Mellon Interim Report

Matthew Christy
Elizabeth Grumbach
Laura Mandell

November 2013

Contents

Overview

The team at the IDHMC has been hard at work over the last year on the Mellon Foundation funded Early Modern OCR Project (eMOP). While we may not find ourselves where we expected to be at this moment, we have accomplished a great deal. Since work began we have encountered unforeseen obstacles and tasks that took longer to complete than anticipated. But at this point we have created a solid, robust OCR infrastructure and a great deal of institutional and technical knowledge that assure success in our future work.

Over the course of year one, the eMOP team has made significant advances towards our OCR goals, but the team has also faced significant challenges in achieving milestones and checkpoints on schedule. The following report will first present a narrative for year one of the grant, focusing on our successes, yet also paying close attention to how we have overcome unforeseen obstacles. We will close this interim report by looking towards the future, past the tenure of this grant, to how our lessons learned can inform further research. Please also see margin annotations throughout the document to identify which sections of the report address specific eMOP checkpoints and milestones.

Data

Wrangling all of the necessary data for this large OCR project proved one of our largest and least anticipated hurdles. We knew a file storage system and database was needed in order to store, manage, and access data and metadata related to over 45 million page images. But, work could not start on building and tuning these systems until we could get all of the data and metadata from various providers. Subsequently, this process was fraught with unexpected delays. Combining data and metadata from several different sources (e.g. ECCO, EEBO, TCP, and more) also required that everything we received had to be normalized before it could be ingested into the storage system or database. For example, because OCR is performed on a page-level process, when we received document-level files in our datasets, a program to break those files up into page-level units had to be written. Operating on large numbers of documents (in total: 45 million page images and ~30,000 transcriptions) meant that these programs took quite a while to complete—and then the results had to be checked for accuracy. In addition, we had to sort through the confusions caused by receiving data in several pieces, as multiple copies, or with incomplete or missing metadata—necessitating more time to organize, request clarification, or contact our providers for additional information/data.

eMOP is also a project involving excellent collaborators from around the country and the world. So creating a mechanism with which we could make all this data and metadata accessible to all

concerned was another task we hadn't anticipated. We were able to create such a system however, which has proven useful to collaborators and others outside of the project.

Finally, we had the task of integrating this database and file storage system with the processes and tools being created by Performant in order to control our workflow, perform the OCR processes, and display our results. This created further work as it demonstrated certain deficiencies in our systems as they stood and required some tweaking as well as some reingestion of data into the database. All of which took some more time, and all of which is standard operating procedure in large, computer-systems driven project.

In all, it took nearly 4 months to complete this task—one which we had not even included as a major task in our original proposal and Gantt Chart. However, I am happy to report that we now have a powerful, versatile, and fast system in place for managing the huge amount of data and metadata involved in this project. Our database is complete and is powering the workflow *for the entire process*. All tools built for eMOP processes write to or from our dedicated MariaDB database system. Our file system is fast and well organized despite its size. We have made this data sharable and open to our current collaborators, and this functionality is sustainable, should the need arise for further manipulation.

Training OCR Engines

During this data wrangling we also worked on creating the proper environment on our computer cluster to perform OCR training and testing with multiple OCR engines. This work involved installing software, establishing security credentials for all necessary personnel including collaborators, and creating secure information exchange channels with our database and file system. Each of these steps involved issues of their own and required a constant dialogue between our team, system administrators, and software developers to ensure that we accommodate all security concerns inherent in a University computing system.

Once we had collected and normalized a sufficient amount of data we began testing OCR systems. We began with Tesseract (an open-source OCR engine from Google) since we believed, based on a review of the OCR literature, that it was the most accurate and fastest open-source OCR engine available, giving us the best chance for success. After a number of tests however, we discovered that our method of training Tesseract to recognize early modern typefaces was insufficient. Tesseract expects to be trained using text documents that meet a certain set of specifications. Training documents must contain real words and sentences created with certain line and word spacing (kerning) requirements. This kind of training is easy to create using word processing software in conjunction with modern fonts. However, it is difficult, if not impossible to create this kind of training using currently available tools and early modern fonts. The team proposed and pursued a variety of methods to create the kind of training documents required by Tesseract. But after further testing with these methods we discovered that Tesseract was not

using its training to recognize glyphs (characters, punctuation, etc.) in the way we had expected it to.

In short, an OCR engine is trained to recognize glyphs in a page image as characters by feeding it, in a pre-defined and proprietary way, representatives of each possible glyph that it will encounter in the page image, along with a "definition" (using the Unicode character set) of what character that glyph describes. The problem with early modern printed documents—the problem at the very heart of this grant project—is that, due to the inconsistency of printing inherent in the early modern period, the glyphs representing a single character can have very different characteristics within a single document or even on a single page. For example, multiple instances of a lower-case letter "a" can look quite different in a single document for various reasons (see Figure 1):

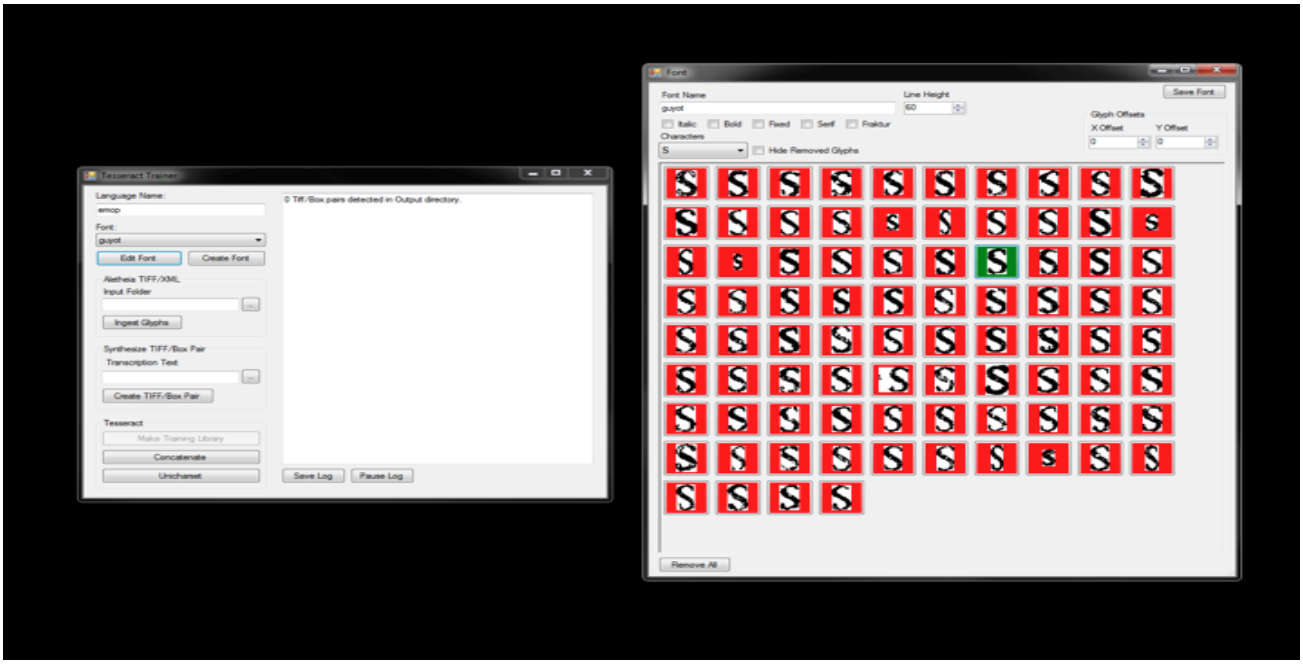
- Noise introduced in the digitization process,
- Noise created by the fading, smudging, tearing, etc. of the original document's pages,
- Warping of the physical pages,
- Skewing of the page images during digitization,
- Wear created on the character's physical print punch through heavy or long-term usage,
- Inconsistencies in the physical application of ink to the print block (over- and under-inking), etc.

Figure 1: A small sample of lower-case "a"s from one document.

Our prior experience with the open-source Gamera OCR engine, indicated that feeding the engine with a variety of different glyphs for a single character improved it's ability to recognize various forms of the glyph as they were encountered in a page image. Our assumption, necessitated by the fact that documentation for Tesseract is sparse at best and can even be misleading or inaccurate, was that Tesseract worked in a way similar to this. However, testing showed that the more, various forms of a character we provided in training, the worse it's character recognition became. After vigorous debate the team concluded that what we needed was a new tool that could be used at an intermediate stage of the training process that would allow us to examine all the glyphs of every character and choose only one, or a small sample of, glyph(s) that best represented each character and then use that set of samples to create a training file for Tesseract[1].

As a result, IDHMC graduate student and member of the eMOP team, Bryan Tarpley, created a tool that we're calling Franken+.

Figure 2: The Franken+ interface, as of October 2013.



Checkpoint 1: With the addition of Franken+, we verified that Aletheia Desktop was needed to complete the project.

To create specific font training for the Tesseract OCR engine, a team of undergraduate student workers lead by another eMOP graduate student, Katayoun Torabi, first process the available page images using Aletheia Desktop[2]. Aletheia Desktop includes several semi-automated tools that identify and define layout regions, lines, words, and individual

[1] **--Note** concerning the meaning of this result for book historians: We have begun working with a theory of font analysis developed by Adrian Weiss.

[2] Aletheia was developed by the Pattern Recognition and Image Analysis (PRImA) Research Laboratory at the University of Salford. Apostolos Antonacopoulos, IMPACT Work Package leader for PRImA, University of Salford, has made Aletheia and other tools available at http://www.primaresearch.org/tools.php. PRImA is an eMOP collaborator.

characters (glyphs) within documents. Aletheia reads the text in the page image (using Tesseract) and assigns a Unicode value for each letter, number, and punctuation mark. As output, Aletheia creates an XML file that contains a set of XY coordinates for each defined region, and for each glyph this file also includes a Unicode value defining what character is represented by that glyph. The data contained in this XML file is then ingested or imported into Franken+. Franken+ uses a MySQL database to associate each glyph image with its corresponding Unicode character. The user can then select any glyph from a drop down menu, see every instance of that character in a window (Figure 1 was created from a screenshot of this window), and choose the best image for each character in that font set. Once the user has isolated the best instance of each character, Franken+ uses a standard text document to produce a set of synthetic TIFF images and XML files, producing a "Franken-text" with only these ideal characters. This Franken-text can then be used to train Tesseract to recognize the typeface being processed.

Checkpoint 1: With the successful creation of a number of Tesseract training sets, that show marked improvement based on initial font training (in Roman typefaces), we have confirmed that further OCR training in this manner is needed.

Franken+ has been well received since it was introduced to the world at the Association for Computing Machinery's (ACM) Document Engineering (DocEng[3]) Conference at Florence, Italy in September, 2013[4]. Franken+ is still in "beta" testing but is currently in use by collaborators in Europe, in the Texas A&M Cushing Memorial Library and Archives, and by our own team of undergraduate student workers. We are very excited about the creation of this new tool to aid in training Tesseract, and can see great potential in it's application for other areas of research related to typefaces. We have already been contacted by several international scholars who are interested in using Franken+ as a tool to aid their own research into early modern typefaces. And the people who have used Franken+ so far have used words like "easy" and "fun" to describe it. Since beta testing began we have ironed out a number of bugs and hope to have it officially released via eMOP's Github repository[5] by the end of 2013.

While we have found that some of our original assumptions about Tesseract were incorrect, leading to a much longer time to reach the stage where we are actively undergoing OCR'ing of our data set, we are confident that we are back on track. Our understanding of Tesseract and how it works is greatly improved. We have created a new open-source tool that not only makes creating training sets for Tesseract simple, but also promises to change how typeface research is conducted. We have created a number of training sets for Tesseract that demonstrate a marked improvement in Tesseract's ability to recognize early modern fonts. We are also working on finding or creating an early modern dictionary with variant spellings that will also improve Tesseract's ability to correctly identify words during the OCR process.

[3] http://www.doceng2013.or

[4] Torabi, Katayoun, Jessica Durgan, and Bryan Tarpley. "Early Modern OCR Project (eMOP) at Texas A&M University: Using Aletheia to Train Tesseract." ACM Press, 2013. 23. *CrossRef*. Web. 31 Oct. 2013.

[5] https://github.com/idhmc-tamu/eMOP

Concurrent to our work on Tesseract, the eMOP team also conducted OCR tests using the Gamera engine. Gamera promised to have superior character recognition to Tesseract, based on what we knew about its training, but also had known issues with other aspects of OCR'ing. For example, Gamera is known to have more problems with line segmentation than Tesseract, especially on page images that are noisy, skewed, or have figures and/or tables. Like with Tesseract, we created training for Gamera on a specific typeface (Baskerville) and then used that training to OCR a number of documents known to be printed with that typeface. Our testing revealed that, in the presence of any of the issues listed above, Gamera took way too long to OCR the documents. In fact, several of the documents failed to complete OCR'ing in the 9-hour window established for it during submission for processing on the computing cluster. The time it takes for Gamera to OCR a document is just too high to offset any gains in character recognition, which were not even clear from the tests. During this same time we discovered that another open-source OCR engine we had intended to test (OCRopus) had discontinued development and was no longer working. We decided to turn our full attention to Tesseract[6].

Checkpoint 1: By proving our general hypothesis that, when specific rather than general letter shapes are sought, the engines distinguish more letters from noise, we proved that a font history database was needed.

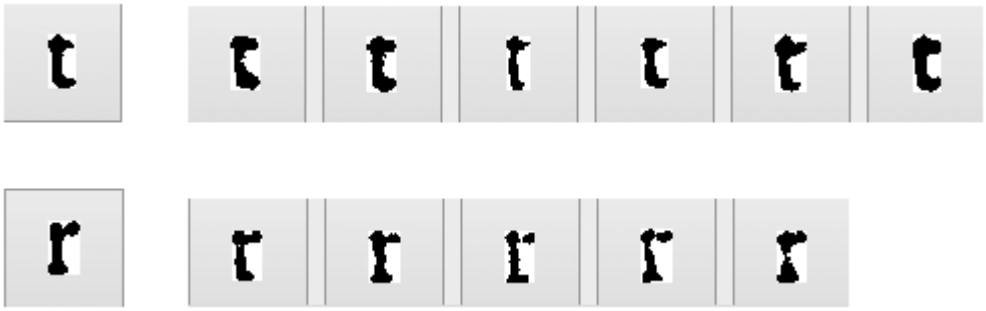
Font History DB

At the beginning of the project the need for a database of typefaces, and when they were used by which printers, was an open question. Once we were able to develop training for Tesseract that yielded meaningful results, we began to investigate this question. We began by testing the efficacy of training Tesseract on basic typeface "families"—broad groups of typefaces that share many general characteristics, which typically were designed and/or produced by the same person or company. For example, we created training for Tesseract based on a typeface specimen sheet produced by the printer François Guyot (circa 1560)[7], which we then used to OCR documents printed from 1550-1700. Our result indicated that for many character glyphs, this broad-based training was adequate for good character recognition. But for some character glyphs, small changes in typeface form combined with some of the problems mentioned above (image noise, worn punches, over-/under-inking, etc.) caused Tesseract to consistently mis-read one character as another. Figure 3 shows several samples of lower-case "r" and "t" glyphs—Tesseract consistently recognized most "t" glyphs from the page image as "r."

[6] Because Gamera was proven in a previous grant to have an excellent capacity for identifying images, we hope to include Gamera in a future version of this project. With our limited time remaining, we will be focusing on Tesseract due to it's ability to OCR quickly, but after our post-processing algorithms detect page images with pictures, we hope to apply Gamera to these documents in order to achieve better OCR results. Please see the "Phase 2" section of this report for more information on our future goals.

[7] http://collation.folger.edu/2011/09/guyots-speciman-sheet/

Figure 3: A sample of "t" abd "r" glyphs from the Guyot Specimen Sheet



This type of mis-reading is not uncommon in OCR engines, especially when trying to read early modern printed documents with their inherent quality variance. But we did see a lower recognition rate for these types of characters as we OCR'd documents that were further removed in time from the training typeface, indicating that we would need to train Tesseract to recognize typefaces at a more granular level. We concluded that a font history database would indeed be necessary, as it would allow us to associate printed documents with typefaces more closely correlated with them based on printer and time period.

We have begun this process by producing a database of printers, locations, and time periods for the EEBO collection. This information is being captured from the imprint lines of the EEBO collection documents as represented in the EEBO metadata we received from Proquest. Owing to the irregular nature of storing data in natural language strings, parsing all of the required information from the imprint lines is an understandably complicated and iterative process. (See Figure 4.) Work is progressing well on this front.

Concurrently, we have begun the process of digitizing and/or mining the ESTC (1473-1800)[8] to create a database of printers and the typefaces they used during different time periods. When we have combined the EEBO/ECCO imprint metadata with the ESTC metadata, we will have a valuable new resource, which will allow us to associated EEBO/ECCO documents via their printers and publication dates with specific typefaces. This resource will allow us to OCR documents with training specific to the typeface(s) they were printed with. And we also feel this will be a valuable new resource for any scholar involved in book and printing history research.

[8] http://estc.ucr.edu

Figure 4: A sample of the imprint lines from EEBO.

eMOP Work ID	Author	Publisher	Publication Date	Title
37332	F. K.	[S.l. : R. Watkins, 1580]	1580	Of the crinitall starre, which appeareth this October and Nouember, 1580
37406	Beaumont, Francis, 1584-1616.	Printed at London : For Thomas Walkley, and are to be sold at his shop at the Eagle and Child in Brittaines Bursse, 1620.	1620	Phylaster, or, Loue lyes a bleeding acted at the Globe by His Maiesties seruants / written by [brace] Francis Baymont and Iohn Fletcher ...
37407	Beaumont, Francis, 1584-1616.	Printed at London : For Thomas Walkley, and are to be sold at his shop at the Eagle and Child in Brittaines Bursse, 1620.	1620	Phylaster, or, Loue lyes a bleeding acted at the Globe by His Maiesties seruants / written by [brace] Francis Baymont and Iohn Fletcher ...
37556	Melanchthon, Philipp, 1497-1560.	[London] : Imprinted at London at the three Cranes in the Vinetree by Thomas Dawson, 1580.	1580	A godly and learned assertion in defence of the true church of God, and of His Woorde written in Latine by that Reuerend Father D. Philip Melancthon, after the conuention at Ratisbona, anno 1541 ; translated into English by R.R.
37636	Rowlands, Samuel, 1570?-1630?	[S.l. : G. Purslowe, 1620?]	1620	[A paire of spy-knaues]
38178	Colville, Elizabeth Melvill, Lady Colville of Culros, fl. 1603.	Edinburgh : Imprinted by Andro Hart, 1620.	1620	A godlie dreame compiled by Elizabeth Melvill, Ladie Culros younger, at the request of a friend.
38213	Corporation of London.	Printed at London : By VVilliam Iaggard, printer to the honourable city of London, 1620.	1620	By the major a proclamation for the prices of tallow and candles.
38230	Lupton, Thomas.	Printed at London : By H. Bynneman, dwelling in Thames Streate, neere vnto Baynards Castell, 1580.	1580	Siuqila too good, to be true : omen : though so at a vewe yet all I tolde you is true, I vpholde you, now cease to aske why? for I can not lye : herein is shewed by way of dialogue, the wonderful maners of the people of Mauqsun, with other talke not friuolous.
38231	Lupton, Thomas.	Printed at London : By H. Bynneman, dwelling in Thames Streate, neere vnto Baynards Castell, 1580.	1580	Siuqila too good, to be true : omen : though so at a vewe yet all I tolde you is true, I vpholde you, now cease to aske why? for I can not lye : herein is shewed by way of dialogue, the wonderful maners of the people of Mauqsun, with other talke not friuolous.

Book History Research

With our training regimen, based on Franken+, in place and our need for a font history database established, the eMOP team recognized the need for more data with which to train the Tesseract OCR engine. What we needed most was a set of high-resolution images of good quality typeface samples. Specimen sheets (good quality printings, sometimes made on modern equipment, using early modern type punches, displaying samples of every glyph for that typeface) provide us with a concise collection of good quality glyphs that can be digitized and ingested into Franken+. Our book history collaborator at the Cushing Memorial Library and Archives at Texas A&M University, visited museums and libraries in Antwerp, Amsterdam, London and Oxford, to generate a collection of these images and to identify other typefaces for which we may request images be made later (due to time constraints or photography restrictions). With this collection we will be able to construct yet another open-source database, which can then be linked to the font history database currently under construction.

In addition to this work, while engaged in research for eMOP, our book history collaborator discovered an article by Adrian Weiss[9] that opened up a new, previously unanticipated line of research. Weiss theorized that typefaces, in a specific period of early modern printing, contain some character glyphs which exhibit two different sets of characteristics. He classified these as S-

[9] "Font Analysis as a Bibliographic Method," *Studies in Bibliography* 43 (1990): 95-164.

face and Y-face characters. The book history team at Cushing is currently using Franken+ to examine the typefaces we have already processed to identify which ones contain S-face and/or Y-face characteristics. When done we can examine our typeface training test data in the context of this new information, and run additional tests if necessary, to determine if this theory can be utilized in training Tesseract.

Triage

In testing Tesseract, and determining the best way to create training for it, we learned a number of things that required us to update our post-processing triage system as described in the grant proposal.

Checkpoint 3

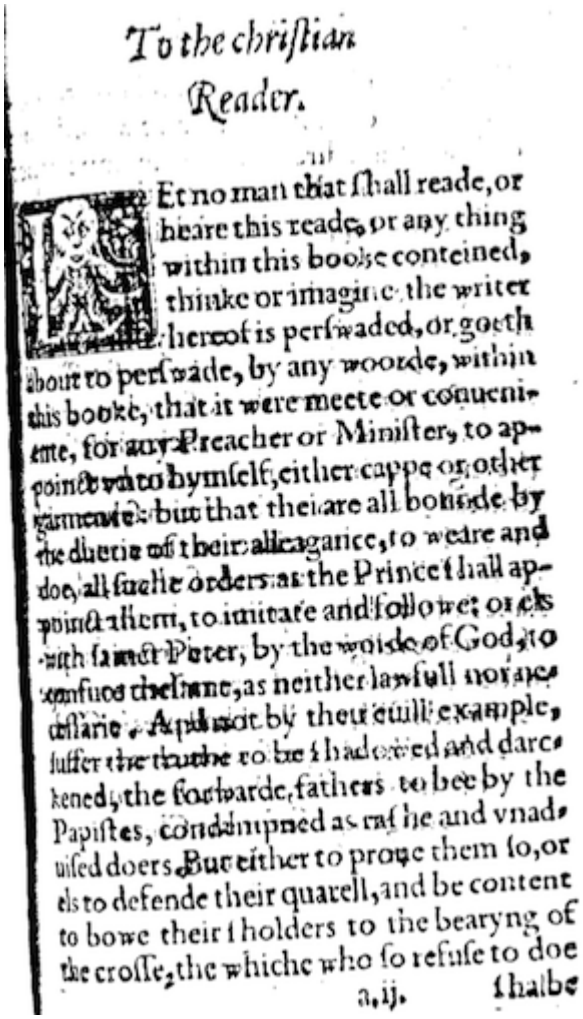
As per Checkpoint 3 in the grant proposal (pg. 28, scheduled for April, 2013), we discovered that with Tesseract there is insufficient correlation between the time it takes to OCR a page and its correctness. This determination requires that we "add further evaluation measures to that of time as a preliminary determinant of OCR performance" (28).

Poor Quality Inputs

During our testing of Tesseract the eMOP team also discovered that many documents, especially in the EEBO collection, are of such poor quality, that they will need to undergo one or more pre-processing steps before they will yield good OCR results. Figure 5 below displays several characteristic problems inherent in many EEBO page images: general noise, bleedthrough, over-inking, skewing, and warping. However, with over 45 million pages to OCR, determining which page images require what kind of pre-processing before OCR'ing is impracticable based on time and manpower requirements.

To that end, the eMOP team, in conjunction with our post-processing collaborators at the Software Environment for the Advancement of Scholarly Research (SEASR) at the University of Illinois, Urbana-Champaign and Ricardo Gutierrez-Osuna of Texas A&M University, are currently developing an expanded triage system. In addition to the automatic triage functionality already laid out in the grant proposal (pg. 27), the expanded system will contain a suite of diagnostic algorithms that will indicate what is most likely wrong with page images that fail to produce usable OCR results.

Figure 5: Example EEBO page image.



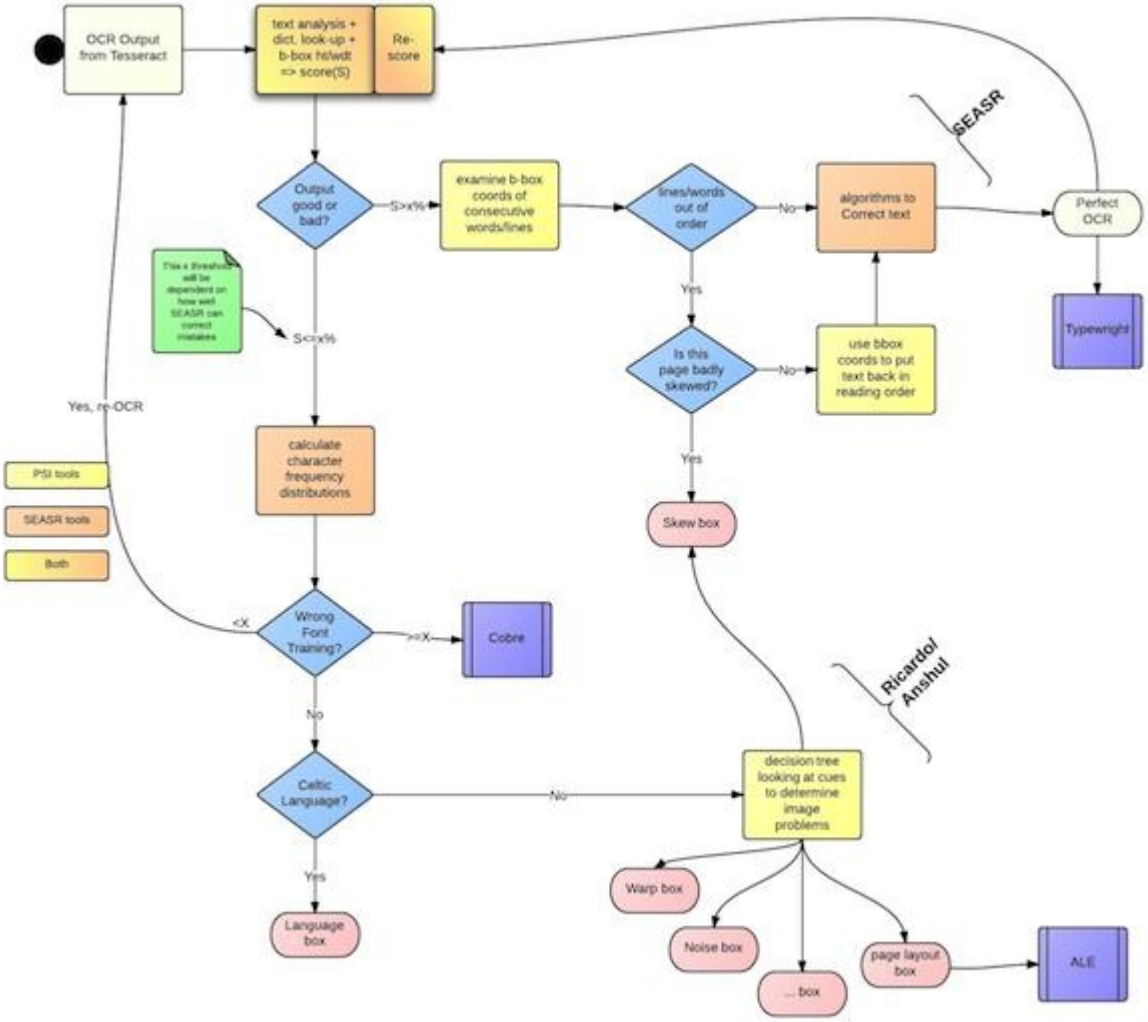
To the christian
Reader.

LEt no man that ſhall reade, or heare this reade, or any thing within this booke conteined, thinke or imagine the writer hereof is perſwaded, or goeth about to perſwade, by any woorde, within this booke, that it were meete or conuenient, for any Preacher or Miniſter, to appoint vnto hymſelf, either cappe or other garmente: but that theiare all bounde by the dueties of their allegaunce, to weare and doe, all ſuche orders as the Prince ſhall appoint them, to imitate and followe: or els with ſainct Peter, by the woorde of God, to confound the name, as neither lawfull nor conuenient. And not by their euill example, ſuffer the truthe to be ſhadowed and darckened, the forwarde fathers to bee by the Papiſtes, condemned as raſhe and vnadviſed doers. But either to proue them ſo, or els to defende their quarell, and be content to bowe their ſholders to the bearyng of the croſſe, the whiche who ſo refuſe to doe

Our post-processing triage system will now result in one of the following for each OCR'd page image (see Figure 6):

- Usable OCR text is ported to 18thConnect and Typewright for improved searching by scholars and additional crowdsourced correction when needed.
- Pages that repeatedly fail (using a metric based on further testing) due to OCR'ing with the wrong trained typeface will be ported to the Cobre tool, modified for eMOP use by a team at the Texas A&M University Libraries, for by-hand font identification by experts.
- Badly warped and skewed pages will be ported to the Aletheia Layout Editor tool (ALE), created for eMOP by collaborators at PRImA, for by-hand line segmentation.
- Pages images that fail based on image quality or language identification issues will be sorted for pre-processing, before they are re-scanned (see *Phase 2*, pg 16).

Figure 6: New eMOP Post-Processing Triage Workflow



The eMOP team is very excited about this new triage workflow. It represents a dramatic new development in OCR processing. The ability to identify the problems that exist in a subset of pages from a larger set of page images is also a big step forward. We have not had this kind of information before, especially for the documents that make up the EEBO collection. Getting this kind of data is the first step to rectifying what has been an intransigent problem for Humanities researchers. We have also received interest in this process from our collaborators in Europe currently working on the SUCCEED and Europeana Newspaper projects, and from members of the HathiTrust organization.

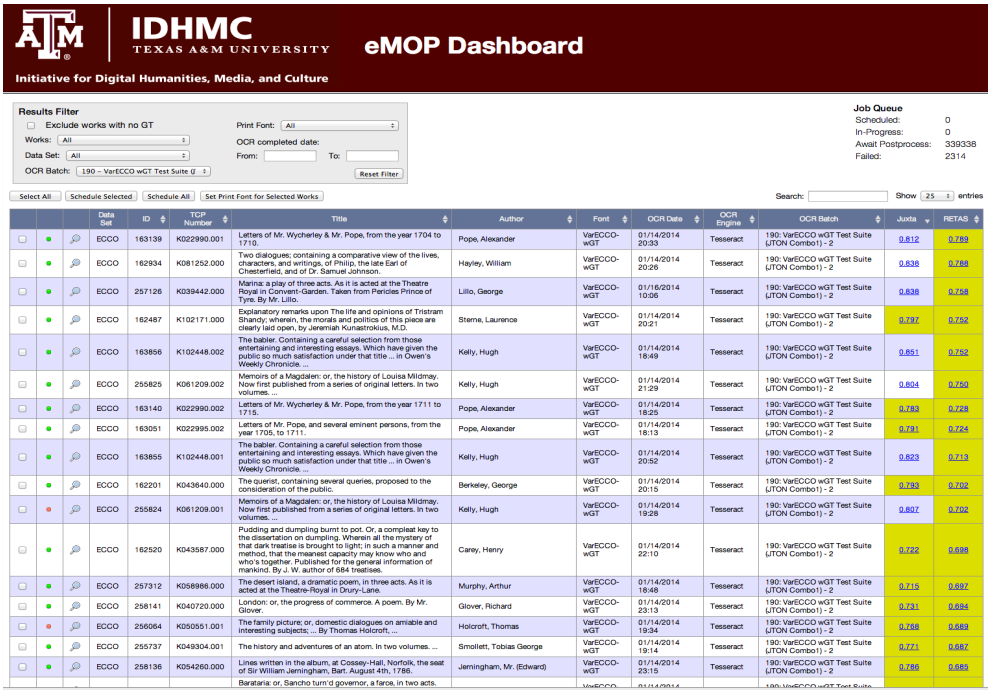
Tools Development

The development of tools both internal and external to the eMOP team is progressing full steam ahead.

- **DB Tools:** In conjunction with the database we created to control the incredible amount of data and metadata we collected, the eMOP team found that we also had to create tools that would allow us to make that information available. The *eMOP Query Builder* is an online database interface, which allows the eMOP team and collaborators to query the database. Once a subset of data has been identified the *Data Downloader* creates an XML file, which can be used with third-party tools to query the database and download the page images, OCR results, and/or text transcriptions of the identified documents. The *eMOP Query Builder* and *Data Downloader* are installed on an eMOP server and can only be accessed by eMOP collaborators (due to the proprietary nature of the data sets).
- **Dashboard:** Created by Performant software in close coordination with the eMOP team, the *eMOP Dashboard* (Figure 7), powered by the eMOP DB, controls the entire OCR process via a table-based web interface. The eMOP Dashboard:
    - Allows users to select documents and begin the OCR process;
    - Sends each selected document by page to the selected OCR engine;
    - Collects the resulting OCR result file(s), stores them in the eMOP file system, and writes this location to the eMOP DB for each page of that document;
    - If ground-truth files are available for that document in the eMOP DB, then they are run through two diff'ing algorithms to create accuracy scores of the OCR results compared to ground-truth; these scores are stored in the eMOP DB for each page of that document;
    - Displays information about each document that has been OCR'd (Title, Author, eMOP ID number, etc.), information about the OCR process (date, batch number), and ground-truth scores (when available);
    - Allows users to drill down into document results to see results by page, including original page images, OCR'd text, and ground-truth scores (when available).

  The *eMOP Dashboard* is installed on and running on an IDHMC virtual machine.

Figure 7: eMOP Dashboard



			Data Set	ID	TCP Number	Title	Author	Font	OCR Date	OCR Engine	OCR Batch	Juxta	RETAS
☐	●	🔎	ECCO	163139	K022990.001	Letters of Mr. Wycherley & Mr. Pope, from the year 1704 to 1710.	Pope, Alexander	VarECCO-wGT	01/14/2014 20:33	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.812	0.789
☐	●	🔎	ECCO	162934	K081252.000	Two dialogues; containing a comparative view of the lives, characters, and writings, of Philip, the late Earl of Chesterfield, and of Dr. Samuel Johnson.	Hayley, William	VarECCO-wGT	01/14/2014 20:26	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.838	0.788
☐	●	🔎	ECCO	257126	K039442.000	Marina: a play of three acts. As it is acted at the Theatre Royal in Convent-Garden. Taken from Pericles Prince of Tyre. By Mr. Lillo.	Lillo, George	VarECCO-wGT	01/16/2014 10:06	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.838	0.758
☐	●	🔎	ECCO	162487	K102171.000	Explanatory remarks upon The life and opinions of Tristram Shandy; wherein, the morals and politics of this piece are clearly laid open, by Jeremiah Kunastrokius, M.D.	Sterne, Laurence	VarECCO-wGT	01/14/2014 20:21	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.797	0.752
☐	●	🔎	ECCO	163856	K102448.002	The babler. Containing a careful selection from those entertaining and interesting essays. Which have given the public so much satisfaction under that title ... in Owen's Weekly Chronicle. ...	Kelly, Hugh	VarECCO-wGT	01/14/2014 18:49	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.851	0.752
☐	●	🔎	ECCO	255825	K061209.002	Memoirs of a Magdalen: or, the history of Louisa Mildmay. Now first published from a series of original letters. In two volumes. ...	Kelly, Hugh	VarECCO-wGT	01/14/2014 21:29	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.804	0.750
☐	●	🔎	ECCO	163140	K022990.002	Letters of Mr. Wycherley & Mr. Pope, from the year 1711 to 1715.	Pope, Alexander	VarECCO-wGT	01/14/2014 18:25	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.783	0.728
☐	●	🔎	ECCO	163051	K022995.002	Letters of Mr. Pope, and several eminent persons, from the year 1705, to 1711.	Pope, Alexander	VarECCO-wGT	01/14/2014 18:13	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.791	0.724
☐	●	🔎	ECCO	163855	K102448.001	The babler. Containing a careful selection from those entertaining and interesting essays. Which have given the public so much satisfaction under that title ... in Owen's Weekly Chronicle. ...	Kelly, Hugh	VarECCO-wGT	01/14/2014 20:52	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.823	0.713
☐	●	🔎	ECCO	162201	K043640.000	The querist, containing several queries, proposed to the consideration of the public.	Berkeley, George	VarECCO-wGT	01/14/2014 20:15	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.793	0.702
☐	●	🔎	ECCO	255824	K061209.001	Memoirs of a Magdalen: or, the history of Louisa Mildmay. Now first published from a series of original letters. In two volumes. ...	Kelly, Hugh	VarECCO-wGT	01/14/2014 19:28	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.807	0.702
☐	●	🔎	ECCO	162520	K043587.000	Pudding and dumpling burnt to pot. Or, a compleat key to the dissertation on dumpling. Wherein all the mystery of that dark treatise is brought to light; in such a manner and method, that the meanest capacity may know who and who's together. Published for the general information of mankind. By J. W. author of 684 treatises.	Carey, Henry	VarECCO-wGT	01/14/2014 22:10	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.722	0.696
☐	●	🔎	ECCO	257312	K058986.000	The desert island, a dramatic poem, in three acts. As it is acted at the Theatre-Royal in Drury-Lane.	Murphy, Arthur	VarECCO-wGT	01/14/2014 18:48	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.715	0.697
☐	●	🔎	ECCO	258141	K040720.000	London: or, the progress of commerce. A poem. By Mr. Glover.	Glover, Richard	VarECCO-wGT	01/14/2014 23:13	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.731	0.694
☐	●	🔎	ECCO	256064	K050551.001	The family picture; or, domestic dialogues on amiable and interesting subjects; ... By Thomas Holcroft, ...	Holcroft, Thomas	VarECCO-wGT	01/14/2014 19:34	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.768	0.689
☐	●	🔎	ECCO	255737	K049304.001	The history and adventures of an atom. In two volumes. ...	Smollett, Tobias George	VarECCO-wGT	01/14/2014 19:14	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.771	0.687
☐	●	🔎	ECCO	258136	K054260.000	Lines written in the album, at Cossey-Hall, Norfolk, the seat of Sir William Jerningham, Bart. August 4th, 1786.	Jerningham, Mr. (Edward)	VarECCO-wGT	01/14/2014 23:15	Tesseract	190: VarECCO wGT Test Suite (JTON Combo1) - 2	0.786	0.685
						Barataria: or, Sancho turn'd governor, a farce, in two acts.							

Checkpoint 2: With the creation of RETAS and OCTO (below), we confirmed that these algorithms can handle the original flawed OCR. We proved that there was no need to use Aletheia to generate semi-automated ground truth.

- **Diff'ing Algorithms:** In order to test the accuracy of our OCR methods, we had collaborators create two diff'ing algorithms which compare OCR results to ground-truth in different ways. *Juxta* was created by Performant and is based on the Juxta Collation software they developed for NINES. *The Recursive Text Alignment Tool (RETAS)* is another diff'ing tool developed by R. Manmatha's team for eMOP (documentation here: http://ciir.cs.umass.edu/downloads/ocr-evaluation/).
Both the *Juxta* and *RETAS* algorithms have been created and installed on the eMOP Dashboard.
- **Voting Algorithm:** In order to compare the accuracy of a given OCR engine on a given character, page, or document, R. Manmatha's team developed the OCR Error Correction Tool (OCTO) (http://ciir.cs.umass.edu/downloads/octo/). This tool aligns three OCR outputs of the same page or document in order to "produce a corrected version." While the OCRopus engine is no longer a viable option, we may use this voting algorithm to compare Tesseract, Gamera, and the original Gale OCR results side-by-side.
- **TypeWright:** Phase 1 and 2 of TypeWright development was completed, in conjunction with Performant Software. The following were accomplished:
  - Finished TypeWright-enabling the 65,000 texts from the ECCO dataset; these texts are now available via 18thConnect.org;
  - Added the ability to add a new red box, above and below each line of text;
  - Added the ability to resize red boxes around lines, for line segmentation;
  - Added the ability for users to mark texts complete, after "editing" every page;

- Added the admin dashboard, which allows the admin to easily sort and view edited documents, see the percentage correct, review how much of a document was edited by any specific user, and confirm a document complete;
- Added exporting capability to the admin dashboard, using the XSLT's prepared by Matthew Christy, for the following formats: Original Gale XML, Original text, Corrected Gale XML, Corrected text, Corrected TEI-A.

These developments have been completed and are now live on 18thConnect.

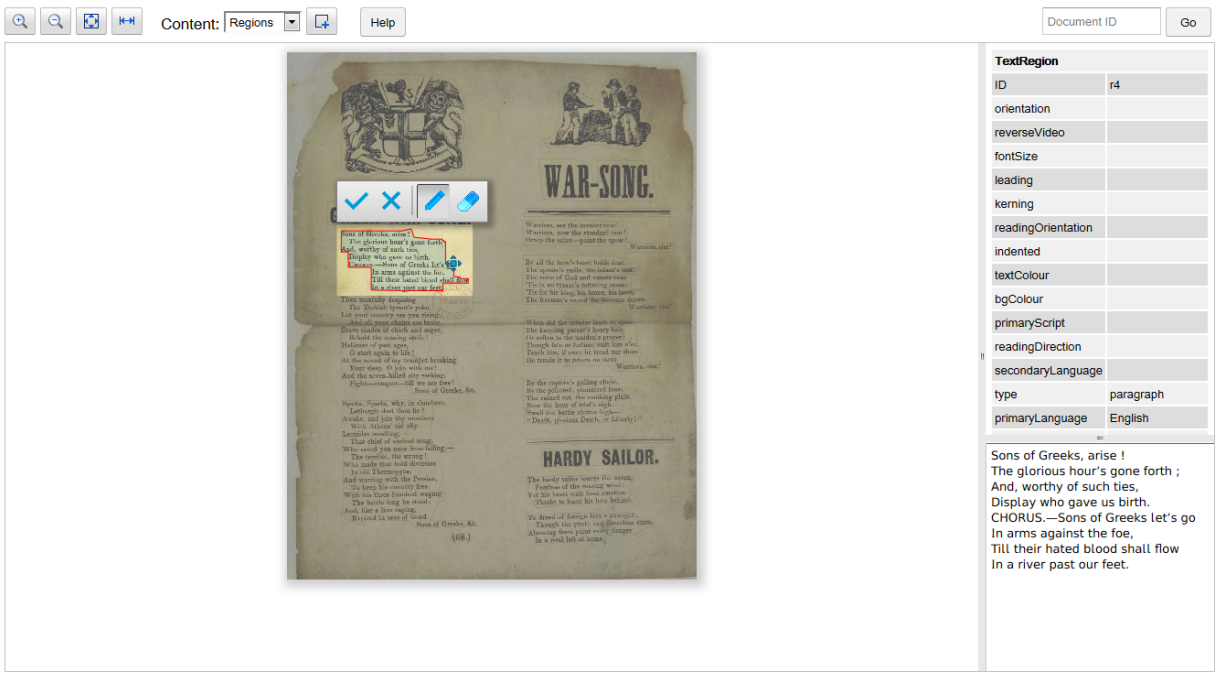
- **Franken+:** As mentioned before, *Franken+* is an exciting new tool which can be used to train Tesseract, research typefaces, and possibly more functionality we have not yet anticipated.

Franken+ is currently in beta testing and the source code should be released on the IDHMC/eMOP github page by the end of this year (2013).

- **AWL editor:** Created by collaborators at PRImA, the Aletheia Web Layout Editor (AWL editor) is a web-based tool which allows users to identify regions on a page image such as paragraphs and lines. This information can then be fed into an OCR engine when the engine is having trouble identifying these regions on its own.

AWL has been built.

Figure 8: Aletheia Web Layout Editor (AWL editor), as currently designed.



The eMOP team needs to refine its design to make the best possible interface, install AWL on one of our servers, and integrate the tool with Collex.

- **Cobre:** Programmers at the Texas A&M University Libraries have adapted their *Cobre* program to create an online interface for book history experts to closely examine document pages. Cobre will allow experts to identify the typeface used to print a page,

Milestone 2: The milestone for releasing our crowd-sourced correction tools was September 2013. As detailed on this page, this milestone is almost completed. While we are waiting on data to propagate the tools, TypeWright development has been completed, and basic technical and graphical design for Cobre and Aletheia Layout Editor (ALE) have been completed. An additional, excellent tool was added to this release, as well (see Anachronaut).

and correlate portions of OCR'd documents with documents that are unable to be OCR'd.

Cobre has been built and the library has created a mechanism for ingesting documents into it. Cobre now needs to be integrated into Collex.

- **Anachronaut:** *Anachronaut* is a game designed for Facebook, which allows players to play the role of an early modern printer. Players are fed images of words cut from the eMOP collection of documents which they must correctly identify in order to receive credits with which they can buy equipment to improve their printing shop. The game was developed by a group of Texas A&M University Computer Science undergraduate students of eMOP collaborator Ricardo Gutierrez-Osuna.

  Development on *Anachronaut* has been completed, and it has been installed on an eMOP server. We are currently working to finalize issues to get the game working on Facebook. While the game does not currently use images from our proprietary datasets, we are in discussions to extend the permissions used for the TypeWright tool to Anachronaut.

Figures 9-11: Anachronaut Prototype.

Computing Platform

As we continue to work, one concern is the time it will take to OCR the over 45 million page images that make up our corpus. While the Brazos Cluster is an amazing and powerful resource, its full power is not available without a sizeable financial contribution ($20,000) to make us stakeholders in the Cluster. As non-stakeholders we are relegated to OCR'ing page images using unutilized computing cycles, making the time that it takes to OCR our collection unpredictable. As such, it could take anywhere from 3-9 months to complete all OCR work. To counter this instability, the IDHMC has just completed negotiations with the Brazos Cluster and the College of Liberal Arts (CLLA) to become a stakeholder by hosting a server rack on CLLA property, which would constitute a remote resource of the Brazos Cluster. This resource will contain 768 core processors and allow the eMOP team to OCR every page in our collection in approximately 2 weeks. This would dramatically increase our ability generate the kind of text results we desire by letting us try a variety of approaches to the problem.

Conclusion

While we have fallen behind the Milestones and Checkpoints that were established prior to beginning work, as laid out in the grant proposal's Timeline Gantt Chart (pg. 35), the eMOP team nonetheless feels we have made tremendous progress to this point. We have successfully tackled issues and pursued lines of research that were not identified prior to beginning. We have created new tools and established communication with researchers that are informing our process as we continue to pursue our stated goals. Our investigations of the optimal options for training Tesseract, in particular, but also Gamera to read early modern fonts will be beneficial to future OCR projects using similar data sets. Due to these intensive investigations, we are getting ever closer to our promised 93% correct (as promised for Milestone 1), and, due to the work of our post-processing collaborators, that number is expected to improve significantly.

List of eMOP Accomplishments for Year One

- The eMOP team collected metadata, page images, XML and OCR files, and transcriptions from all of our eMOP partners, including ECCO and EEBO, for review and organization.
- eMOP graduate student Bryan Tarpley built the eMOP Database, which combines ECCO, EEBO, TCP, ESTC, and MARC record metadata, page images, transcriptions, and more. This eMOP database is extremely interoperable, built on a MariaDB server, which is the most robust, reliable, and scalable MySQL server, and the center of all eMOP functionalities, software, and tools.
- eMOP graduate student Bryan Tarpley built the eMOP Query Builder, an interface that automatically builds mysql queries for the user, based on a series of options, in order to output a specific subset of data needed for a collaborator. The Data Downloader tool allows any collaborator with proper permissions and access to successfully and easily download subsets of the eMOP data based on certain classifications (MARC record number, eMOP work ID, TCP number, and more).
- James Raven and Robert Hume, two book history experts, came to consult with the eMOP team on the overall premise of the eMOP workflow, and how to interact with other book history scholars through eMOP tools. To this end, Drs. Raven and Hume were the test

subjects for our Cobre user study, in which the response from Raven and Hume led us to adopt different terminology for the Cobre tool.

Figure 12: James Raven and eMOP team.

Figure 13: Robert Hume, Dir. of Cushing J. Larry Mitchell, and Distinguished Professor Margaret Ezell

- Phase 1 and 2 of TypeWright development, in conjunction with Performant Software, in which the following were accomplished:
    - Finished TypeWright-enabling the 65,000 texts from the ECCO dataset; these texts are now available via 18thConnect.org;
    - Added the ability to add a new red box, above and below each line of text;
    - Added the ability to resize red boxes around lines, for line segmentation;
    - Added the ability for users to mark texts complete, after "editing" each page;
    - Added the admin dashboard, which allows the admin to easily sort and view edited documents, see the percentage correct, review how much of a document was edited by any specific user, and confirm a document complete;
    - Added exporting capability to the admin dashboard, using the XSLT's prepared by Matthew Christy, for the following formats: Original Gale XML, Original text, Corrected Gale XML, Corrected text, Corrected TEI-A.
- The eMOP team and ARC Project Manager, Liz Grumbach, collaborated to move the ARC and TypeWright servers from the Rackspace and Softlayer clouds, respectively, to dedicated VMs at the IDHMC at Texas A&M.
- The Cushing Memorial Library and Archives team, lead by Dr. Anton DuPlessis, finished development and customization of the Cobre tool for the eMOP project; the Cobre tool enables scholar-experts to compare, re-order pages, and annotate the metadata for multiple

printings of documents in the eMOP dataset. Further work needs to be done to incorporate Cobre into Collex, for use in 18thConnect and REKn.

- The PRImA research team finished initial development of the Aletheia Web Layout Editor (AWL editor), the tool is now web-ready and allows the user to re-draw regions on problematic OCR'd pages, such as Title pages, multi-columned texts, image-heavy documents, and more. Further work needs to be done to design the user-facing interface for the tool and for incorporation into Collex.
- The Aletheia training team at the IDHMC, lead by eMOP graduate student Katayoun Torabi, worked with PRImA research to improve the Aletheia Desktop tool. In effect, Torabi's team was able to produce multiple Tesseract training libraries with specific typefaces (including the popular Baskerville and Guyot).
- Bryan Tarpley finished development and beta release of the Franken+ tool, which enables the creation of an "ideal" typeface using glyphs identified in scanned images of documents from the early modern period. Franken+ also exports these typefaces to a training library for the open-source OCR engine Tesseract.
- Dr. Ricardo Gutierrez-Osuna and a team of computer science undergraduate students at Texas A&M developed the Anachronaut tool, a Facebook game that uses the power of social media (and many layers of user confidence testing) to correct single words and phrases. Anachronaut has been moved to TAMU servers, and we plan to have the tool running on Facebook by the end of the year. Further discussions with ECCO and EEBO will be needed to extend the permissions provided to TypeWright to Anachronaut.

Figure 14: The Aletheia desktop tool is used by a team of undergraduate student workers and a graduate student supervisor to identity specific glyphs within a font type.

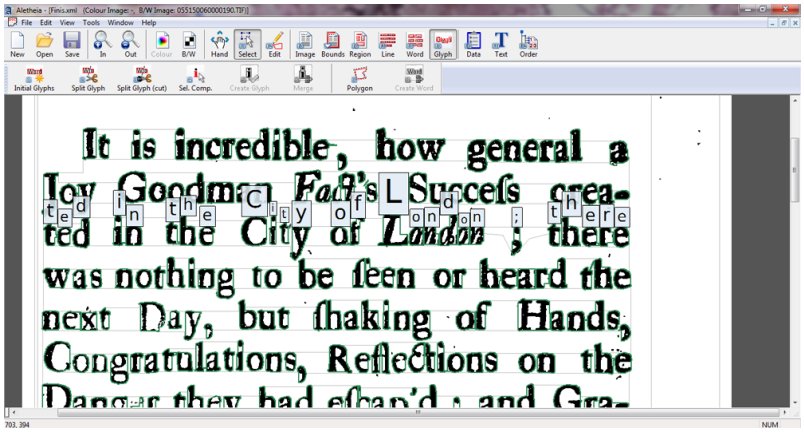


Figure15: Dr. Laura Mandell, PI, with Performant Software Solutions Director, Nick Laiacona, at MLA 2013.

- Our two book history experts, Dr. Jacob Heil (also eMOP Project Manager, Year One) and Dr. Todd Samuelson of Cushing Memorial Library and Archives published the article "Book History in the Early Modern OCR Project, or, Bringing Balance to the Force" in the Journal for Early Modern Cultural Studies[10].
- Mandell presented eMOP at MLA on a division panel where Performant had a booth demonstrating TypeWright at the publishers exhibit (Fig. 8).

[10] Heil, Jacob and Todd Samuelson. "Book History in the Early Modern OCR Project, or, Bringing Balance to the Force." *Journal for Early Modern Cultural Studies* 13.4 (2013): 90-103. Web. 30 Oct 2013.

- Mandell presented eMOP at the European Science Foundation Exploratory Workshop "Knowing about Mediation" convened by James Raven at Cambridge University, 15-18 September 2013.
- Mandell presented eMOP at the SSHRC-funded Social, Digital, Scholarly Editing Conference held by Prof. Peter Robinson 13 July 2013.
- eMOP graduate students Bryan Tarpley and Katayoun Torabi, along with eMOP volunteer Dr. Jessica Durgan, presented a paper on eMOP at DocEng 2013: the 13th ACM Symposium on Document Engineering[11].
- Katayoun Torabi, Bryan Tarpley, and Dr. Jessica Durgan published their eMOP presentation in the DocEng 2013 proceedings.
- Mandell published an article on the necessity of eMOP, and projects like it, in the *Journal for Early Modern Cultural Studies*[12].
- Dr. Jacob Heil (book history consultant and eMOP Project Manager, Year One) was invited to present on "eMOP and Book History" at the SHARP 2013 conference's Digital Projects Showcase.
- Book history consultant Dr. Todd Samuelson of Cushing Memorial Library and Archives scheduled, planned, and travelled to Europe (Antwerp, Amsterdam, and the UK) to complete research into early modern typefaces. He aims to do some more intensive research tracing the movement of European fonts into England, culminating in the photographing of specific font specimen sheets, as researched and requested by the eMOP team.
- In conjunction with Performant Software, the eMOP team designed and implemented the eMOP Dashboard, which runs on the Brazos High Computing Cluster at Texas A&M. The Dashboard:
    - Allows users to select documents and begin the OCR process;
    - Sends each selected document by page to the selected OCR engine;
    - Collects the resulting OCR result file(s), stores them in the eMOP file system, and writes this location to the eMOP DB for each page of that document;
    - If ground-truth files are available for that document in the eMOP DB, then they are run through two diff'ing algorithms to create accuracy scores of the OCR results compared to ground-truth; these scores are stored in the eMOP DB for each page of that document;
    - Displays information about the document that has been OCR'd (Title, Author, eMOP ID number, etc.), about the OCR process (date, batch number), and ground-truth scores (when available);
    - Allows users to drill down into document results to see results by page, including original page images, OCR'd text, and ground-truth scores (when available).
- In coordination with IDHMC systems administrator Trey Dockendorf, the eMOP team has customized and improved the "eMOP controller" code to optimize the availability of the Brazos High Computing Cluster.
- IDHMC Lead Programmer Matthew Christy (also Co-Project Manager for eMOP, Year Two) tested variations of font training combinations for Tesseract and Gamera in order to confirm that training OCR engines with early modern fonts would make significant improvements in the OCR of early modern documents. For training with Tesseract, he

[11] Katayoun, Torabi, Jessica Durgan and Bryan Tarpley. "Early modern OCR project (eMOP) at Texas A&M University: using Aletheia to train Tesseract." *Proceedings of the 2013 ACM symposium on Document Engineering*. New York: ACM, 2013.

[12] Mandell, Laura. "Digitizing the Archive: The Necessity of an 'Early Modern' Period." *Journal for Early Modern Cultural Studies* 13.2 (2013): 83-92.

consulted with the Google development team and the Google groups for Tesseract in order to customize Tesseract for academic projects. For training with Gamera, he was able to customize a set of instructions for using Gamera for OCR (it is originally a glyph recognition software)[13].

- eMOP collaborator R. Manmatha, and his graduate student Zeki Yalniz, developed a customized version of their RETAS (recursive text alignment) tool for this project[14]. This tool is being used to evaluate the effectiveness of our OCR, when that OCR has ground truth available. This algorithm is currently being used in the eMOP dashboard to test our testing and training progress, and it will eventually be used to help us determine our overall OCR correctness score. Manmatha and Yalniz have also released an API for use with the tool, which will be incorporated into the final OCR workflow.
- eMOP collaborator R. Manmatha, and his graduate student Zeki Yalniz, developed the OCTO tool, the OCR Error Correction Tool, which aligns three OCR outputs of the same page or book in order to produce a corrected version[15].
- eMOP collaborators at Performant Software developed a Juxta algorithm for eMOP, in order to evaluate the effectiveness of our OCR, when that OCR has ground truth. This tool was adapted from Performant's current Juxta Web Service (https://github.com/performant-software/juxta-service) and Juxta Commons (http://www.juxtacommons.org) implementations. Performant built the JuxtaCL tool, which is a command line version of the change index tool based on the Jaro-Winkler distance, but the visualizations that are produced by the Juxta Web Service tool can also be seen through the eMOP dashboard, and are generated on command.
- Dr. Jacob Heil and eMOP undergraduate Abhay Ohri began forming a database of printers and publishers in early modern England, based on the imprint line from ECCO and EEBO. This database is intended to be the basic structure from which the early modern font database will be built. Currently, Matthew Christy and Liz Grumbach are combining data from this imprint line, the ESTC, and the STC for further construction of this resource.
- Matt Enis wrote an article in *The Library Journal* about eMOP project goals and the importance of eMOP work[16].
- ASECS has offered TypeWright full-day, pre-conference workshops over the last three years, 2012, 2013, and will do so again in 2014, paid for by the IDHMC / Texas A&M.
- Elizabeth Grumbach submitted to the Digital Humanities Conference 2014 an abstract for a panel in which participants on the eMOP project will discuss how we had to change the direction of the project very quickly in order to keep abreast of our discoveries about OCR engines, training them, and data differences. It has received very high reviews, and so we are planning on presenting the story of the eMOP grant at DH2014 in Switzerland this summer.

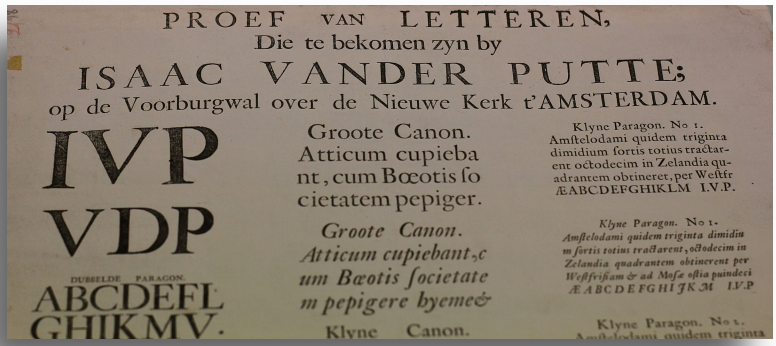
Figure 16: Specimen sheet of Canon from Antwerp (detail).

[13] http://emop.tamu.edu/node/49

[14] http://ciir.cs.umass.edu/downloads/ocr-evaluation/ and https://github.com/idhmc-tamu/eMOP/tree/master/RecursiveTextAlignmentTool_release_v1_1

[15] http://ciir.cs.umass.edu/downloads/octo/

[16] http://www.thedigitalshift.com/2012/11/digitization/next-gen-ocr-project-reaches-back-into-early-english-history-and-databases/


PROEF VAN LETTEREN,
Die te bekomen zyn by
ISAAC VANDER PUTTE;
op de Voorburgwal over de Nieuwe Kerk t'AMSTERDAM.
IVP
VDP
DUBBELDE PARAGON.
ABCDEFL
GHIKMV.
Groote Canon.
Atticum cupieba
nt, cum Bœotis so
cietatem pepiger.
Groote Canon.
Atticum cupiebant, c
um Bœotis societate
m pepigere hyemeq
Klyne Canon.
Klyne Paragon. No 1.
Amstelodami quidem triginta
dimidium sortis totius tractar-
ent octodecim in Zelandia qu-
adrantem obtineret, per Westfr
ÆABCDEFGHIKLM I.V.P.
Klyne Paragon. No 1.
Amstelodami quidem triginta dimidiu
m sortis totius tractarent, octodecim in
Zelandia quadrantem obtinerent per
Westfrisiam & ad Mosæ ostia puindeci
Æ A B C D E F G H I J K M I.V.P
Klyne Paragon. No 1.

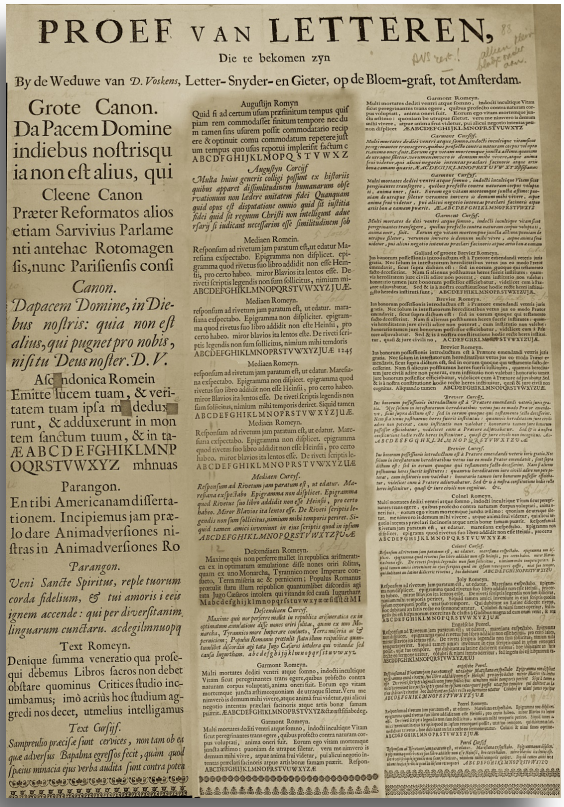
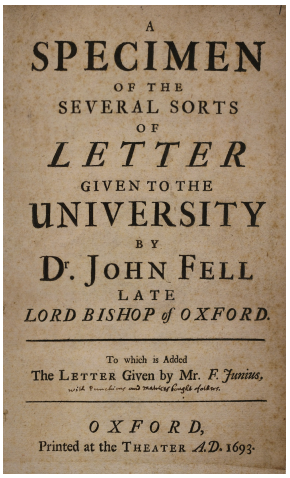

PROEF VAN LETTEREN,
Die te bekomen zyn
By de Weduwe van D. Voskens, Letter- Snyder- en Gieter, op de Bloem-graft, tot Amsterdam.
Grote Canon.
Da Pacem Domine
indiebus noſtrisqu
ia non eſt alius, qui
Cleene Canon
Præter Reformatos alios
etiam Sarvivius Parlame
nti antehac Rotomagen
ſis, nunc Pariſienſis conſi
Canon.
Dapacem Domine, in Die-
bus noſtris: quia non eſt
alius, qui pugnet pro nobis,
niſi tu Deus noſter. D. V.
Afc ndonica Romein
Emitte lucem tuam, & veri-
tatem tuam ipſa m deduc
runt, & adduxerunt in mon
tem ſanctum tuum, & in ta-
ÆABCDEFGHIKLMNP
OQRSTVWXYZ mhnuas
Parangon.
En tibi Americanam diſſerta-
tionem. Incipiemus jam præ-
lo dare Animadverſiones ni-
ſtras in Animadverſiones Ro
Parangon.
Veni Sancte Spiritus, reple tuorum
corda fidelium, & tui amoris i eis
ignem accende: qui per diverſit. anim.
linguarum cunctaru. ac degisimnuopq
Text Romeyn.
Denique ſumma veneratio qua proſe-
qui debemus Libros ſacros non debet
obſtare quominus Critices ſtudio in-
cumbamus; imò acriùs hoc ſtudium ag-
gredi nos decet, utmelius intelligamus
Text Curſiv.
Sumprelio præciſæ ſunt cervices, non tam ob ea
quæ adverſus Bapalma egreſſos fecit, quam quod
ſpeciæ minacia ejus verba audita ſunt contra poten

Figure 17: Full Canon Specimen Sheet.


A
SPECIMEN
OF THE
SEVERAL SORTS
OF
LETTER
GIVEN TO THE
UNIVERSITY
BY
Dr. JOHN FELL
LATE
LORD BISHOP of OXFORD.
To which is Added
The LETTER Given by Mr. F. Junius,
OXFORD,
Printed at the THEATER A.D. 1693.


FINIS.

Figures 18-19: Specimens of Fell type; printer's ornament.

541	*Conference Track / Type of Submission:* **Panel / Multiple Paper Session** **Navigating the Storm: eMOP, Big DH Projects, and Agile Steering Standards** Grumbach, Elizabeth M (1); Christy, Matthew J (1); Mandell, Laura (1); Neudecker, Clemens (2); Auvil, Loretta (3); Samuelson, Todd (5); Antonacopoulos, Apostolos (4) *Organization(s):* 1: Initiative for Digital Humanities, Media, and Culture (IDHMC), Texas A&M University, United States of America; 2: KB National Library of the Netherlands, Netherlands; 3: Illinois Informatics Institute (I3) at the University of Illinois at Urbana Champaign, United States of America; 4: Pattern Recognition and Image Analysis (PRImA) research Lab at The University of Salford, United Kingdom; 5: Cushing Memorial Library and Archives, Texas A&M University
	1st file dh2014-abstract447-2.pdf (1st Nov 2013, 07:26:50pm CET)

Figure 20: Future Presentation of the eMOP Project proposed for the major Digital Humanities Conference.