid to critical tasks and milestones that are important and also serve as a foundation for subsequent tasks. For example, we cannot start running the 260,000 EEBO and ECCO documents through the OCR engines until these engines have been optimized and the optimizations tested. Testing requires having a test set, the documents that Manmatha is preparing. If in January it looks as if he is having trouble creating the test set due to too many errors, team members at the IDHMC will, with expert advice from Clemens Neudecker at IMPACT, create a sample test set using Aletheia desktop. We need to put our contingency plan into effect long enough before we start running the documents in 4/13/12 that we can create a test set. The documents must start running during that 4th Quarter because it will take 8 months to run all the documents through the Brazos High Performance Computing Cluster.

Periodically, throughout grant tenure, by quarter, IMPACT leader Hildelies Balk will discuss the grant's progress and projected quarter with PI Laura Mandell, via Skype. Hildelies will be able to spot inconsistencies and pitfalls in planning. Within 90 days after the grant-work begins, a user group will be established consisting of individuals and organizations outside of the project that will help to evaluate the product in terms of their own individual needs. We will select members of the scholarly societies ASECS, SSA, and MLA who take a deep interest in preserving early modern cultural artifacts.

VIII. Personnel (Staffing)

In this section, we summarize the work on the project that will be done by each team and person More biographical information on each team member is available in the Appendix, pp. 16).

A. Texas A&M University

PI:

Mandell:

Director of the IDHMC, Professor Laura Mandell will be in charge of overseeing grant work per the Management Plan above, including writing all reports, as well as overseeing the tasks that will be performed by the IDHMC.

IDHMC, including 2 graduate students and 2 undergraduates student paid for by the grant:

The IDHMC's program manager Mary Farrington and Administrative Assistant Liz Grumbach will be assisting Mandell in arranging meetings and travel. The IDHMC's programmer Matthew Christy will be performing all work involving XSLT, including white-space analysis and transforming document and metadata outputs into forms needed by various tools and agencies. The IDHMC has hired Dr. Jacob Heil, currently a Postdoctoral Associate at Texas A&M University, for 12 months beginning September 2012 to build the font database and to assist Todd Samuelson in performing the research, photography, and sampling necessary for populating that database.

Two MA students per year for two years will optimize page images, create a ten-document test set in Aletheia Desktop, create font training libraries using Aletheia Desktop, and serve as test subjects for interfaces. One undergraduate student per year will work on layout analysis using Aletheia desktop and Aletheia web. The other

undergraduate will assist Dr. Jacob Heil and Todd Samuelson in photographing and scanning rare books.

Computer Science and Engineering Department

Gutierrez-Osuna + PhD student: Co-PI: **Dr. Ricardo Gutierrez-Osuna** is an associate professor in the Department of Computer Science at Texas A&M University. Dr. Gutierrez-Osuna and his Ph.D. student (to be named) will create document evaluation metrics and the automatic triage system.

Furuta + PhD student:

Co-PI Dr. **Richard Furuta** is a faculty member at Texas A&M University where he is a Professor in the Department of Computer Science, Director of the Center for the Study of Digital Libraries (CSDL), and Director of the Hypermedia Research Laboratory. Dr. Furuta and his Ph.D. student (to be named) will optimize the workflow system in terms of allocating what works best when done by machines to machines, and what works best when done by humans to humans.

HPC—Brazos Cluster

Dr. Guy Almes, Director of the Brazos High Performance Computing Cluster owned by the Academy of Telecommunications and Learning Technologies, will oversee and provide system administration—the work of Trey Dockendorf—to the early modern OCR project work as well as installation and upkeep of the servers purchased for the grant and running the cluster (see the letter of commitment, Appendix p. 5). Dr. Almes will be consulting with us constantly to make sure that the Cluster can run all our documents on time.

Cushing Library / TAMU Libraries Digital Initiatives Group:

Anton Duplessis, MA, Curator of the Colonial Mexican Collection at Cushing Library and Director of the Primeros Libros Project, and principle developer of the Cobre tool, will oversee the development of Cobre, for approximately 10% of his time per year. Scott Phillips, Alexey Maslov, and James Creel from the Digital Initiatives Group will be developing Cobre. They will work for two intensive five-week sessions, one per year, on adapting D-Space and Cobre to our needs. During the first five-week session, a Django/Djakota server configuration will be set up by Yixuan Li only for 18thConnect's instance of the Cobre tool and its log-in policies: anyone who logs into 18thConnect will be able to get an account. The library sysadmin Li will continue to manage that server during grant tenure, until we transfer the server to the 18thConnect infrastructure during the second year of the grant, when he will help us with that. During that first cycle, the changes to Cobre listed in the Appendix (p. 110) will be made. During the second year's development cycle, the usability studies conducted by Dr. Furuta's team will be used to make changes in the interface and functioning of Cobre by this team.

Dr. Todd Samuelson, Curator of Manuscripts and Rare Books at Cushing Library, for 15% of his time, will work together with Dr. Heil and the undergraduate assistant, both from the IDHMC, to do the research necessary for creating the Font Importation Database and create the database itself. After bringing in book historian James Mosley, Dr. Samuelson will travel to Antwerp and London to work at two major repositories for early modern typeface, both to do research and to request images for fonts not available in Cushing.

Outside Consultants: Retired Director of the St. Bride Library and Institute, James Mosley will travel to College Station to give Dr. Samuelson advice as they create

the Font History Database, and to make certain that Dr. Samuelson makes the best use of his research time in Antwerp and London. Book History expert James Raven and Restoration scholar Robert D. Hume will travel to College Station to be trained in how to use Cobre for four days. Dr. Furuta's team will observe how difficult this training is, and we will get feedback from Dr. Raven and Dr. Hume.

B. Subcontractors

The subcontractors for this project have each included Statements of Work, but here, we will describe their work more generally. All of them except Performant included travel money to come to College Station: the travel money for our overseas partners will not be paid by Mellon, as noted on those Statements of Work, but by the IDHMC. As programmers for ARC (Advanced Research Consortium), Performant comes to College Station twice per year anyway, so we did not need to add extra travel monies for them.

Performant Software: Nick Laiacona, Paul Rosen, Ed Zavada, and Kristin Jensen comprise this group that has been the primary programming company for NINES, 18thConnect, and MESA, and it is now undertaking to put up REKn (a Renaissance version of NINES) and to support ARC (the Advanced Research Consortium, currently sponsored by Texas A&M University's IDHMC). They will travel here to College Station in conjunction with an ARC meeting to help us work on document evaluation metrics. Performant set up the ARC Web Service that makes use of the Solr indexer which will ultimately take in all corrected OCR. They will during the course of this grant tweak and run our OCR engines on the Brazos Cluster, write the scripts we need to run large batches of documents, restructure the Solr indexer full-text search mechanism to scale it up and index by page image instead of by document, make changes to TypeWright, provide design and programming services for everything that the IDHMC cannot do, and systematize all the tools that we are building and workflows that we are establishing within the NINES/18thConnect/REKn/MESA universe. Finally, they will help us establish systems to export texts to all the necessary parties: Gale, ProQuest, and the TCP.

IMPACT group at the National Library of the Netherlands: **Hildelies Balk, Clemens Neudecker**, and the IMPACT team (IMproving ACcess to Texts) of the National Library of the Netherlands have worked for the last ten years on OCR'ing European texts. They have created the Center for Competence as a way of offering and sustaining services to European libraries who have OCR needs. IMPACT will be helping us in three major ways:

1. Assessing our progress, making sure that we are doing things in the proper order and on time to meet deliverables. IMPACT has ten years experience in working on OCR, and Hildelies Balk heads the group: her advice and oversight will be invaluable.

2. Helping us with training OCR engines in European fonts. Even though IMPACT cannot give us their font training because it is embedded in a commercial OCR engine, ABBYY FineReader, they have – again, 10 years – experience in creating OCR font training libraries and have done so for some of that time using Aletheia Desktop outputs.

3. Helping us set up our OCR process in Taverna, which they use for this purpose in the Centre for Competence developed out of IMPACT as its sustainable business model, established at the conclusion of their grant funding.

The SEASR group at the University of Illinois: **Loretta Auvil** works at the Illinois Informatics Institute (I3) at the University of Illinois at Urbana Champaign. **Boris Capitanu** is a research programmer working in the Illinois Informatics Institute (I3). Loretta Auvil and Boris Capitanu, will be working with Dr. Gutierrez-Osuna on creating document evaluation metrics via textual analysis. From work on the MONK project, the SEASR team excels at creating and testing n-gram principles and edit distances involved in dictionary lookups. This form of textual analysis is crucial both in evaluating documents and in preparing them for being loaded into TypeWright. during the second summer of grant tenure. Loretta Auvil will have a face-to-face meeting with Dr. Gutierrez-Osuna and his student early in the process in order to begin working on document evaluation metrics. Also at the University of Illinois, in the English Department, is Professor Ted Underwood who works with SEASR on early modern data sets for his topic modeling projects.[41] Dr. Underwood will provide us with the period-specific variant spelling lists and name gazetteers for persons and places that he is developing in Python.

PRImA Lab at the University of Salford (www.primaresearch.org): **Dr. Apostolos Antonacopoulos** is director of PRImA Labs, Pattern Recognition and Image Analysis Research Laboratory at the University of Salford, Manchester, UK. He developed Aletheia, a tool for the semi-automated layout analysis of documents. He is Senior Lecturer in the School of Computing, Science and Engineering at the University of Salford. PRImA will be fixing Aletheia Desktop to correct some usability issues, creating Web Aletheia for manual line segmentation, and helping us with font training using Aletheia Desktop. Apostolos will travel here once during the first year of the grant.

R. Manmatha + Ph.D. student: R. Manmatha, research associate professor in the Dept. of Computer Science at the University of Massachusetts, Amherst, and part of the Center for Intelligent Information Retrieval, and a student (to be named) will be helping us in two ways: 1) to create testing data (see Goal 3 above); and 2) to create a voting algorithm to compare and utilize multiple OCR outputs (see Goal 9 above). Manmatha and his student will travel to Charlottesville, VA, to work with Performant software on implementing the voting algorithm into the OCR workflow.

IX. Sustainability

After the end of the grant period, the corrections of OCR and triage of documents will continue in the paths and via the methods developed through the support of the Mellon Foundation. These tools should be sustainable with some maintenance and grow in usage as the community becomes aware of their existence and utility. The consortium built through this support will also look individually and collectively for additional resources to expand the effort and further refine the functionality of the tools that we

[41] Dr. Underwood has written extensively about his experiments: http://tedunderwood.wordpress.com/

have developed. Sponsorship for text-specific efforts in particular sectors might also be solicited where appropriate. The platform developed with support from the Mellon Foundation will have a long lasting impact on the community.

Once documents have been corrected in TypeWright, they will be periodically exported a) to the user-correctors; b) to the proprietors (Gale, ProQuest); and c) once 99% correct, to the TCP and the ARC catalog.[42] Additionally, corrections to the metadata made in Cobre will be stored in D-Space on a server at TAMU Libraries, sent to ProQuest, Gale, and TCP. These corrected metadata records will be sent to Brian Geiger at the University of Riverside, California, Director of the North American English Short Title Catalog (ESTC). The ESTC will review the corrected records and ingest them when appropriate. Cobre currently uses Dublin Core Metadata categories, but those can be mapped onto the Marc records that are needed for ESTC (this is the form currently needed, per an email correspondence with Brian Geiger, but that may change).[43]

For any text corrections made either in TypeWright or Cobre, 18thConnect, Gale, ProQuest, and the Text Creation Partnership will all hold copies of corrected texts, with regular updates as the crowd continues working on them.

Our crowd-sourced correction tools will continue to be used in 18thConnect and ARC which are supported by Texas A&M University, Mandell, and other faculty and students. Younger scholars are already contributing, and we are building in plans for new leadership. For instance, Mandell could pass the directorship of 18thConnect to Associate Director Brad Pasanek at the University of Virginia which has been supporting NINES continuously and so would be able to support 18thConnect as part of that commitment. Like NINES, in other words, 18thConnect will not live and die with one person but will become incorporated into the scholarly community. Community engagement is the best insurance for being sustained is, as we have learned from Jerome McGann's conference on sustainability, *The Shape of Things To Come* (http://shapeofthings.org/).

The IDHMC is building a business model for ARC in order to sustain it beyond the five years promised by Texas A&M University. While we would insist that the OCR outcomes would be available free and unfettered, there is a significant opportunity for building a not-for-profit model which is either based upon the service and/or outside investment into specific segments of the output. The underlying principle of preserving and making these documents freely available will help to guide the model.

The code-base for the tools that we build will be sustained by Performant Software and enhanced as they are used by18thConnect and others. For instance, JISC Collections has applied for a grant to incorporate TypeWright into its Historic Collections interface, as well as to build a smaller widget made to contribute nonce corrections by

---

[42] The ARC Catalog (http://catalog.performantsoftware.com/) is the SOLR indexer and Lucene search engine that currently serves data to NINES, 18thConnect, and that will be serving data to MESA and REKn.

[43] Per a recent email from Brian Geiger about revisions to the ESTC, 30 April 2012, the ESTC in North America and at the British Library are considering alternate "linked" data models like that being used for the British National Bibliography (Tim Hodson, "British Library Data Model: Overview," Talis Systems 22 July 2011 (http://talis-systems.com/2011/07/british-library-data-model-overview/), accessed 28 May 2012.

incidental users, especially on mobile devices, and this adoption will enhance and update the tool. And the leaders of Bamboo Corpora Space have applied for a grant to incorporate TypeWright into TextShop (per an email exchange with Neil Fraistat; letter from David Greenbaum coming).

X. Reporting

As mentioned in the management section, this effort will develop a real-time communications wiki to insure that the various elements of the project and their responsible organizations are kept in communication. This wiki-based platform will also include a space for posting progress again the milestones and outcomes for the various checkpoints. We have proposed a total of four CHECKPOINTS scheduled for 11/1/12; 1/28/13; 4/15/13; 3/15/14. With each checkpoint there is a metric against which progress will be measured. These outcomes along with other metrics will be reported to members of the community and available publicly on the wiki.

We will write interim and final reports to the Mellon Foundation, due October 15, 2013 and December 15, 2014. The Interim Report will show that we passed CHECKPOINTS 1-3 and attained MILESTONES 1 and 2 by 10/1/13; the Final Report will show that we attained MILESTONE 3 by 10/31/13, MILESTONE 4 by September 30, 2014, and that MILESTONE 5 is an outcome of the grant (results will be tracked from 9/30/14-12/1-14). We will include in the final report a full explanation of how the budgeted monies were spent. These reports will take the form of extended descriptions of progress made, research performed, lessons learned, and goals for the future. We will discuss each milestone delineated in this proposal, assessing how well we have met our goals in each case.

We will advertise the availability of our tools and databases on library listservs and in library journals, as well as in the Arts-Humanities Index sponsored by CenterNet. Martin Mueller and Laura Mandell are in the process of submitting a proposal to CLIR for a co-authored report on Early Modern OCR.

XI. Intellectual Property

Intellectual property will be developed and managed with the goal of having a nearly seamless set of tools available to all scholars.

We have proprietary issues concerning 1) Voting technology, 2) OCR engines and enhancements, 3) early modern texts, and 4) tools.

1) Prime Recognition announces its "voting technology" on its home page, http://www.primerec.com/ (17 May 2012). We have no access to, nor do we wish to have access to, Prime Recognition's proprietary voting algorithms. R. Manmatha is developing pair-wise comparison techniques for us that we will make freely available. While Prime Recognition has successfully deployed the "voting" technique, the concept of doing so is not theirs—it is widely known in machine learning and pattern recognition

fields, and it has been recommended for use in the OCR literature.[44] Texas A&M would never knowingly infringe upon a patent, and our Office of Technology Commercialization will do a patent search for the voting process should the project be funded so that we may work to particularize our own methods in a way that prevents patent violations.[45]

2) Out of two of the best OCR engines, Google's Tesseract and ABBYY FineReader, one is proprietary (the latter). According to Tesseract developers, there could be some question as to who owns the OCR output generated by commercial engines, whereas Tesseract itself, made available through an Apache 2.0 license, poses no such threat. Thus, though ABBYY FineReader currently may give slightly better results than Tesseract, it seems to us possible with less financial investment than it takes to purchase ABBYY and less risk to use Tesseract and two other open-access engines—Gamera and OCRopus—with similar Apache 2.0 licenses. All training and improvements of these engines funded by the Mellon Foundation will be freely available under an Apache Foundation license via github.

Moreover, instructions about when to use which component will be more than mere user-documentation: it will consist in two databases. A Re-Scanning Database will list the texts for which ECCO and EEBO page images are inadequate for preservation, texts that need to be re-scanned in whole or part. The other will list the fonts used when specific printers and booksellers in the trade are invoked either on the title pages or in the marc records. The Font Database will allow users of Gamera and Tesseract to know which Font Training Libraries to use on which texts. Unlike the source code for the engines, these databases will be made available on a website hosted by the IDHMC about the early modern OCR problem, and it will contain links to libraries / software components on github.

What we most hope to make available to libraries and collectors, large and small, is our workflow for best "massaging" early modern texts and thereby getting them into machine-readable, correctable form, whether one uses inadequate images or not. To that end, we will publish a freely available Taverna Workflow linking the processes that we built and concatenated. It will be made available under the Lesser GNU License (see item 3 below).

---

[44] William B. Lund, Daniel D. Walker, Eric K. Ringger, "Progressive Alignment and Discriminative Error Correction for Multiple OCR Engines," *2011 International Conference on Document Analysis and Recognition* pp. 764-768 contains a history of the idea as well as citations to the papers that have discussed it through 2011.

[45] Peter Schuerman, Ph.D., Director of Licensing and Intellectual Property in the Office of Technology Commercialization at Texas A&M (pschuerman@tamus.edu), explained the OTC's procedures.

Our two MySQL databases and their contents will be usable and downloadable by all under a Creative Commons Attribution-ShareAlike 3.0 Unported License. Any post-processing dictionaries and gazetteers that we develop will be made fully available via the IDHMC site, also under a Creative Commons Attribution-ShareAlike 3.0 Unported License.

3) Texts in the EEBO and ECCO databases were often given as images to these proprietary companies, ProQuest and Gale respectively, as partial payment for their preservation techniques. However, usage of many of the page images is still controlled by the libraries who have contracts with these companies, as is the case for all the early modern materials in the HathiTrust. It matters little whether the company is for-profit or not if usage of their page images is dictated by holding libraries.

We have negotiated a contract with Gale Cengage Learning who owns the ECCO catalog, reproduced in the Appendix (see p. 7), and summarized in an 18thConnect press release (http://www.18thconnect.org/news/?p=19#more-19). A recently negotiated addendum (Nov. 9, 2011) leaves the agreement thus: Gale has given us access to their page images and OCR results for use in TypeWright tool where one can see two-inch strips of the image and 3 lines of text at a time. We are also allowed to use a snippet of 50 words from the OCR in the SOLR/Lucene search index, though we only ingest corrected OCR. Most important, we are allowed to give any scholar who corrects a TypeWright text the TEI-encoded and plain-text versions of those texts. 18thConnect encourages scholars to make digital electronic scholarly editions, which they are free to do with the texts they have "earned" by correcting – Gale exerts no further claim on that text, but is willing to consider print-on-demand publishing of such ventures. The scholar is given the text, but not the page images, though those can be included in any volume printed by Gale itself.

At a meeting on April 24 with ProQuest who owns the EEBO Catalog, Mary Sauer-Games gave us a verbal agreement to sign the same contract as we have worked out with Gale-Cengage learning, allowing anyone to perform corrections and those who perform substantial work to receive copies of the texts they correct. Per an email dated May 3, 2012, she is currently in the process of putting that contract through her legal department (Appendix, p. 29).

Per the letter from Rebecca Welzenbach, the TCP will give us all 45,500 (latest figure) double-keyed texts to use for training the engines, a great boon. They can give us these texts because Texas A&M University is one of the partners with the TCP in Phases I and II (Appendix, pp. 7).

4) All tools created and used by Performant Software, including TypeWright, JuXta, and Collex, are open access via an Apache Foundation license and available on github: https://github.com/collex. Cobre is similarly open access, and we will release our modifications of it in the same way. Aletheia is freely available as a desktop tool, but the source code for that tool is proprietary to PRImA Labs. The source code for our version of it as a web tool will be released as open access code with an Apache Foundation license, via github, per the letter from PrimaLabs to be found in the Appendix (see p. 24). SEASR's post-processing workbench, Meandre, is available as source code and modules on the SEASR web site: http://www.seasr.org. The modules that we create in SEASR's workbench for post-processing will be made available through the HathiTrust Research center as well as the SEASR site itself. Taverna is available in all versions through the GNU Lesser General Public License v2.1 (http://www.oss-watch.ac.uk/resources/lgpl.xml). Our creation of a workflow in it will be made available through the same license, per its requirements. "General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish); that you receive source code or can get it if you want it; that you can change the software and use pieces of it in new free programs; and that you are informed that you can do these things." We will not contribute our work back to Taverna directly as the University of Manchester requires transfer of copyright to them for that purpose. We will only be using the system as currently constituted, and so copyright remains open access.

Thus our grant request meets the following terms of the Mellon IP Policy:

- Represents and warrants that it will solely own all intellectual property created with grant funds, either as work made for hire or as a result of a contractual agreement;

Per the contracts with Gale and ProQuest, the correction work done by faculty is theirs (A.23) to publish as they see fit, Gale having only first right of refusal (A.2.4) to the standard publishing contract. In the Gale Contract (Appendix, pp.19), the only product to which Gale Cengage Learning retains IP rights are **OCR Results** *(item 6), only one of the products produced by the grant, but* **the OCR plain text** *(1.2) and Metadata (1.4) are granted to NINES/18thConnect (2.1) through a "limited, non-exclusive, royalty-free right." Gale has the right to determine our display and use of it, but the* **OCR plain text** *is intellectual property retained by NINES/18THCONNECT, the property itself maintained in perpetuity on our servers. We choose to contribute those results back to the proprietors of EEBO and ECCO and to the Text Creation Partnership (TCP, where the plain-text data is subject to the contracts that the TCP has with Gale Cengage Learning and ProQuest, which means that they can release the textual data to the public in 2015, just as they have released 2,000 plain text files from the ECCO collection in 2011.*

- Represents and warrants that it has obtained the necessary licenses for third-party content and that the project will not infringe on third-party rights;

Mandell has negotiated with Gale Cengage Learning and ProQuest, and they in touch with their contributing libraries, primarily the British Library. As Caren Milloy's letter points out, the JISC Advisory Council is made up of representatives from the British Library who want 18thConnect to do this work.

- Will make software available, wherever possible, according to the terms of an open source license and in open source repositories, and will publicize its creations;

Please see 1 & 3 in Intellectual Property above, and Workplan D.12.

- Provides the Foundation the right to review the pricing and distribution of any software services, content, and digital products developed with Foundation funds;
- Will maintain any software created for a number of years beyond the term of the grant; and

Per Mandell's agreement with Texas A&M University, all work she creates for ARC during the first five years of her employment here, her term as director of the IDHMC, will be sustained by the IDHMC. Mandell's "year two" begins June 2012. Mandell's contract as director may be renewed for an additional five years; if so, the ARC tools will be supported an additional five years by IDHMC. Because the ARC servers will be purchased by the Mellon Grant, all tools, data, and processes will be made available on them for at least for the life of the servers. Should the IDHMC wish to use those servers beyond the duration of the grant, it must support, sustain, and augment them. Additionally, ARC catalog and tools will be preserved and supported beyond that time by NINES, REKn, 18thConnect, MESA, and ARC, whatever institution(s) support them in the future (currently Indiana Univ., Univ.of Virignia, North Carolina State Univ., University of Victoria, Dalhousie University, and Texas A&M University). Finally, through its "Data Management Plan," Texas A&M commits to sustaining data and software produced by any faculty member per the terms of the grants they receive in the Institutional Repository of TAMU libraries.

- Grants the Foundation a nonexclusive, royalty-free, worldwide, perpetual, irrevocable license to distribute any Foundation-funded software and/or digital products for scholarly and educational purposes, in the event the grantee cannot complete or sustain the project.

OCR plain text results will be owned and distributed by Faculty and Proprietors until such time as released open access via the TCP contracts:

> Currently, EEBO-TCP Phase II texts are available to authorized users at partner libraries. Once the project is done, the corpus will be available for sale exclusively through ProQuest for five years. Then, the texts will be released freely to the public.[46]

The current release date for EEBO is 2015; the ECCO release date is 2017. At that time, the Mellon Foundation may have nonexclusive, royalty-free, worldwide, perpetual, irrevocable license to the EECO / EEBO materials in the ARC catalog. The Mellon Foundation is immediately granted such rights to all

[46] http://www.textcreationpartnership.org/tcp-eebo/

open-access tools upon their creation and modification during grant tenure.

XII. Budget Narrative

[removed]