September 30, 2015

To: Donald Waters, Senior Program Officer, Scholarly Communications, The Andrew W. Mellon Foundation

From: Laura Mandell, Professor of English, Director, Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University
Matthew Christy, Lead Developer, IDHMC
Elizabeth Grumbach, Project Manager, Advanced Research Consortium (ARC)

A. Summary of the Project

The goal of the Early Modern OCR Project (eMOP) was to create a process for OCRing—mechanically typing—digital page images of early modern texts published between 1473 and 1800 so that they would no longer be opaque to the eyes of the machine. Machine-readable (typed, keyed) texts are necessary for full-text searching documents, many of which have metadata that incompletely or even erroneously describes them. These digital page images came to us from the ProQuest company's Early English Books Online (EEBO) and Gale-Cengage Learning's Eighteenth-Century Collections Online (ECCO), together totaling 307,000 documents, about 45 million pages. Because it is precisely *modern* printing practices (beginning between 1820 and 1830) that make it possible for machines to read digital page images and create text files from them, a new set of training and post-processing techniques were needed to handle *early modern* texts, as well as to compensate for the poor quality of the images in each collection: originally microfilmed from British Library Collections, the digitized page images from EEBO and ECCO are low in information density, imaging pages that themselves were beset by problems endemic to early modern printing and preservation practices: uneven type, page bleed-through, missing, torn, or blackened pages, Blackletter fonts, variant spellings, printing errors, imperfect inking, to name a few. In order to confront these challenges, we intended to marshal the newest and best, open-access OCR engines, creating a workflow and set of tools around them that would make it possible to achieve correct transcriptions of these page images.

B. Progress Made Toward Expected Outcomes of the Grant

The expected outcomes of the grant were of six types:

1) OCR training for early modern typefaces

We created over 40 training libraries for Tesseract, using typefaces produced from the 17th and 18th centuries, all of them available on our Github site (http://early-modern-ocr.github.io/) for free to anyone who wishes to use them. Some of them, the Enschede and Caslon, for instance, were produced using specimen sheets that we obtained by sending book historian Todd Samuelson of TAMU's Cushing library to Europe in order to get pictures of these sheets from the Plantin-Moretus Museum (Antwerp), the Museum Enschedé (Amsterdam), and St. Bride's (London).

The open-access OCR engine Tesseract, which we used for the eMOP project, can only be trained by using font sets. Due to Tesseract's native training mechanism, Tesseract can only read images of fonts that have an equivalent in Microsoft Word, for instance, or Adobe Illustrator. Of course early modern typefaces have no such digital instantiations: we had to fabricate them using page images of the fonts. English graduate student Bryan Tarpley created

the FrankenPlus Tool for creating digital font sets out of page images that have been processed using Aletheia, software developed for the IMPACT project in Europe (IMProving ACcess to Texts) and modified for us by our partners at the Univ. of Salford. FrankenPlus is also available on the eMOP website and GitHub.

Thus, we created many training libraries for the Tesseract OCR engine as well as tools for creating more of those training libraries. Please note that some word and character lists—ligature lists (unichar-ambig files) as well as dictionary/variant spelling lists (dawg files)—are used as part of the Tesseract OCR engine's process. Those files are described and made available on the eMOP site as well, both as part of font training libraries and separately (fully explained here http://early-modern-ocr.github.io/ TesseractTraining/). One can see how much better the OCR engine performs if one uses our training than it does with the built-in, default training (Appendix A).

Additionally, we have created a printer's database, listing all the early modern printers we could discover in the EEBO and ECCO documents at https://github.com/Early-Modern-OCR/ImprintDB: our goal is to release it to the book history community and begin the research process of determining which typefaces were used by which printers. While currently only available in XML format, this database will eventually be released in a dynamic, collaborative user interface. This database has been carefully prepared through several iterations of context analysis and was then normalized. It is an iterative reading of data from the imprint lines of those two collections, with an aim towards pulling out information about who the document was printed by, printed for, sold by, and the detailed location of creation.

2) Triage and post-processing routines

Prof. Ricardo Gutierrez-Osuna and his graduate student Anshul Gupta at TAMU created an hOCR denoising algorithm that works on OCR outputs to remove noise—usually outputs produced in the absence of text (maps, printers' ornaments, blackened or torn pages). This team at TAMU has also determined a process for triaging documents that have extensive issues that prevent or complicate OCR. This process uses machine learning to measure skew, diagnose typeface families, and identify coordinates for multiple columns on a page. The SEASR team at the University of Illinois created a Page Evaluator tool and a Page Corrector tool that evaluates the OCR output correctness (highly incorrect pages are removed from the OCR outputs) and then uses dictionary files and a google three-gram database to correct OCR output. These as well as two tools for determining correctness in relation to ground truth, Juxta-cl and Retas, are available via the eMOP Github Repo (http://early-modern-ocr.github.io/).

3) High percentage correctness of OCR results

We have not yet achieved the 97% level of correctness that we had hoped to by the end of the grant period, but our results using only one OCR engine, as opposed to multiple engines plus a voting algorithm to determine the word most often selected, are very high indeed, only 3% lower than the OCR that was made for Gale-Cengage Learning by PrimeRecognition, a company that uses a voting mechanism, hand correcting, and six of the top commercial OCR engines. Using our own measures, we estimate our level of correctness on ECCO documents at 86%, and Gale-Cengage Learning's at 89% (they claim 90-95% on their website but are simply using different measures). Our level of correctness on EEBO is significantly lower: 68%. We have determined that a high number of documents in the EEBO collection are poorly imaged, and therefore not preserved digitally, as libraries might presume, given that a specific title appears in the EEBO collection. We will release a database of unreadable pages so that libraries can know what texts

in their special collections do NOT have adequate digital surrogates, encouraging those libraries to create such digital archival materials and allow us to OCR them.

Although the average level of correctness for EEBO was 68%, approximately half of the page images in that collection range in correctness from 80 to 100%, the statistic we get when 30,000-40,000 EEBO documents are measured against ground truth (see Appendix C). Page images that are deeply flawed, roughly one quarter of the set, will include many title pages, pages with columns, pages that have been skewed, and pages too full of noise to be readable by OCR (maps, images, torn and blotched pages, pages with significant bleed-through, images that are too light; see Appendix D for samples of uncorrectable digital images). As we ran these documents, our system recorded both skew and noise measures of each individual page: skewed pages will be de-skewed and re-run; the de-noiser scores suggest which pages have columns, and those will be segmented and re-run; pages that have insuperably high noise levels will be recorded in the database of unreadable pages, and libraries will be asked to scan copies of those texts if they have them. In other words, through simply continuing our work with the tools and processes that we have built during grant tenure, we will be able to improve the correctness of EEBO documents.

4) Large number of documents OCRed

In our grant proposal, we estimated that we would be able to OCR 23.7 million page images. We have far exceeded that number, thanks to the parallelization efforts of the Brazos High Performance Computing Cluster personnel and IDHMC's Lead Developer Matthew Christy, efforts that are made available to others in the form of the "eMOP Dashboard" application in the eMOP Github Repo. To date, we have run almost all of the 307,000 documents from EEBO and ECCO through our OCR and post-processing routines: we have OCRed approximately 41 million pages. In doing so, we have also collected a wealth of interesting, non-consumptive data about the distinguishing features of these early modern documents, e.g., data concerning common OCR engine errors, digitization issues, and the material description of the page. We intend to make this data available for analysis, in the hopes that future development work will help us more fully understand how our cultural heritage is being and should be preserved.

5) OCRed texts made available online

The EEBO and ECCO texts will be made available in TypeWright and Cobre. ECCO is loaded into TypeWright; EEBO is now loading and will be available in 18thConnect's TypeWright tool by December 2015 at the latest. Despite EEBO's 68% correctness overall, all EEBO documents are being uploaded into TypeWright for crowd-sourced correction. This is because anyone correcting a document can easily flag an unreadable page with the "Report this Page" button at the upper to mid-left. Eventually, we will hook that button up to the AWL editor, not yet available: there, people will be able to identify images and columns using the layout editor, again helping us figure out how and what to re-run. When users report to us a completely unreadable document, we will encourage them to look for it in Cobre, the edition-comparison tool that we modified during grant tenure. Editions are currently being uploaded into Cobre, a task which should be completed by early next year: when users click on the "Works" button, they see a list of work titles that, when selected, reveal all the editions of that title. A transcription page available in the comparison tool will allow users to copy transcriptions from one good edition to an edition with poor OCR and then make any changes necessary.

All text transcriptions generated by the eMOP workflow of pages with adequate de-noising and skew scores will be made searchable in 18thConnect (and soon, ReKN). For the sake of transparency, we will indicate precisely what percentage of the EEBO collection and what

percentage of ECCO texts are being searched at any given time. That percentage will rise as we continually re-run documents that have been pre-processed (de-skewed) and re-scanned (from libraries willing to help). These texts will be searchable by page, giving users the opportunity to determine how often a search term or phrase appears in a specific document. The OCRed text will also be returned to ProQuest for EEBO and Gale for ECCO on a regular schedule as corrections are made, in hopes of making these texts more fully searchable via their interfaces as well.

6) Font Training, Tools, and workflow made freely available

a) Font Training and Tools:

Over 40 font training sets and 6 tools are freely available on the eMOP Github site (see "G. Intellectual Property," below). Almost all of EEBO and ECCO—all the pages that could be processed—have been OCRed (41 million pages), and we are loading those pages plus their OCR into the ARC catalog and the 18thConnect TypeWright tool. They will be available for crowd-sourced correction using the TypeWright tool by the end of 2015—they are loading now, in a continuously running process. We have completed enhancing TypeWright (http://www.18thconnect.org/typewright/documents) and building Cobre (http://emop.tamu.edu/cobre). The Cobre tool, which we presented at the DH2014 Conference in Lausanne, allows for comparing editions and transposing correctly typed text from one edition to another while making manual corrections. More work remains to be done on the Aletheia Web Layout Editor (AWL), and the tools still need to be connected in one continuous workflow.

b) Workflow:

There are actually two workflows: the internal workflow for running digital images through our OCR process, and the externally-facing correction workflow (mentioned just above) that involves web-accessible tools used by "the crowd" of scholars willing to correct OCRed transcriptions. The external workflow currently includes an automated email notification system that allows us to know when a document has been corrected in TypeWright as well as when a page has been reported as a problem. We perform vetting and trouble-shooting on a person-by-person basis (see Appendix E, the *TypeWright Administrator's Handbook*, pp. 10-20). The version of external workflow in which users are automatically given other options (Cobre, AWL) when they encounter a problematic page image was not completed because we simply need to purchase more expert programming time in order to finish it. We had hoped to create an internal workflow completely loaded into Taverna that we could then allow others to download in order to create their own OCR systems. We were not able to get expert advice concerning the Taverna workflow organizing system (see D.3 below), but, beyond that, we found the complications in setting up our own internal workflow to be formidable and thus to go beyond systematizing in that way. In lieu of setting up a downloadable workflow, we have put on the eMOP website under the "Instruction" tab a series of videos and instructional materials to help people use the eMOP OCR workflow and tools (http://emop.tamu.edu/software), including the eMOP Dashboard (http://early-modern-ocr.github.io/emop-dashboard/) which runs OCR processes, sends OCRed documents through three post-processing analysis and correction routines, and saves those results while recording RETAS-Juxta-cl correctness scores. We have also put online numerous diagrams depicting a possible workflow for others to use when they acquire those tools (http://emop.tamu.edu/workflows). Because of how complicated the workflow is, and how dependent it is upon file and server locations, providing a Taverna workflow is ultimately less helpful than providing consulting services as people try to set up their own OCR systems on their

own servers. We have had to do a lot of troubleshooting in getting our own tools and files to work together, and we can discuss our strategies with others when they need it.

c) Dissemination and Educational Outreach:

In addition to all the conference presentations listed at the eMOP website, we have presented seminars on OCRing early modern documents at the Folger Shakespeare Library (Early Modern Digital Agendas, Summer 2015, http://folgerpedia.folger.edu/EMDA2015), for the Texas A&M Programming4Humanists class (Spring 2015, http://www.programming4humanists.org), and at a pre-conference workshop at the 2014 Society of American Archivists Annual Meeting. For the last three years, we have given full-day, pre-conference workshops on creating digital editions using TypeWright at the annual meeting of the American Society for Eighteenth-Century Studies (ASECS; http://idhmc.tamu.edu/asecs/2015-workshop-agenda/); we plan to include further workshops at the Shakespeare Association of America (SAA) and the Renaissance Society of America (RSA) as soon as the Renaissance Knowledge Network (ReKN) has been launched. For these courses and seminars, we have used all the "Instruction" documents and videos available on the eMOP website: http://emop.tamu.edu/software. Furthermore, to generate interest in crowd-sourced correction, we have started a "Liberate the Text" campaign, launched at DH2014 in Lausanne, Switzerland, and taken to many conferences among the Texas digital library community as well as Early Modern Scholars.

d) Sustainability:

In order to enhance and sustain eMOP, we have begun advertising OCR Services for themed library collections: http://idhmc.tamu.edu/arcgrant/devproject/libraries/. While we help for free any group who wishes to have our guidance in OCRing their materials, we will charge when doing it for them.

C. Setbacks or Challenges

The setbacks and challenges had to do with

1) training multiple OCR engines

At the outset of the grant, neither we nor the IMPACT team, nor the Google group with whom we consulted, really understood the best way to train Tesseract. The IMPACT group in Europe had used a tool for training their ABBY FineReader engine that was developed by IBM Haifa, a tool called "Concert." Concert allowed you to give the OCR engine as many instances of a letter as possible within a given document: it seemed to IMPACT, and consequently to us, that the more instances you gave your OCR engine, the better it would do in analyzing texts. With Tesseract, that turns out not to be true—in fact, the opposite is true. The more instances of "l" that you give to Tesseract, the more everything begins to look like an "l" to it. Tesseract needs just a few of the most perfect instances of a letter, and the letter cannot be fed to it in isolation but must be part of a page of text. It was to create faux early modern documents, documents woven together out of the most perfect instances of the letters, that we created FrankenPlus.

2) time for handling 41 million page images and texts

Running page images and post-processing routines can be very time consuming, as mentioned above: during February and March of 2015, we used more computing time on the Brazos High Performance Computing Cluster than the Physics High Energy Group at Texas A&M, and they have participated in finding the "God particle" (see Appendix B). The problem now is loading texts and page images into TypeWright, our crowd-sourced correction tool, and into the ARC SOLR server, for page-level full-text search. We are now continuously running these indexing and loading tasks but do not anticipate that they will be completed until the end of 2015. EEBO

and ECCO will be available in 18thConnect.org, in TypeWright, by December. We had originally hoped to have EEBO available in TypeWright via ReKN, the Renaissance Knowledge Network, but they are not yet launched.

3) EEBO/ECCO results versus scholar, librarian, and collector needs

Our focus on making EEBO and ECCO machine readable meant that we did not create a full OCR process: the images for EEBO and ECCO could not be manipulated or pre-processed, because they are so low in information density—they are black-and-white, very low resolution files. Most of the people who want to use our OCR workflow will have much higher quality images and will need to pre-process them: deskew, binarize, etc. We have created instructions for them on the Instruction section of the eMOP web site.

4) Data wrangling

We now have the most correct database of metadata containing a mesh of TCP, ECCO, EEBO, ESTC data, connecting all of these important digital resources in one place. This database is integral to the further study of these datasets and our early modern cultural heritage.

D. Significant Board, Management, or Staff Changes

1) James Moseley: we did not bring in the book historian James Moseley to help us with Cobre and font identification, but we did bring James Raven and Robert Hume, as planned.

2) Jacob Heil: Dr. Heil was supposed to do font research for year one of the grant. Because it took us the duration of the grant period to extract and regularize printer information from the EEBO and ECCO documents, he was unable to work on that research beyond helping us to select fonts for which we would make training sets. Dr. Heil was project manager for the first year of the grant; we did not budget any money to keep him for the last year and a half.

3) The National Library of the Netherlands (KB), the Improving Access to Texts (IMPACT) team: the head of the IMPACT team, Clemens Neudecker, left his job at the KB for another job, and so we did not use IMPACT's services except for one advisory meeting held at Texas A&M.

E. Plans and Goals for the Period Subsequent to the Grant

1) New OCR engines

Because of the amount of time it took to train Tesseract, we have only as yet trained that one OCR engine. We are very happy that the percentage correctness achieved through one open-source OCR engine so closely approaches Gale's OCR output, which again, made use of six top commercial engines plus a voting algorithm. (One can see a comparison of Gale's OCR and our OCR output here: http://juxtacommons.org/shares/RWOqVQ). But we will not of course stop there: we do intend to get our percentage correctness up into the 90th percentile by continually applying for grants to add other OCR engines into our workflow. The developers of Aletheia, our partners in the UK, have gone partway toward creating an export function that will allow us to use all the training that we created for Tesseract in other open-source OCR engines such as Gamera and OCRopus. Moreover, while we were working on the eMOP project, a team at Stanford University (subsequently spearheaded at Carnegie-Mellon) developed an OCR engine designed specifically for historical OCR. As one step in the process of setting up multiple OCR engines, we have just been awarded an NEH Implementation Grant, partnering with the University of Texas at Austin, in order to add Ocular to our OCR workflow ("Reading the First Books: Multilingual, Early-Modern Optical Character Recognition for Primeros Libros," PI Dr. Sergio Romero at Austin, Laura Mandell at A&M). Given that we have achieved 86% correct for ECCO using one engine, and close to that for roughly half of the EEBO documents, we think

that we can achieve a percentage correctness in the 90s by having libraries re-scan books and ingesting new images, developing a new OCR engine (one per year at the least, more if we receive more grants), crowd-source correcting, re-running the OCR process multiple times, and voting among engines. We project being able to achieve this level of correctness between 5 to 10 years, depending upon the levels of funding that we receive.

2) EEBO and ECCO documents uploaded and searchable in multiple ways

As mentioned in C2 above, we are uploading EEBO and ECCO documents into TypeWright and Cobre; as described in G below, as soon as these documents are loaded, they will be searchable in 18thConnect.org (and ultimately in ReKN) by page.

3) Automatic Font Detection

Another MS student working under Dr. Gutierrez-Osuna has taken up work on page detection where Anshul Gupta left off. He will be creating a tool that will allow us, with input from a small number of users, to separate Blackletter from Roman and Italic Fonts; it is a supervised training tool using machine-learning algorithms. This tool will also be made available on our Github site.

4) apply for grants to continue approaching 97%

As with the NEH Implementation Grant, we will continue to seek partners to propose grants in order to add new OCR engines to our suite of tools.

5) Apply eMOP workflow to other digital library projects

We are currently working with Notre Dame and the Visualizing English Print project, among others.

6) Re-OCR

We will run through our OCR process all pages on the Brazos Computing Cluster after applying pre-processing techniques, based on the triage system output generated for us by Dr. Gutierrez-Osuna at TAMU.

7) Further publications

Mandell, Laura; Grumbach, Elizabeth. "The Business of Digital Humanities: Capitalism and Enlightenment," *Scholarly Research and Communication* (forthcoming, 2015).

"The Business of Digital Humanities: Capitalism and Enlightenment."

Mandell, Laura; Christy, Matthew; Grumbach, Elizabeth. "Data Preparation" In *Computation for Literary Analysis*. Eds. James O'Sullivan and Ray Siemens. Forthcoming from Penn State Press, 2016.

F. Publications

Gupta, Anshul. *Assessment of OCR Quality and Font Identification in Historical Documents*. MS Thesis submitted to Texas A&M, August 2015.

Heil, Jacob, and Todd Samuelson. "Book History in the Early Modern OCR Project, or, Bringing Balance to the Fore." *Journal of Early Modern Cultural Studies* 13.4 (90-103).

Mandell, Laura. "Brave New World: A Look at 18thConnect." *Age of Johnson* 21 (2012): https://earlymodernonlinebib.wordpress.com/2012/06/06/laura-mandells-brave-new-world-a-look-at-18thconnect/

---. *Breaking the Book: Print Humanities in the Digital Age*. Wiley-Blackwell Manifestos. 2015 (contains some discussion of data correctness and eMOP).

---. "Digitizing the Archive: the Necessity of an Early Modern Period." *The Journal for Early Modern Cultural Studies* 13.2 (Spring 2013): 83-92.

---, et. al. "Navigating the Storm: IMPACT, eMOP, and Agile Steering Standards." Under consideration at *Digital Scholarship in the Humanities* (formerly *LLC*).

Torabi, Katayoun, et. al. "Early Modern OCR Project (eMOP) at Texas A&M University: Using Aletheia to Train Tesseract." Proceedings of the DocEng 2013. 4/13/09. http://dx.doi.org/10.1145/2494266.2494304.

G. Intellectual Property

We produced 11 tools during grant tenure: the eMOP dashboard to coordinate an OCR workflow, FrankenPlus, a denoising algorithm that works with OCR outputs, two ground-truth comparison tools (Juxta-Cl and RETAS), a page evaluator and page corrector, Cobre, the Aletheia Web Layout editor, and TypeWright. All these tools have been assigned an Apache License Version 2.0. We have created training sets for using Tesseract on early modern documents. Each of the font sets have been assigned a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. All these licenses are published on the eMOP GitHub site (http://early-modern-ocr.github.io/) along with the tools and fonts, as a "LICENSE.md" file. Our publishing imprint database is freely available for download (http://early-modern-ocr.github.io/ImprintDB/) with an Apache License Version 2.0, as will be our uncorrectable pages database, though we would like to set them up as web accessible databases as well.

H. Budgetary Variances

Two small variances are: we spent a little less on personnel salaries than we expected, and a little less on personnel benefits; these small variances were not the result of any changes in staffing but simply the result of over-estimation at the outset. The largest budgetary variance is the discrepancy between the budgeted amount for "Software" and the amount actually spent, a $5,000 variance. Via an email dated 5/12/15, the Mellon Foundation gave me permission to vary the Software costs and explain the variance here. The additional programming costs that we had budgeted turned out to be more than we expected: we paid Performant Software Company, using interest accrued by the grant funds, in order to write the programs we needed to a) load EEBO into TypeWright and EEBO metadata into our SOLR server (rake tasks); and b) make the EEBO and ECCO OCR searchable via our Lucene search engine at the level of the page: currently, our texts are indexed in SOLR by full document, not by page. Performant successfully created the rake task – we are running it now, and will be until December, according to our estimates – and successfully adjusted our search capacity for both EEBO and ECCO. That change will not be visible in our search interface until we have re-indexed ECCO and indexed EEBO for the first time, but Performant did successfully complete the job.

I. Signature of Principal Investigator

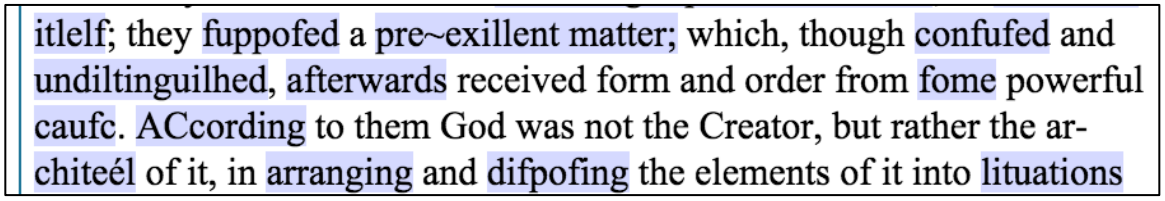
Signature:

Name: Laura Mandell
Title: Professor of English; Director, Initiative for Digital Humanities, Media, and Culture
Date: September 30, 2015

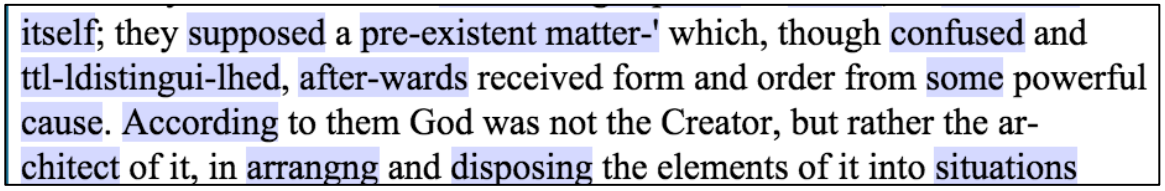
Appendix A

Images of raw OCR results using our training libraries, before applying post-processing routines, as visualized in the Juxta Commons Diff engine.

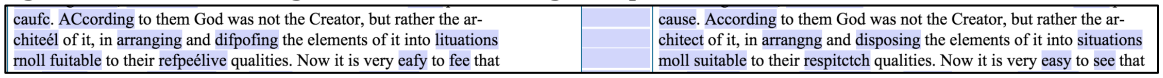
Default:


itlelf; they fuppofed a pre~exillent matter; which, though confufed and undiltinguilhed, afterwards received form and order from fome powerful caufc. ACcording to them God was not the Creator, but rather the ar-chiteél of it, in arranging and difpofing the elements of it into lituations

Using the training libraries that we used for the EEBO Collection:

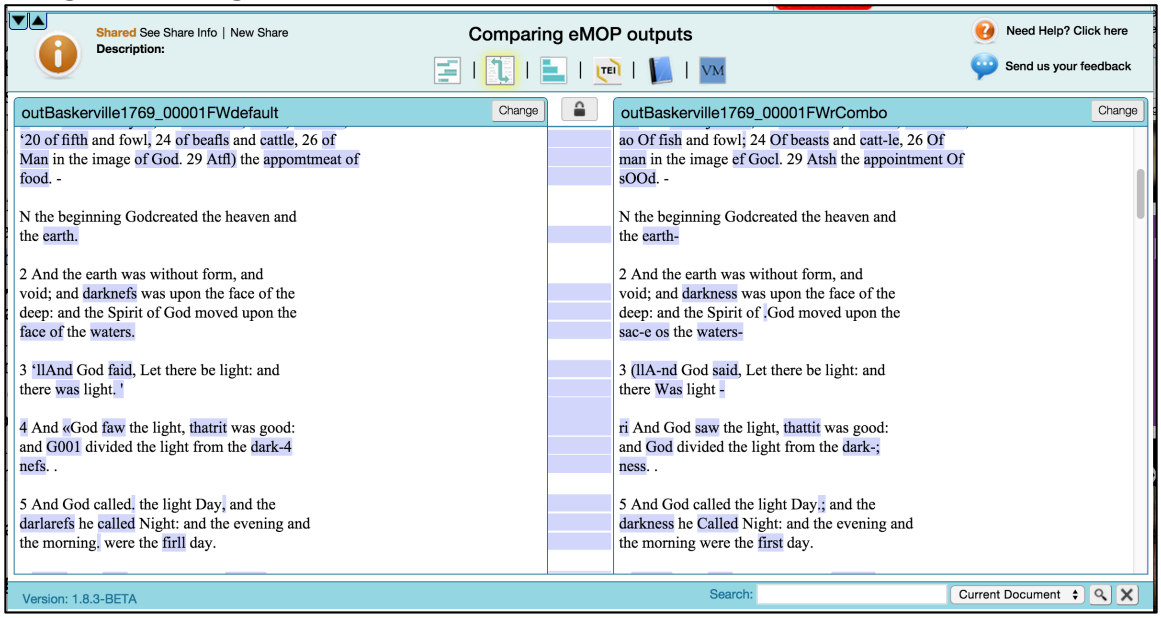

itself; they supposed a pre-existent matter-' which, though confused and ttl-ldistingui-lhed, after-wards received form and order from some powerful cause. According to them God was not the Creator, but rather the ar-chitect of it, in arrangng and disposing the elements of it into situations

Again, default training and EEBO training compared:



caufc. ACcording to them God was not the Creator, but rather the ar-chiteél of it, in arranging and difpofing the elements of it into lituations rnoll fuitable to their refpeélive qualities. Now it is very eafy to fee that	cause. According to them God was not the Creator, but rather the ar-chitect of it, in arrangng and disposing the elements of it into situations moll suitable to their respitctch qualities. Now it is very easy to see that

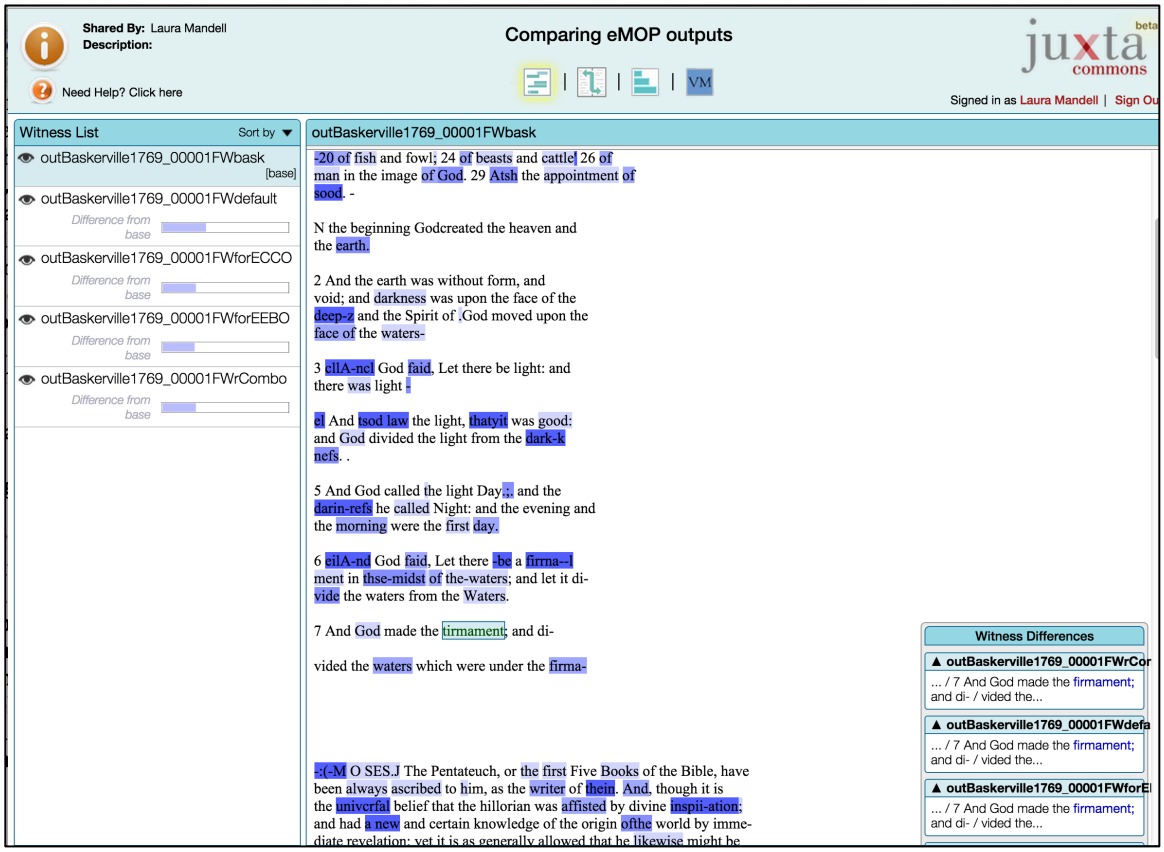
Using the training libraries that we used for the ECCO Collection:



Notice in verse 2, the default is better in the case of "face of": "God moved upon the face of the waters." You can see that we got "sace os" because, with increased capacity to detect the long-s, there is also the increased risk that f's will be

misinterpreted as “s.” The “f” isn’t always misinterpreted as “s” in our historically-specific libraries—notice “Of fish and fowl” at the top of the pages, as well as “first day” at the very end.

Here Juxta Commons shows many of the different font training libraries that we created (all available on our eMOP GitHub Repo, https://github.com/Early-Modern-OCR/TesseractTraining, indexed on the eMOP site, http://early-modern-ocr.github.io/TesseractTraining/):



Here, in the lower left, you can see that three out of the four training libraries loaded into Tesseract read the word “firmament” properly: we now know that it is possible to use our voting algorithm upon different training-set results as well as upon outputs from different OCR engines. Paradoxically, the only engine that misread the word comes from the training set that was built *using that very document*. What we now know is that there is as much variance within a printed text made from a specific font as there is between printed texts – within a specific range of font types.

These OCR samples are available for viewing publicly on JuXta Commons:

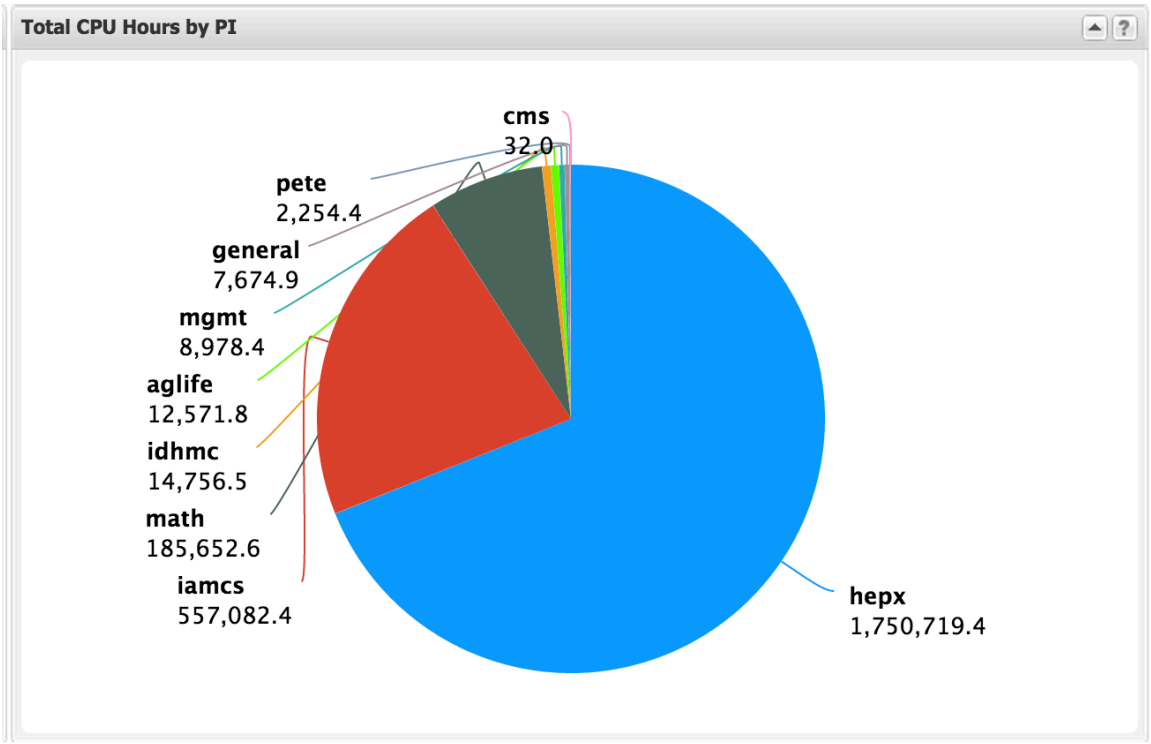
> http://juxtacommons.org/shares/YjG1Ku
>
> Training Libraries Used for texts in JuXta Commons:

SC8b-R7-D2b-forECCO => “out Baskerville1769_00001FWforECCO.txt"
RI5Combo-R8-D2b => “out Baskerville1769_00001FWrCombo.txt”
BASK1769RI5Combo-R8-D2b => “out Baskerville1769_00001FWbask.txt”
SC8b-R8-D2b-forEEBO => “out Baskerville1769_00001FWforEEBO.txt"

Appendix B:

Usage of High Performance Computing Cluster for the eMOP Project:

January 2015:



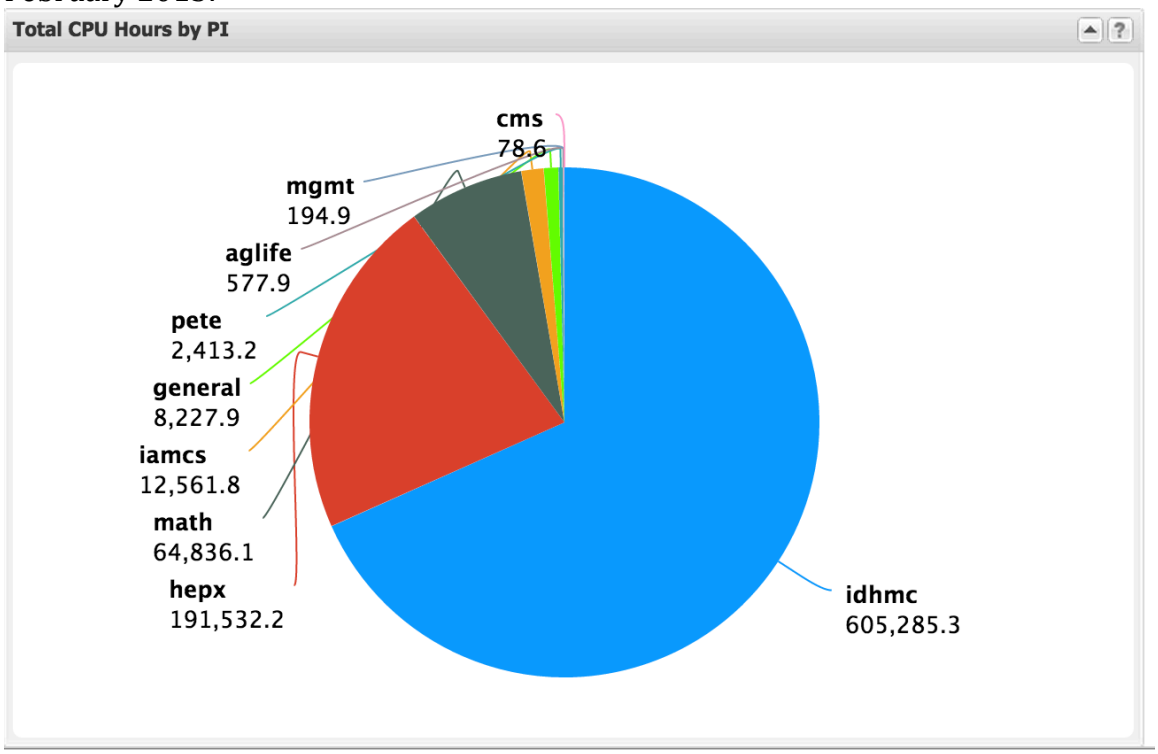
Notes about groups using the CPUs:

"**idhmc**" = The Initiative for Digital Humanities, Media, and Culture; we are running the eMOP project.

"**hepx**"= High Energy Physics Research (http://physics.tamu.edu/research/he.shtml):

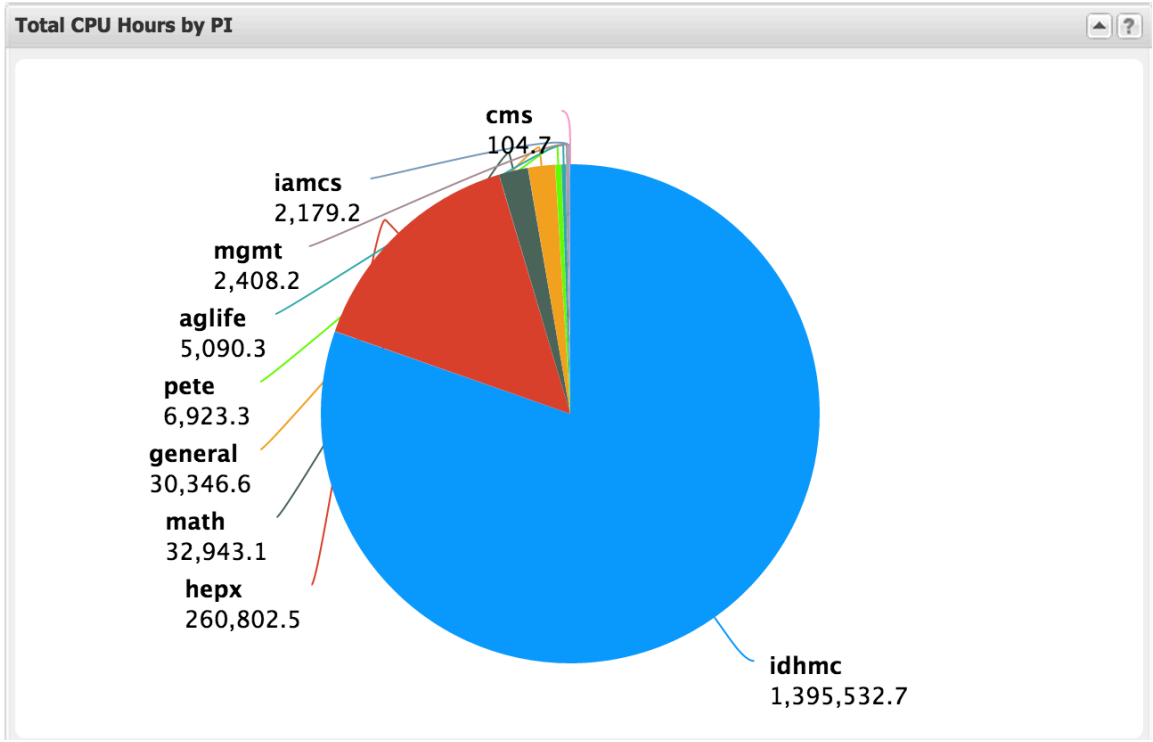
The High Energy Physics group at TAMU participated in the recent discovery of Higgs boson, the "God particle." "**pete**," also listed above, is part of this group.

February 2015:



The idhmc overtakes hepx (and pete)!

March 2015:



The idhmc overtakes hepx for the second month in a row, finishing our OCR run in late March.

Appendix C

EEBO and ECCO correctness measures were made using "groundtruth," keyed texts from the Text Creation Partnership, a total of **1,815,556*** pages.

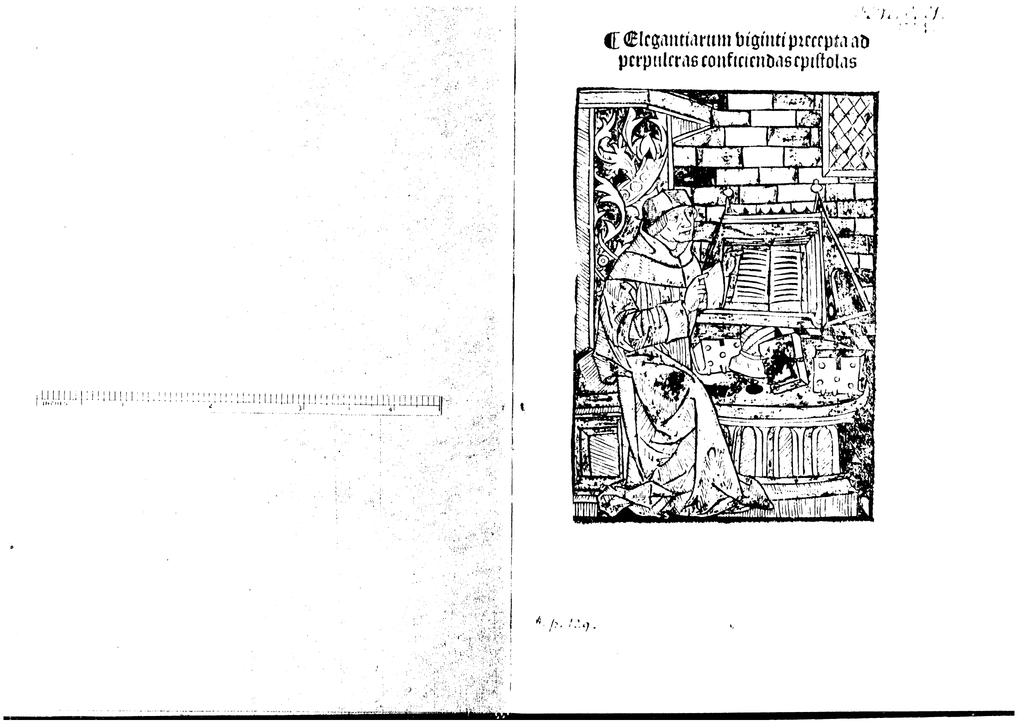
The bulk of them, approximately 86% or 1,561,378, are EEBO documents;
The remainder, approximately 14% or 254,178, are ECCO documents.

Percentage Correctness Ranges	Number of Pages	Percentage of total Pages with Groundtruth at these correctness levels
90-100%	354,358	**19.5%**
80-89%	524,825	**28.9%**
70-79%	316,511	17.4%
60-69%	193,708	10.7%
50-59%	125,060	6.9%
40-49%	85,875	4.7%
30-39%	61,992	3.4%
20-29%	48,539	2.7%
10-19%	43,776	2.4%
0-9%	60,912	3.4%
Total	1,815,556	100%

48.4% of the pages, close to half, are in the 80 to 100% correctness range.

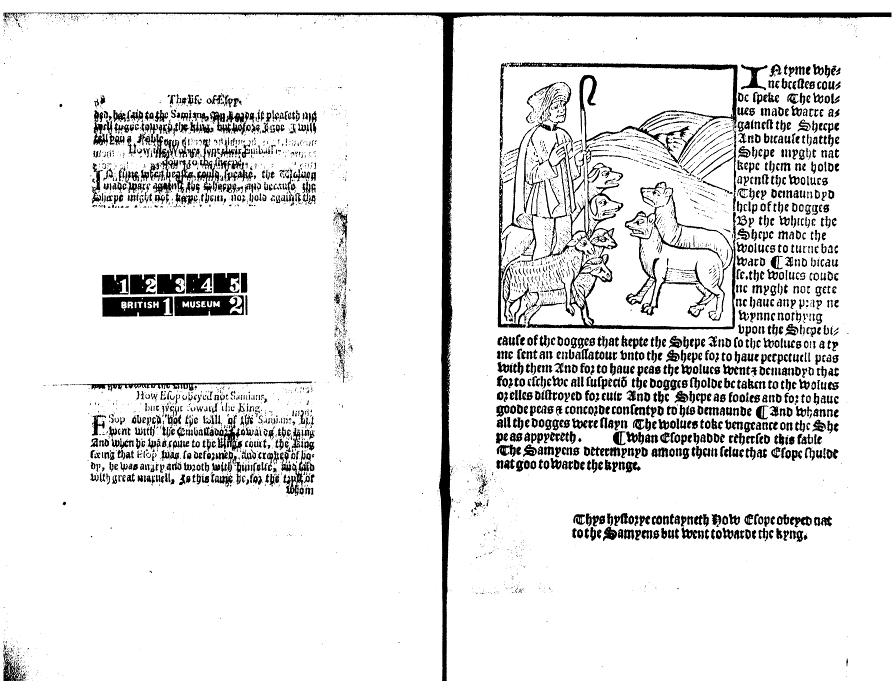
*We were given 2,345,524 pages of ground truth but only 1,815,556 were used to measure correctness. We are not yet sure why. We are working to address this issue and will post updated figures on the eMOP website as soon as we can.

Appendix D
Sample Page Problems



Elegantiarum viginti precepta ad perpulcras conficiendas epistolas

Image pages



The life of Esop.

ded, he said to the Samians, My Lords it pleaseth me well to goe toward the King, but before I goe I will tell you a Fable.

How the Wolves sent their Embassadours to the sheepe.

In time when beasts could speake, the Wolues made warre against the Sheepe, and because the Sheepe might not keepe them, nor hold against the

How Esop obeyed not Samians, but went toward the King.

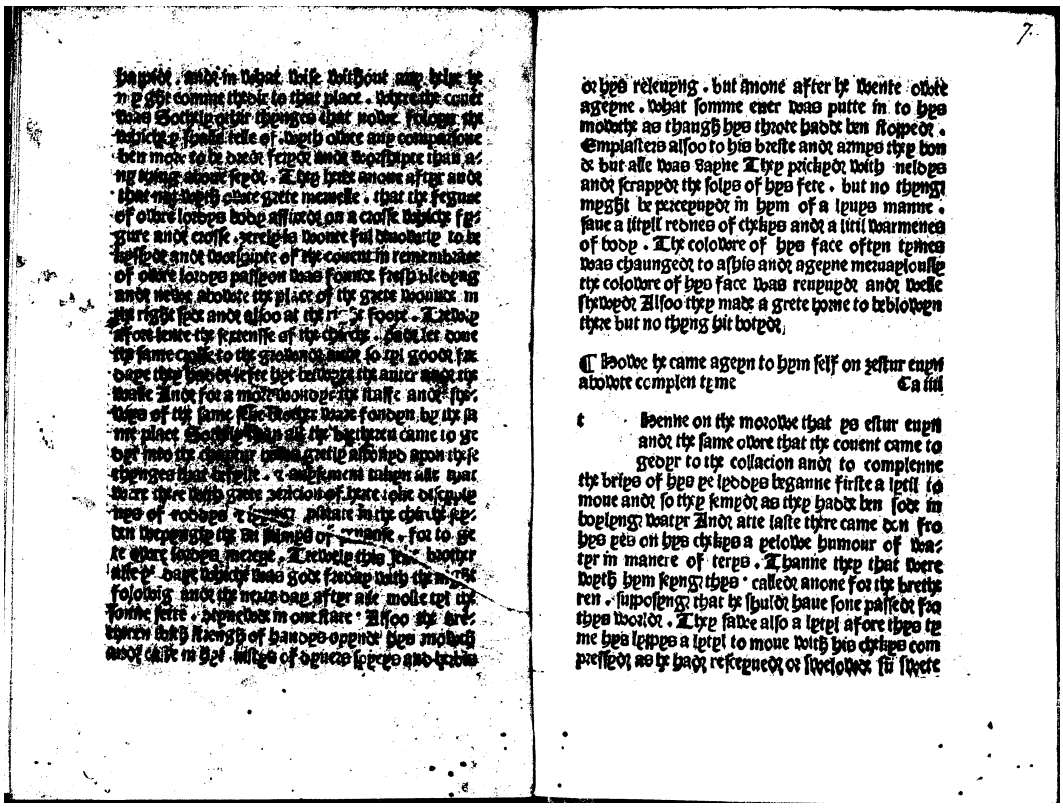
Esop obeyed not the will of the Samians, but went with the Embassadours towards the King. And when he was come to the Kings court, the King seeing that Esop was so deformed, and crooked of body, he was angry and wroth with himselfe, and said with great marvell, Is this same he, for the which of whom

BRITISH 1 MUSEUM 2

In tyme whanne beestes coude speke The Wolues made warre agaynest the Shepe And bicause that the Shepe myght nat kepe them ne holde ayenst the Wolues They demaundyd help of the dogges By the whiche the Shepe made the Wolues to turne bacward And bicause the Wolues coude ne myght not gete ne haue any pray ne wynne nothyng vpon the Shepe bicause of the dogges that kepte the Shepe And so the Wolues on a tyme sent an enbassatour vnto the Shepe for to haue perpetuell peas with them And for to haue peas the Wolues went & demandyd that for to eschewe all suspectiō the dogges sholde be taken to the Wolues or elles distroyed for euir And the Shepe as fooles and for to haue goode peas & concorde consentyd to his demaunde And whanne all the dogges were slayn The Wolues toke vengeance on the Shepe as appyereth. Whan Esope hadde rehersed this fable The Samyens determynyd among them selue that Esope shulde nat goo towarde the kynge.

Thys hystorye contayneth How Esope obeyed nat to the Samyens but went towarde the kyng.

Mixed image/text



Bleedthrough

A Prognostication.

Dayes.

25 at Whitchurch, count. Waterford. / at Iland Kierie, count. Kierie. / at Ballanrobe, count. Mayo. / at Newtowne, at Omye. — S. Iames.
at Dundanouan, count. Corke.

26 at Tomhaggard, count. Weixford. / at Galballie, count. Limbrick. — S. Annes day.

Fayres in August.

1 at Belfast, count. Antrim.
at Ballikillie, Queenes countie.
at Knocke Anie, count. Limbricke.
at Wicklow, count. Wicklow.
at Kil-cocke, count. Kildare.
at Carrig-drumruske, count. Letrim.
at Kil-kenic, count. Westmeath.
at Afaune, count. / at Tallow, count. / at Kil-macktomas, count. — Waterford.
at Slygo, count. Slygo.
at Killishanan, count. Dunnagall.
at Inchiloghfe, count. Farnamanay.
at Ryres, count. Mayo.
at Muckarcke, count. Typ. — Lammas day.

2 at S. Tannaues greene, count. Weixford.

3.&4. at Innifkilling.

10 at Knocklong, count. Limbricke. / at Palmerstowne count. Dublin. / at Ardee, count. Lowth. — S. Laurence.
at Moy-mors, count. Westmeath.

at Rakeele,

Skew

Faded:

fore, by the Management of those Pyrating Fellows, rather than Sailors, the *Indigo* I could not take on Board; nor could I trust any of them with the Sailing such a Ship, as could neither keep me Company, nor make a tight Steerage-way.

In p. 33. He says truth as to Mr. *Observator's* Ship and the Monkey that was left in her; *but as to Firing at any thing but the Ship*, and the whole Crew, who before we could make up to her, had made into their Boats, and nothing beside the Cargo and the *Monkey* were indeed on Board, *which should have been brought along, had we not Observators enough at Home.*

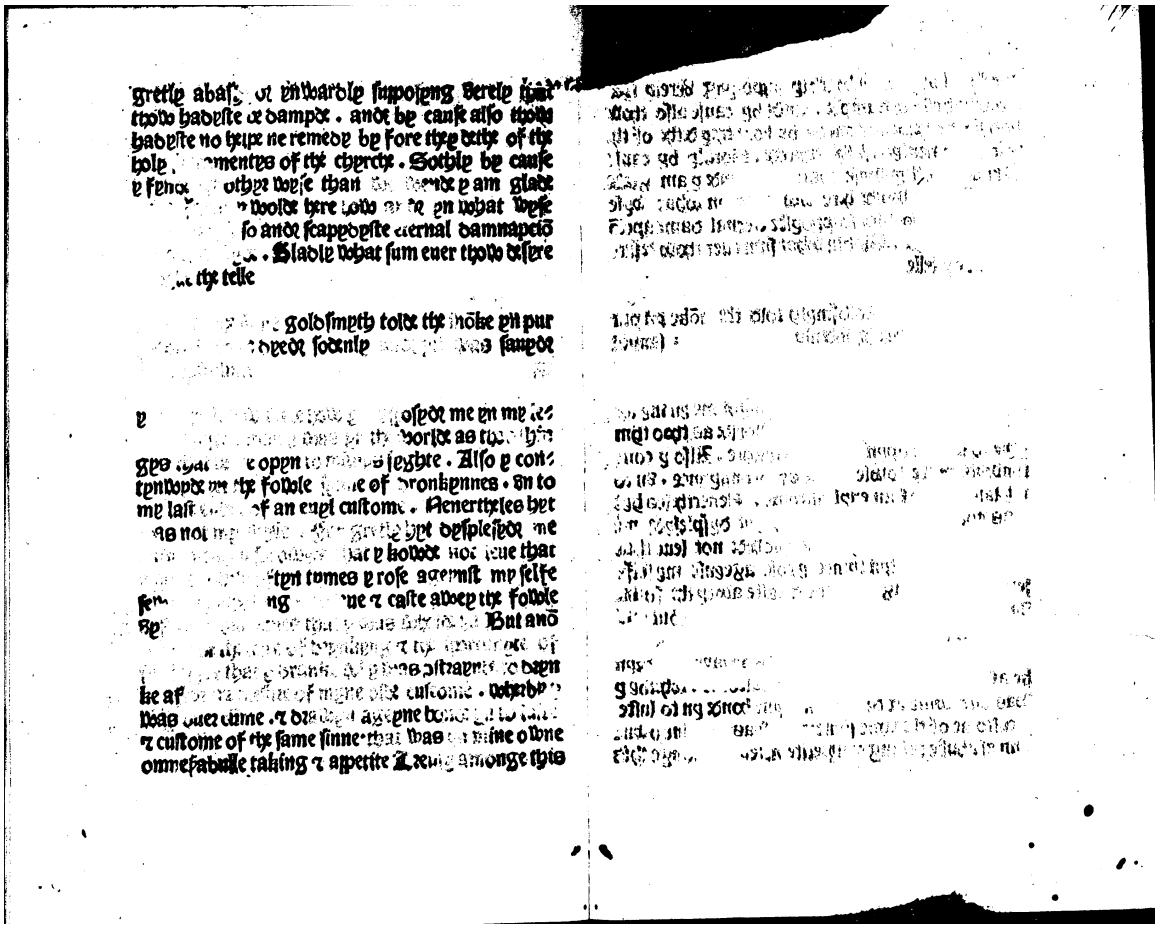
The next thing they Charge me with, especially Mr. *Funnell* in p. 41. in the Gulph of St. *Michael*, says, when we were in the Barque last taken, and Indian Canoe sailed by, and that I order'd her to be Fir'd upon. *The Contrary is very plain, for when I saw some of them that had Fir'd without my Orders, I was very Uneasy and Troubled at it, knowing the Consequence of it.*

The Second Thing that's Material is, that the Ship mentioned by Mr. *Funnel* in p. 45. 46. where he seems to Exclaim 8000 Dollars was Hid in the Run of the Ship, *and I slipt the Opportunity of taking it, and turning her Adrift*; So this I answer, *That I had evident Proof she had Landed her Money at Truxillo*; and as to Provisions, we took as much as would provide us for One Year, and much longer, if well manag'd, that was, what our Ship could well Stow, and this was the *Steward* and the *Crew's* Calculation. Now as to a Report that they make about Town 8 or 9000 Dollars that should be offer'd for her Ransom; First I had no convenient Road to Ride in, and the strong *Southerly Winds* were set in, and so if I had Loyter'd for her, must certainly been Imbay'd for 3 or 4 Months: Besides the *Winds*, thro' the Treachery of the *Spaniards*, I have had the Experience of it before in a like Case, Riding those for Ransom with Capt. *Swan* and Capt. *Davis*, for instead of keeping their Faith, they came off with a Fire-ship in the Night and 14 *Periagoes*; and tho' we had much better Crew, and Stouter Men. We came narrowly off.

Now, the they are Judges in my Case and Conduct, a Parcel of Fellows who were perpetually drunk, and very fit, you'll say, for Guarding a Ship in the Night, or being kept in any Decorum.

Again in p. 46. 47. Whereas Mr. *Funnell* frequently would Insinuate, that I could agree with no body; and so says that parted this way with Capt. *Spradling*. I say, I Depos'd Capt. *Spradling*; say not only that, but at *Juan de Fernandos*, when all his Men left him, I reconcil'd them and him again; there for

Fading



Fading and torn pages

nes; al our bleſsings, health, wealth, and proſperitie to the increaſe of Satans kingdome, are there abuſed: that not onely they are tearmed, as of late The ſchoole of abuſe, *by one* [a]; The ſchoole of Bauderie *by another* [b]; The neſt of the Diuel, and ſinke of al ſinne, *by a third* [c], *ſo long agoe,* The chaire of peſtilence, *by Clement Alexandrinus* [d]; *by Cyril* [e], *and Saluianus* [f] The pompe of the Diuel; the ſoueraigne place of Satan, *by Tertullian* [g].

And albe I cal them, A ſecond and third blaſt, &c. yet do I not ſo, as though there were no moe blaſtes, or dehortations fro them, or inuectiues againſt them beſide. For in al ages the moſt excellent men for learning haue condemned them by the force of eloquence, and power of Gods worde (as I am to proue vpon anie good occaſion offered). But ſo do I tearme them

[a] M. Gosſon in his inuective against plaies, called The Schoole of abuſe.

[b] 3. Blaſt of retrait fro plaies.

[c] M. Spark in his rehersal ſermon at Paules croſſe, 29. of April. Ann. 1579.

[d] *Clement. Alexand. li. 3. Pedag. cap. 12.*

[e] *Cyril. Catech. 1. Myſtagogica.*

[f] 2. blaſt of retrait fro plaies.

[g] *Tertul. lib. de ſpectaculis.*

Bleedthrough and fading

TypeWright Administrator's Handbook

Version 1.4 (Anne Arundel Locker-Thaddeus, Liz Grumbach, Tim Duguid)

Table of Contents:

I. Introduction

You have been assigned the task of administrating the TypeWright tool for correcting OCR that has been generated from scanned images of documents. This handbook is designed to give you the tools to manage TypeWright.

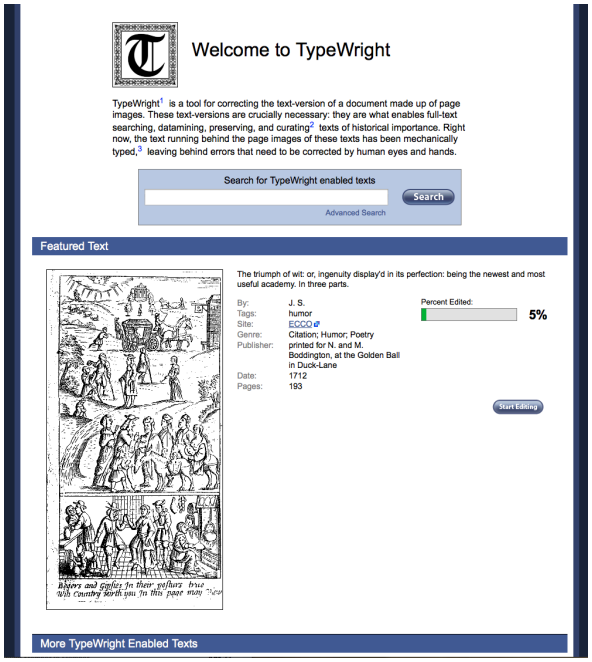


Figure 1 TypeWright Home

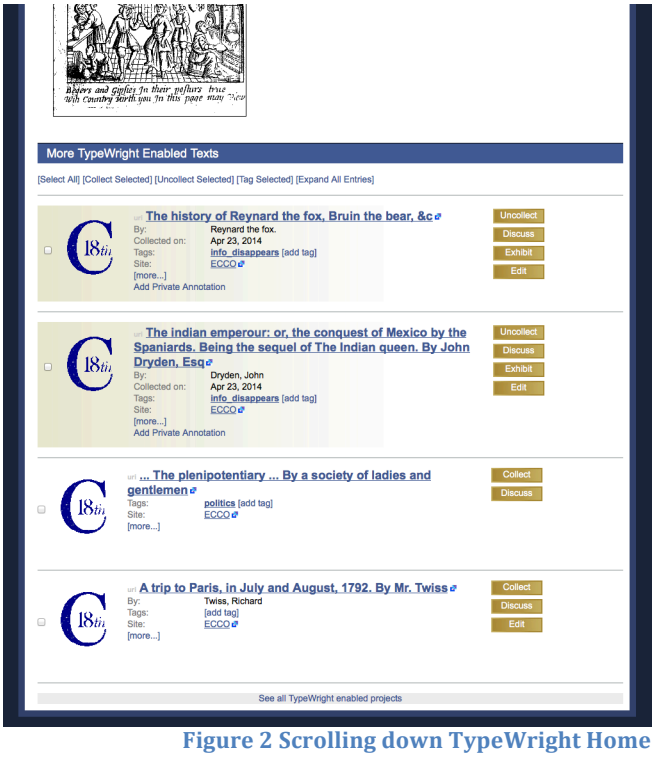


Figure 2 Scrolling down TypeWright Home

II. Administration Dashboard

As a site administrator, you will be able to control many aspects of TypeWright using this dashboard. You can access these controls through your site's own administration dashboard; Figure 3 shows the access to the TypeWright dashboard in 18thConnect.

Figure 3 Getting to the TypeWright Administrators' Dashboard

The TypeWright administration dashboard has three starting options: "Main," "Documents," and "Users."

A. Main

From this page you can get the existing text file that underlies a TypeWright enabled document, you can assign and work with the past and current TypeWright Featured documents, you can create new TypeWright Featured documents, and you can access some statistical data about TypeWright use.

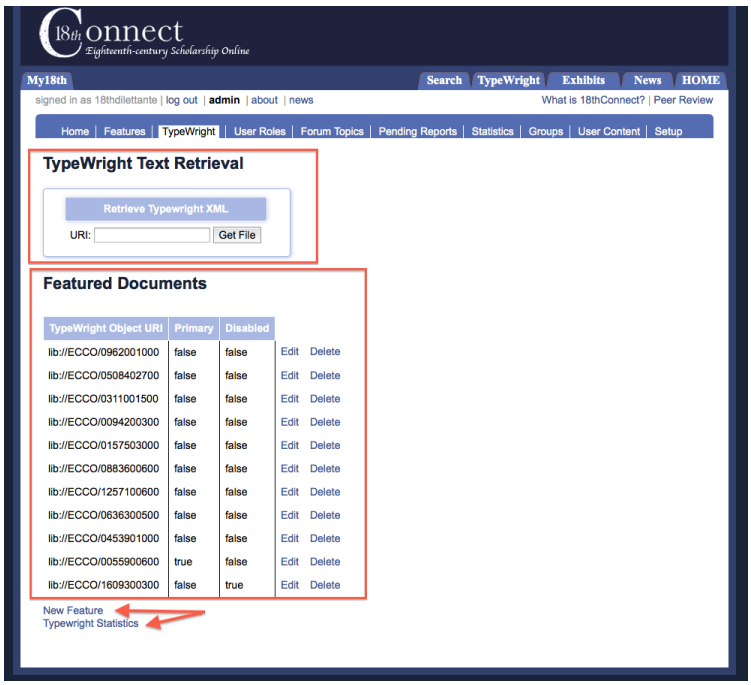


Figure 4 The TypeWright: Admin: Main page with sections and links highlighted in red.

1. TypeWright Text Retrieval Box

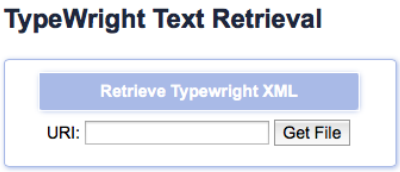


Figure 5

To access the text underlying a document, you will need to know the URI for that document, which you can find in either of two places:

- From any list of documents such as a search or your collected items.
- From the URL for the Document Home page for that document, conveniently at the end of the URL following the "uri=" statement.

Copy and paste the URI into the text box and click on the "Get File" button. Like with any download, you will be prompted to open or save the XML file as you prefer.

2. Featured Documents List Box

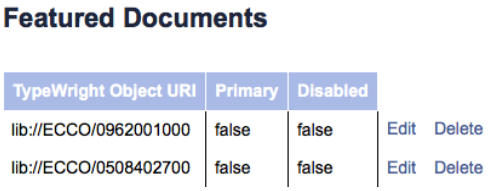


Figure 6

This list accumulates all the previous featured texts along with the current text.

a. The current featured text will show as "true" in the "Primary" column, and the other documents will show as "false". When a new feature is added, it appears at the bottom of the list.

b. The "Edit" link produces a pop-up box that is the same "Feature" box as that produced by the "New Feature" (see II/A/3. below).

- The text box is live, so the URI can be changed if necessary.
- Check boxes control the "Primary" and "Disabled" status, with unchecked boxes showing as "false" in the list, and checked boxes showing as "true"; to retire a feature text once you have chosen a new one, un-check the "Primary" box.
- In some cases, you may want to remove a past featured text from the public list on the TypeWright Home page, but leave it in the list on the Administrator's TypeWright Main page; to do this, check the "Disabled" box here.

c. The "Delete" link will produce a box asking if you really want to delete the item; deleting will remove the item from both the public and the administrator's list of previous featured documents.

3. New Feature

First, look for this as one of two small links in the bottom, left-hand corner of the page. Clicking on this "New Feature" link produces the "Features" box, but with an empty "Object's URI:" text box.

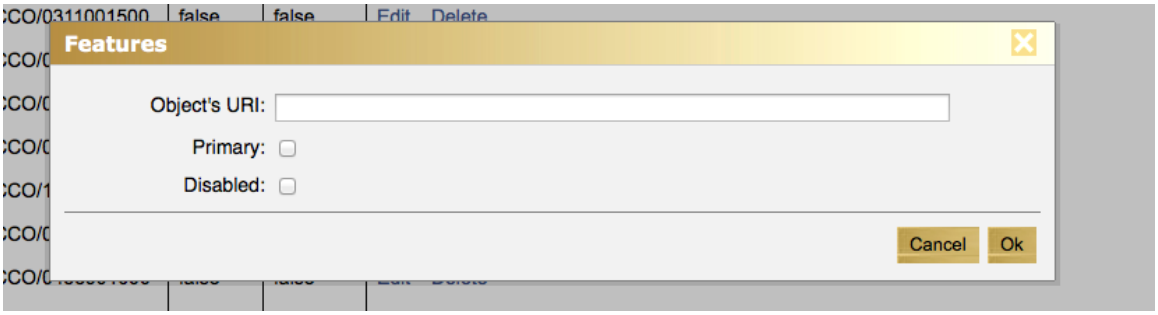


Figure 7 The "Features" pop-up awaiting the URI for the new "Featured Text"

You will need to paste or type your new feature text's URI into the text box. (See 2.1.1)

You will check the box marked "Primary" so that this document will appear as the "Featured Text" on the TypeWright home page. Remember to edit the previous document by un-checking its "Primary" box!

4. TypeWright Statistics

Look for this as the other of the two small links in the bottom, left-hand corner of the page. Clicking on this link gathers live statistics—which can take some time—and displays them in a new page.

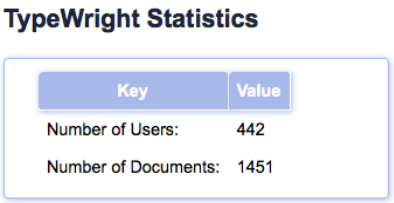


Figure 8

The counts of "Users" and "Documents" on this page are counted using broader definitions, and thus are higher than the counts available from the "Documents" page and the "Users" page, the subjects of the next two major sections.

B. Documents

"Corrected TypeWright Documents"

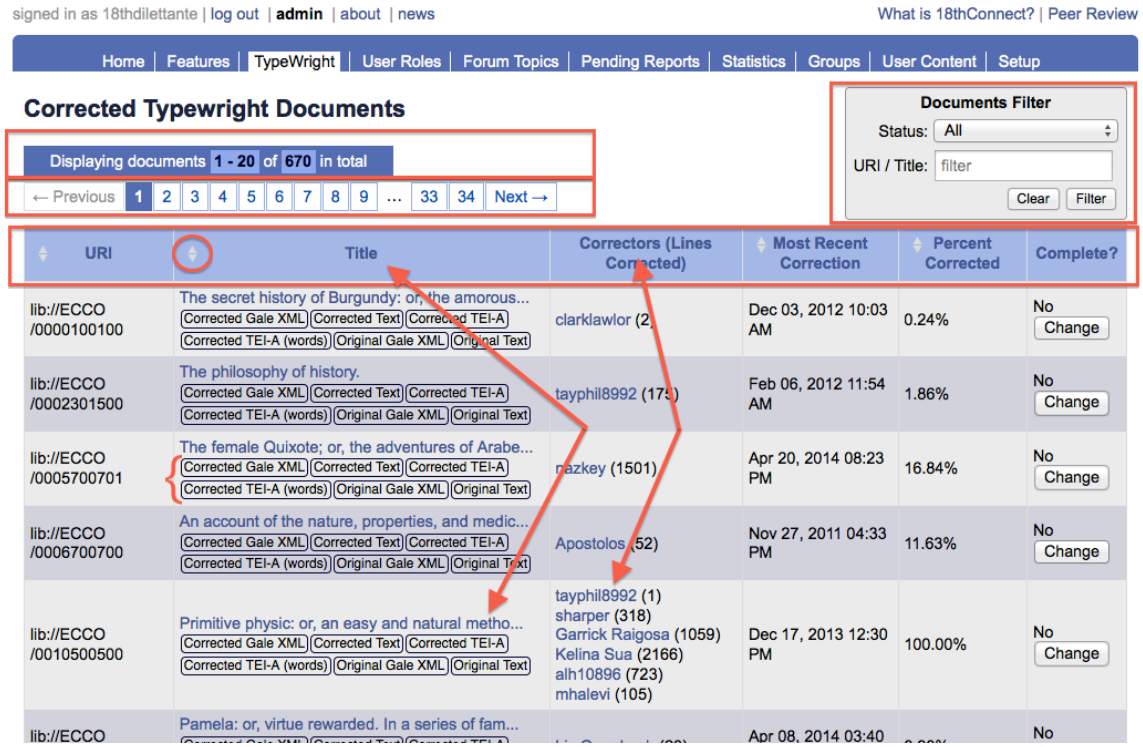


Figure 9 The TypeWright: Admin: Documents page with details highlighted in red

This page displays the contents of the TypeWright database of corrected documents organized by individual documents.

1. THE UPPER PART OF THE PAGE (indicated with the topmost three red rectangles in Figure 9), above the actual database listing, has information, navigation links for the pages of the listings display, and list filtering capabilities by status of completion, by title, or by individual document:

- Displaying Documents __ - __ of ____ in total (the actual numbers here will be determined by how you have or have not filtered your list)
- Navigation links to pages of the list
- Documents Filter
  - Status | [drop-down: All; Not Complete; User Complete; Confirmed Complete]
  - URI/Title | [text box]
  - Clear button | Filter button

2. THE ACTUAL DATABASE LISTING HEADER (indicated with the fourth red rectangle in Figure 9) uses the following column names, and can be sorted (ascending or descending) on columns that display the sort icon (circled in Figure 9):

- URI [sortable]
- Title [sortable]
- Correctors (lines corrected)

- Most recent Correction [sortable]
- Percent Corrected [sortable]
- Complete ?

3. IN ADDITION to the information indicated by the headers, the listed information includes several links. The listed "Title" (arrow on the left in Figure 9) links to the Document Home page in TypeWright and in the "Correctors" list each name (arrow on the right in figure 9) links to a page of statistics for that user. Under each Title you will see buttons (indicated in Figure 9 with a curly bracket) that allow you additional access to the underlying text in several different forms:

- Corrected Gale XML
- Corrected Text
- Corrected TEI-A
- Corrected TEI-A (words)
- Original Gale XML
- Original Text

4. IN THE "COMPLETE?" COLUMN, you also have a "change" button for changing the status regarding the completion for the document.

C. Users

"Corrected TypeWright Documents" is also the title for this page, even though this page displays the contents of the TypeWright database of documents organized by individual Users who have made corrections.

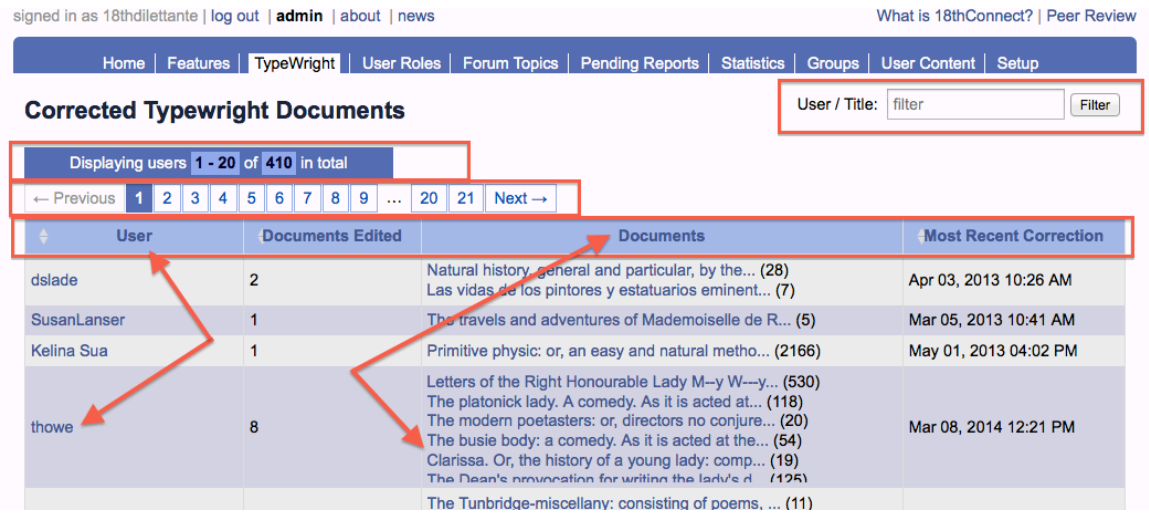


Figure 10 The TypeWright: Admin: Users page with details highlighted in red

1. THE UPPER PART OF THE PAGE (indicated with the topmost three red rectangles in figure 10), above the actual database listing, has information, navigation links for the pages of the listings display, and list filtering capabilities by user:

- Displaying Users __ - __ of ___ in total (again, the actual numbers here will be determined by how you have or have not filtered your list)
- Navigation links to pages of the list
- User/Title: | [text box] || Filter button

2. **THE ACTUAL DATABASE LISTING HEADER** (indicated with the fourth red rectangle in Figure 10) uses the following column names, and can be sorted (ascending or descending) on columns that display the sort icon:
    - Users
    - Documents Edited
    - Documents
    - Most Recent Correction

3. SEVERAL OTHER LINKS not indicated by the column names are available, as they are in the "Documents" list:
    - The listed "User" name (marked with the red arrow on the left in Figure 10) links to a page of statistics for that user,
    - The "Document" name (marked with the red arrow to the right in Figure 19) links to the Document Home page within TypeWright.

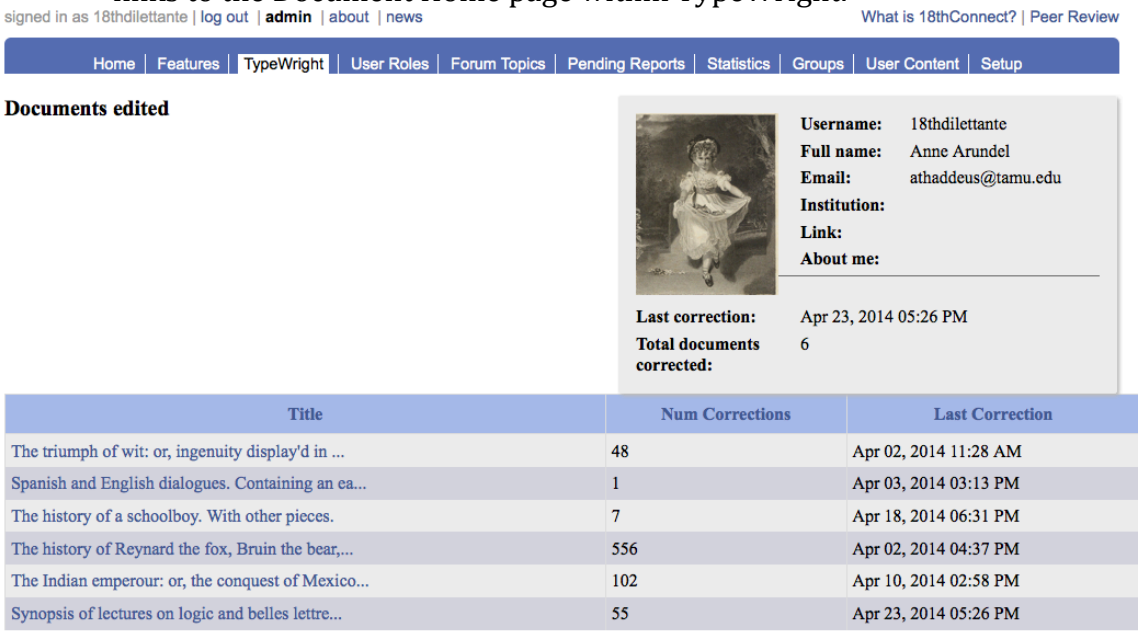


Figure 11 The page of user statistics linked from the Correctors columns of the TypeWright: Admin: Documents page and the Users column of the TypeWright: Admin: Users page

III. Information for Administrators on the TypeWright editing pages

As an administrator, the editing page that you see will include more information than the pages seen by regular users. Below the editing instructions you will have a section titled "DEBUGGING INFO: Word statistics by occurrence" (see Figure 12). This section has dropdown lists of the instances of "odd punctuation" and the words from the original OCR text of the document in five self-explanatory sections. The information is provided for the current page and for the first 100 pages of the document. You can use this information to make whole-page checks on the interim progress of corrections to a document, or as a supplement to the Workflow for Completed Document Evaluation in the next section.

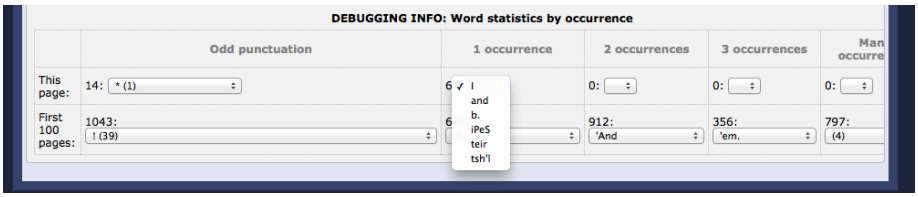


Figure 12 Debugging Info from Edit screen with the "This Page:" "1 occurrence" menu dropped

IV. TW completed documents evaluation

A. Workflow

(available from Basecamp Docs and LucidChart)

This flowchart lays out the path a document will follow through its evaluation, beginning when a user marks the document “complete,” and terminating in a decision whether to send the corrected text to Gale, or to reactivate the document in TypeWright for further editing.

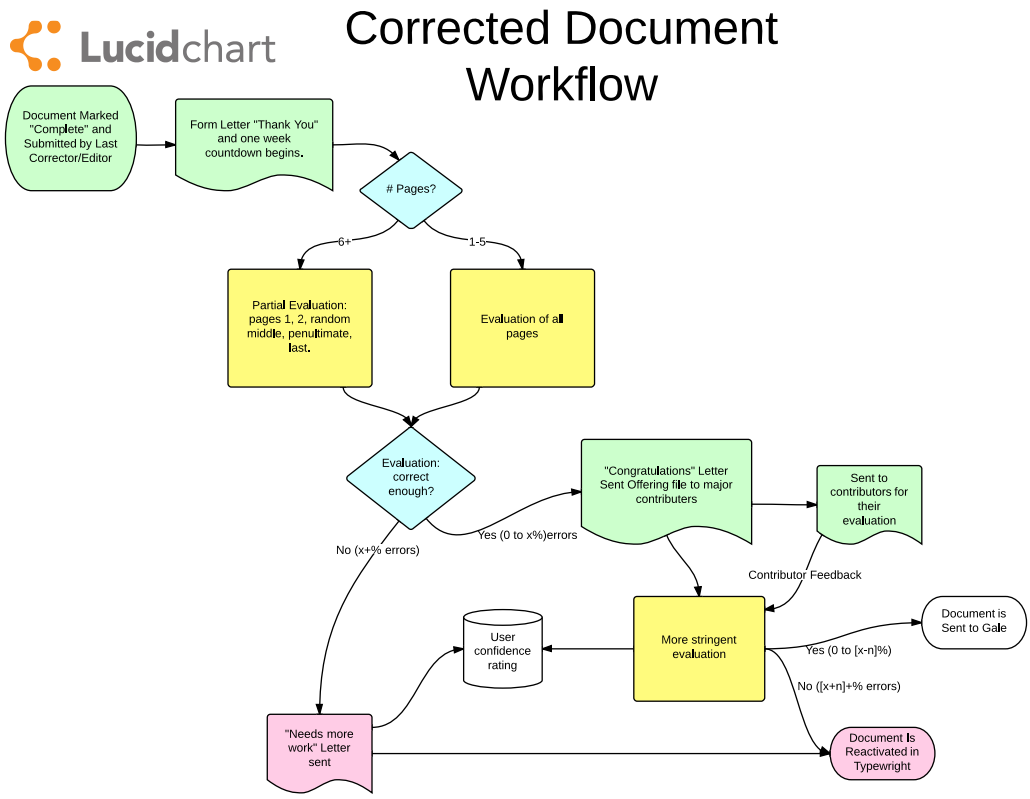


Figure 13 Workflow from "Document Marked Complete..." to either of two final terminations

B. Evaluating TypeWright User-Completed Documents

(most current edition is available from Basecamp Docs)

When a TypeWright user marks a corrected document as complete, it is the administrators' job to evaluate said document to ensure quality correction and content. To begin evaluating user-completed documents, the administrator must understand how to calculate a document evaluation score.

1. Calculating a User-Completed Document Evaluation Score

With a basic understanding of the eight types of weighted mistakes, an administrator is able to compile a total number of mistakes per document in order to come up with an objective evaluation score.

a. Evaluate specific portions of the document

Since an administrator does not have the time to evaluate every single page of a user-completed document, there are specific pages that should be assessed in order to generate an evaluation score. These pages include:

- First page of document
- Second page of document
- Estimated middle page of document
- Second to last page of document
- Last page of document

When selecting the middle page of a document for evaluation, a general method can be to divide the total pages of the document by two, then choosing a page closest to the divided amount. For example, a good middle page for a 50-page document would be page 25, while a good page for a 127-page document would be either page 63 or 64.

b. Make a tally of total mistakes counted

Since there are seven different types of evaluated mistakes, it is recommended to use the "TW Correct Doc Score Sheet" document on Google Docs in order to compile each type's score. An effective method for tallying mistake scores would be look over a page for a certain type of mistake, count the total number of those type of mistakes found on the page, then recording that number on the spreadsheet. This way, every page gets evaluated for each type of mistake. After finished compiling the total number of each of the seven mistake types on a page, the administrator can then go to the next designated page for evaluation and continue the tallying process.
(See Figure 14 for an example of what the spreadsheet looks like with tallied mistake scores.)

F	G	H	I	J	K	L	M
Text Not Corrected / Junk Not Removed (3)	Blank/Junk Lines Not Deleted (1)	Missed Lines Not Added (2)	Spelling Normalized (2)	Incorrect Spacing Not Fixed (1)	Didn't convert long-s to 's' (2)	Catch words Not Preserved (1)	Total Score
4	1	0	0	0	2	0	17
3	19	0	0	0	2	0	32
110	5	2	5	0	5	0	359
0	3	0	0	0	0	0	3
0	1	0	0	0	0	0	1
1	7	3	0	0	0	0	16
4	50	4	0	0	0	1	71
0	92	35	1	0	0	1	165
5	39	6	0	0	2	0	70
3	4	3	1	0	2	0	25
1	8	2	1	0	0	0	17
1	6	0	0	0	0	0	9
5	4	7	0	1	0	0	34
6	12	1	0	1	0	0	33
2	5	5	42	5	4	0	118
61	5	0	10	1	3	0	154
24	3	4	0	2	6	0	39
0	67	2	0	0	0	0	71
2	16	1	0	0	4	0	32
4	8	0	0	0	0	0	20
5	10	1	0	0	0	0	27
7	7	0	0	0	6	0	40
2	5	4	0	0	0	0	19
3	5	0	0	0	0	0	14

Figure 14 A section of the "TW Correct Doc Score Sheet

c. Multiply weighted mistake points

Once an administrator has tallied up the mistake score for each type of mistake on every specified page of the document, he or she can now begin to multiply scores by their weighted amounts. These weighted amounts are as follows:

- Text not corrected/Junk not removed (3)
- Blank/Junk lines not deleted (1)
- Missed lines not added (2)
- Spelling normalized (2)
- Incorrect spacing (1)
- Failure to convert long ‘s’ (2)
- Catch words not preserved (1)
- Unauthorized TEI markup used (no penalty)

For example, if a document’s total number of mistakes for not correcting text were 10, this score would be tripled to 30 because it has a weighted amount of 3. The other types of mistake totals will be either left alone (because they are only weighted once) or doubled depending on which type.

d. Add weighted scores together to obtain evaluation score

Once each mistake type has its weighted score, these weighted scores can be added to come up with the final evaluation score. This evaluation score, depending on its number and the total number of pages in a document, can be used to see whether or not a user-corrected TypeWright document is adequately completed.

2. The Eight Types of Weighted Mistakes

These common mistakes make up the scoring system used to evaluate whether or not a user-marked completed document meets the standards of an 18thConnect completed

document. Once a basic understanding of these mistakes is met, the process of calculating an evaluation score becomes a much easier process.

a. Text not corrected/Junk not removed (3 points)

This mistake is the most heavily penalized with a weighted value of three points per error. These mistakes appear as either nonsensical characters or misspellings by the computer not fixed by the TypeWright user in the editing box. Every 'junk' character and word not corrected are three points each.

In this example, there would be two mistakes counted for not correcting 'doating' and 'Fool' and one mistake counted for not removing the apostrophe, giving a total score of nine points counted for this line.

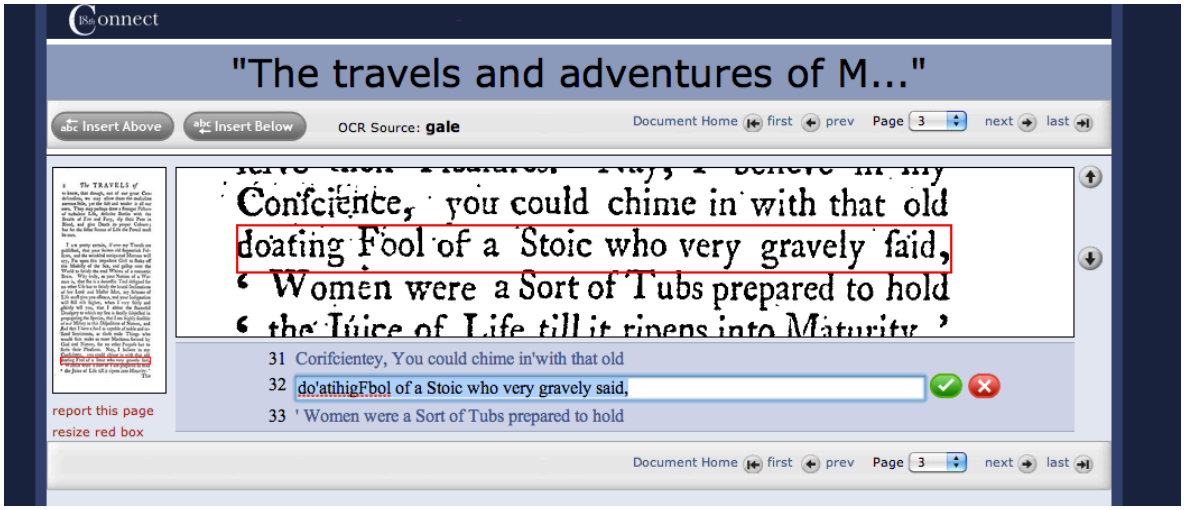


Figure 15: Text not corrected/Junk not removed

b. Blank/Junk lines not removed (1 point)

This is by far the most common mistake of all, in which users do not delete blank lines or lines that the OCR mistakenly reads as text. These mistakes are most common on title pages as well as documents with tables, various margins, and images. Each line not removed is counted as one point.

In this example, both lines 5 and 6 would be counted as lines not removed and each receives one point.

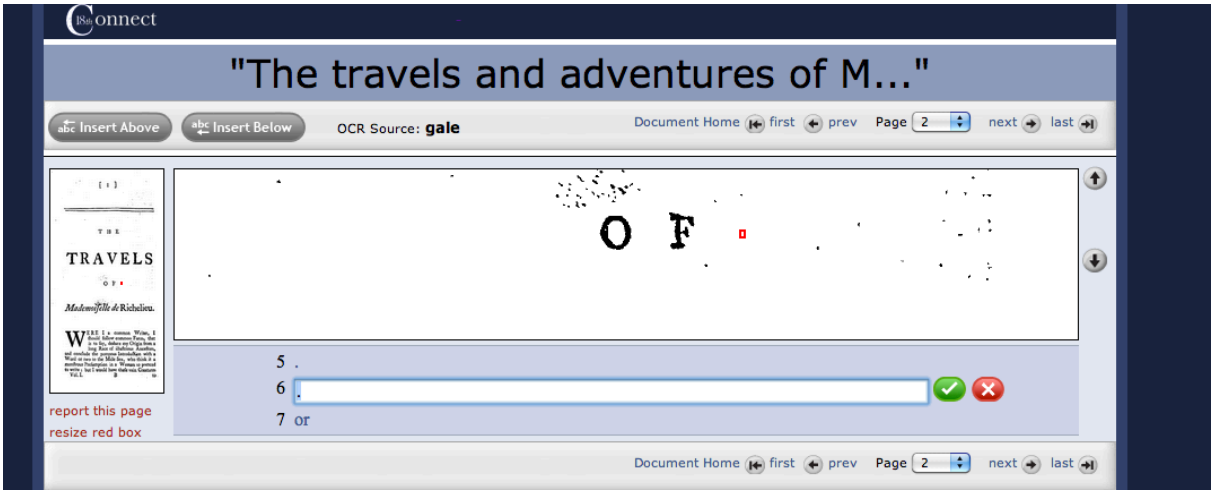


Figure 16 Blank/Junk lines not removed

c. Missing lines not added (2 points)

This mistake is counted when the TypeWright user has not manually added lines of text that the OCR did not catch. These errors are most common with headings, title pages, and page numbers located at the top of the page. Each line not added is counted as two points.

In this example, the TypeWright user has not added a line to include the text found at the top of the page, resulting in the penalty of two points.

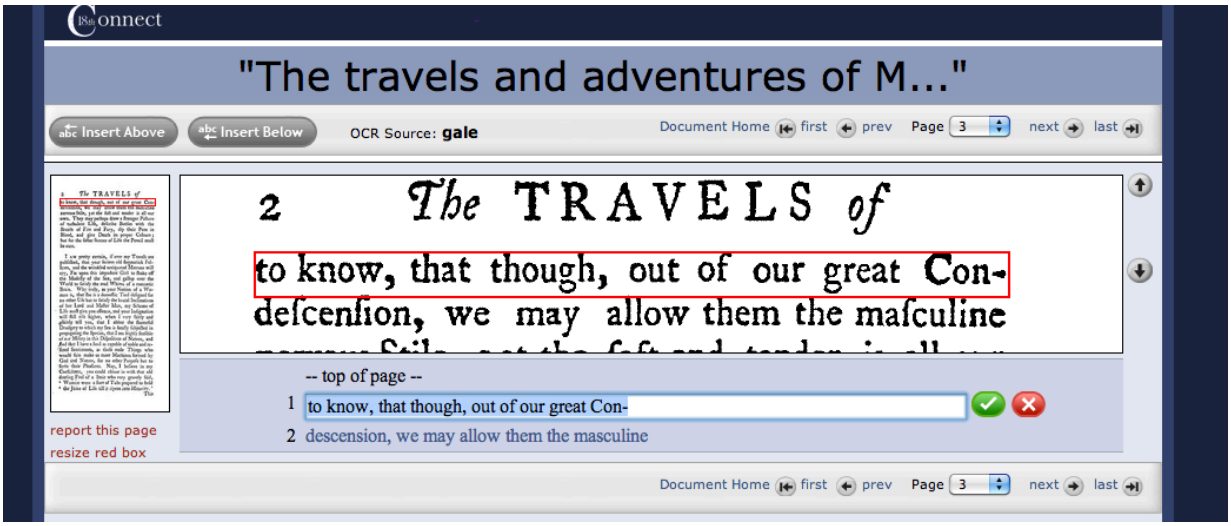


Figure 17 Missing lines not added

d. Spelling normalized (2 points)

This is a common mistake resulting from a TypeWright user 'modernizing' the text from a document. Typically users will correct the older variations of words to fit their present-day English spellings, failing to preserve the original text. This mistake is worth two points for every word normalized.

In this example, the original word ‘suprize’ has been normalized to the present-day spelling ‘surprise’; this mistake results in the penalty of two points.

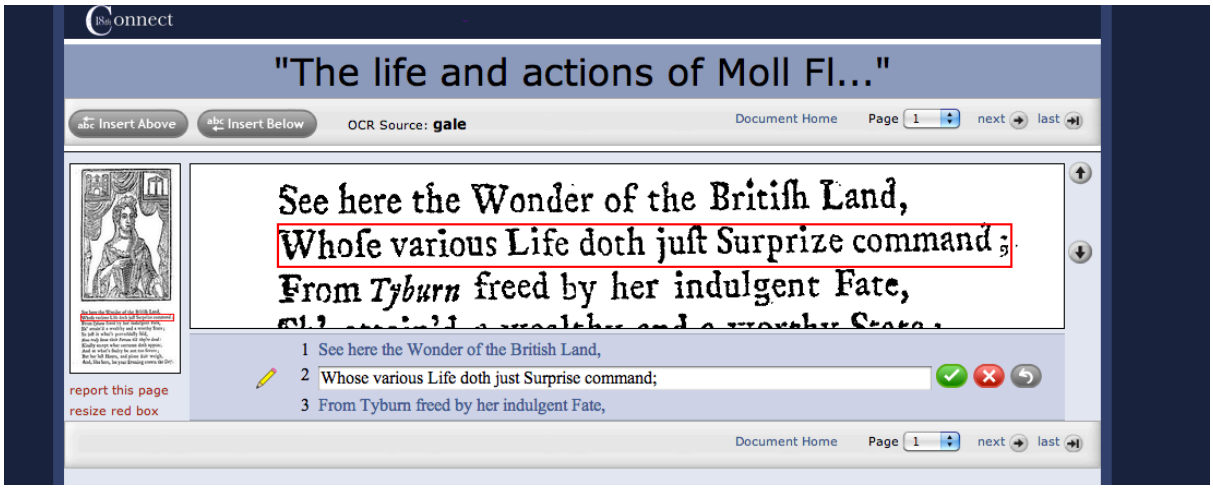


Figure 18 Spelling normalized

e. Incorrect spacing (1 point)

This type of mistake occurs when a TypeWright user does not correct/preserve the spacing to reflect the original document. This mistake is usually failing to separate words or removing excess spacing that the OCR has mistakenly read. Each instance of incorrect spacing has a penalty of one point.

In this example, there is too much space between the words “being” and “debauch’d”. Additionally, there is no space between the words “Master’s” and “Eldest”. These mistakes would result in a penalty of one point each.

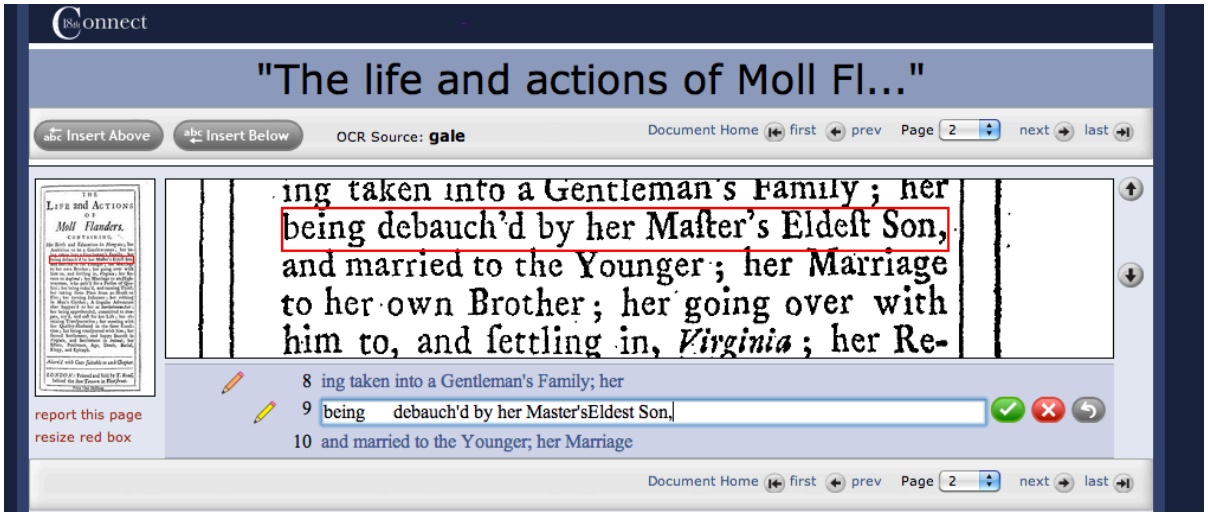


Figure 19 Incorrect spacing

f. Failure to convert long ‘s’ (2 points)

In almost every Typewright document, the long 's' symbol, ſ, is used instead of the present-day letter 's'. Generally, the OCR attempts to correct this symbol with the letter 'f' which should be corrected to the letter 's' by the user. Every instance where the user has not converted the long 's' to the letter 's' results in a penalty of two points. In the case of the TEI <subst> tag sequence used to designate the long 's', the text should reflect the long 's' within the <del> tags, and the modern 's' within the <add> tags.

In this example the OCR has interpreted the word 'settling' as 'fettling'; the user has failed to correct this mistake, resulting in a penalty of two points.

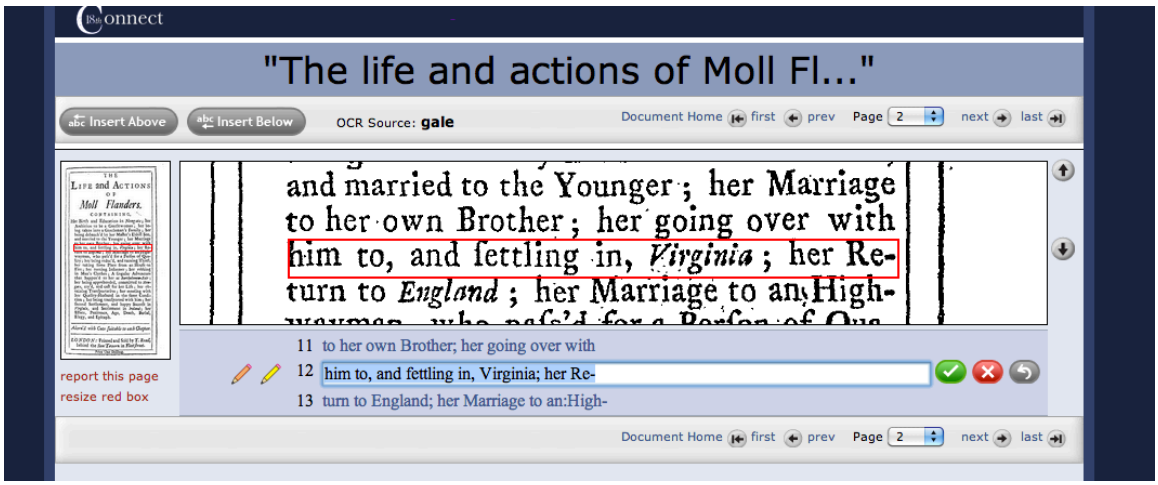


Figure 20 Failure to convert long "s"

g. Catch words not preserved (1 point)

In most of the documents found on 18thConnect, there is a 'catch word' at the bottom of every page, originally designated for bookbinders to correctly order pages for binding. Oftentimes users will delete these catch words thinking they are extraneous, when in fact they need to be preserved because they are part of the original document. Every catch word not preserved is penalized with one point each.

In this example, the user has mistakenly deleted the line with the catch phrase, failing to preserve the original text, resulting in a penalty of two points.

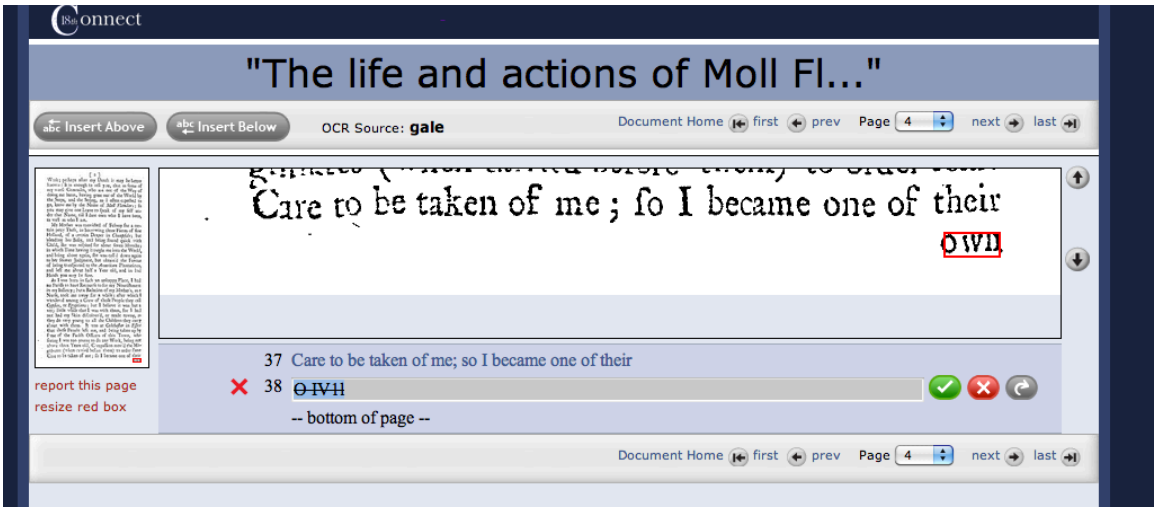


Figure 21 Catch words not preserved

h. Incorrect Use of TEI tag(s) (0 points)

Users are given the option to insert any of a list of accepted TEI markup tags in order to designate unusual formatting or character glyphs. There are two errors that merit comment. Such errors are NOT weighted, only tallied and commented on in the correspondence with the user.

1. Tags that extend over several lines of text: Tags must be opened and closed on the same line.
2. Use of unauthorized tag-attribute-value combinations that do not appear on the following list of TypeWright Accepted Tags, Attributes, and Attribute Values.

MARKUP	TAG/ATTRIBUTE
Italics	<hi rend="italic">…</hi>
Boldface	<hi rend="bold>…</hi>
Superscript	<hi rend="sup">…</hi>
Subscript	<hi rend="sub">…</hi>
Smallcaps	<hi rend="smallcaps>…</hi>
Underlining	<hi rend="underline>…</hi>
Centered Line	<hi rend="center">…</hi>
Drop Cap	<hi rend="dropcap">…</hi>
Page Header	<fw type="header">…</fw>

MARKUP	TAG/ATTRIBUTE
Page Number	<fw type="pageNum">…</fw>
Signature Mark	<fw type="sig">…</fw>
Catchword	<fw type="catch">…</fw>
Strikethrough	<del rend="overstrike">…</del>
Inverted character	<c rend="inverted">…</c>
Substitution of glyph	<subst> ~~ſ~~ <add>s</add> </subst>

C. User Correspondence Form Letters

(available from Basecamp: 18thConnect: Text Documents: TypeWright Form Letters)

1, USER COMPLETES DOCUMENT:

Dear TypeWright User,

Thank you for using the TypeWright interface in 18thConnect to edit **[insert full title here]**. Your work will aid in the effort to correct the "dirty OCR" that runs behind this document and improve the searchability of the 18th-Century digital archive.

We'd also like to thank you for taking the time to mark the text complete, and this document is now under review by the 18thConnect team. Please give us a few days to review the document.

As you may know, our contracts with Gale-Cengage Learning allow us to release corrected text and XML to the scholar-editors who strive to correct the text of our TypeWright documents. These released texts (liberated texts!) can then be used by scholars for digital editions, text-mining, and more. After the review period, we will contact you with the results of your TypeWright editing experience. If the text has been corrected above a certain percentage, and the guidelines stated on the editing interface page have been followed, we will be contacting you with further information on receiving the corrected text.

Please let us know, at your earliest convenience, if you will be requesting plain text or XML from 18thConnect for this text. If you do not respond, we will not send the corrected plain text or XML to this email address.

In the meantime, please consider taking the TypeWright User Survey (http://bit.ly/1awiUjR). As a TypeWright SuperUser, your feedback is very valuable!

Thank you, again, for your interest in 18thConnect, and we look forward to speaking with you again soon.

2. USER DOES NOT RESPOND WITH TXT OR XML REQUEST:

Dear TypeWright User,

Please notify us if you will be requesting plain text or XML from 18thConnect for this text. If you do not respond, we will not send the corrected plain text or XML to this email address.

The 18thConnect team hopes to hear from you, and thank you for using TypeWright to improve the 18th-Century digital archive!

3. USER REQUESTS XML OR TXT:

Dear TypeWright User,

Thank you for contacting us to confirm that you would like to receive [**plain text** or **XML**] for the TypeWright corrected text "[**insert title here**]."

Our team is currently reviewing the corrections made to this text. After the review period, we will contact you with the results of your TypeWright editing experience. If the text has been corrected above a certain percentage, and the guidelines stated on the editing interface page have been followed, we will be contacting you to release the [plain text or XML] for your text! Please give us up to one week from the date you marked the document complete for the evaluation period.

In the meantime, please consider taking the TypeWright User Survey (http://bit.ly/1awiUjR). As a TypeWright SuperUser, your feedback is very valuable!

4. USER CORRECTIONS ARE APPROVED:

Dear TypeWright User,

Thank you for using TypeWright to edit "**[INSERT TITLE]**." Your work has been reviewed by the 18thConnect team, and I'm very pleased to announce that it has passed our review process.

Though this document passed our review process, the 18thConnect team wants you to be aware of the following errors found in the edited document. We like to communicate this information to our users, in case they intend to use the released text for further digital work.

[examples below; adjust for each user]

- Document was not 100% completed.
- Final ~10 pages of the document were not corrected/edited.
- Some "garbage" characters (or non-alphanumeric characters) were remaining in the document in suspicious patterns. This is an indicator that some OCR'd lines were not corrected by hand.
- The following TEI markup tag(s) used in the document did not come from our list of accepted tags, and thus will not be processed by our conversion protocols: [**INSERT TAG**]. If this tag will be important to your future projects, please contact the project manager by emailing liz@18thconnect with a request that we add this tag to our list.

Attached to this email is a corrected text version of "**[INSERT TITLE]**." Please let us know, by responding to this email, if you are interested in receiving an XML version of this corrected text, in TEI-A format (http://www.tei-c.org/index.xml).

We look forward to seeing how this released text (liberated text!) forwards your research, digital edition, or project. Please stay in contact with the 18thConnect team via this email, technologies@18thconnect.org, or @18thConnect to update us on your progress. We're always happy to hear and publicize the accomplishments of our 18thConnect users!

If you have not done so, please consider taking the TypeWright User Survey (http://bit.ly/1awiUjR). As a TypeWright SuperUser, your feedback is very valuable.

5. USER DOES NOT MEET MINIMUM EDIT REQUIREMENTS

Dear TypeWright User,

Thank you for using TypeWright to edit "**[INSERT TITLE]**." Your work has been reviewed by the 18thConnect team, and we have determined that the corrections do not meet our standards for a "corrected document."

Unfortunately, this means that we are not able to release the plain text or XML of this document, at this time. We have reset the TypeWright text, so that you (and/or other users) can make further corrections and mark the document complete.

If you choose to correct the document further (and we hope that you do!), we have a few recommendations below:
[examples below; adjust for each user]

If there's a reason for this, or you think there's some kind of error, please let us know so we can take that into consideration. Otherwise please address the issue(s) listed above so that we can release the document's text or XML to you.

We look forward to hearing from you in the future, and please let the team know if you have any further questions about correcting documents, receiving (liberating!)

TypeWright texts, or 18thConnect issues in general.

If you have not done so, please consider taking the TypeWright User Survey (http://bit.ly/1awiUjR). As a TypeWright SuperUser, your feedback is very valuable.

D. Gale correspondence forms?

We have no cover letter at this time, but theoretically one will be needed.

V. Reporting Bugs

A. Please report any bugs that arise to the Project Manager for 18thConnect by sending an email to either technologies@18thconnect.org or liz@18thconnect.org.

B. Helpful information includes:

User's name
The date and time
Which page of the site is involved?
What is wrong?

- ____ is not where it should be.
- ____ is where it should not be.
- To be especially helpful, include one or several screen shots!!

What were you doing?

- Opening a TypeWright enabled document text to correct (Start to edit).
- Correcting document text in TypeWright.
- Resizing a "red box."
- Completing (100%) text correction for a document.

What is the extent of the problem?

- Does the problem affect all of your recent activities on this page, or only some?
- Does the problem affect your activity on other related pages?
- Does the problem affect any of your previous work on the site?
- List the URI(s) for the affected item(s)?
- Does switching to another browser (Mozilla, Chrome, Safari, etc.) solve the problem?

C. EXAMPLE (from pivotal tracker):

TITLE
If your Collex session has expired and you click the "collect" button on an object, it will appear that you have successfully collected it when in fact you haven't.
DESCRIPTION
I discovered this bug while testing an update to the session cleaning daemon. I logged in to edge.18th and did a search. Then I waited an amount of time that I knew would cause my session to be terminated. Then I clicked the "collect" button on an object. The usual spinner appeared and the object appeared to have been collected. When I clicked on

"My18th," I was logged out, and after logging in again I checked "My18th" and found that the object had not been collected. Ideally, if the user has been logged out and tries to collect an object, they should get some feedback indicating that they're logged out and the object has not been collected.

VI. Helpful Links

A. TypeWright

- *Information*
  http://www.18thconnect.org/about/typewright/
- *"TypeWright Users" group in 18thConnect*
  http://www.18thconnect.org/groups/20
- *TypeWright in the Blogosphere*
  http://digitalhumanistbeginner.wordpress.com/category/typewright/
  http://www.performantsoftware.com/portfolio/typewright/
  http://ltroost.wordpress.com/2014/01/21/what-lurks-beneath-the-ecco-page/

B. TypeWright Help Resources

- *"TypeWright" in "What Is 18thConnect?"*
  http://www.18thconnect.org/about/typewright/
- *"Exhibits" tab "Show:" dropdown options*
  http://www.18thconnect.org/communities
  - "Exhibits": "Working with TypeWright" by R. Wilson
    http://www.18thconnect.org/exhibits/Working_with_TypeWright
  - "Groups": "TypeWright Users"
    http://www.18thconnect.org/groups/20
  - "Discussions"
    http://www.18thconnect.org/communities (look for related discussions)
- *18thConnect "News" tab: any posts tagged "typewright"*
  http://www.18thconnect.org/news/

C. IDHMC and Projects

- **IDHMC**
  http://idhmc.tamu.edu/
- *Advanced Research Consortium (ARC)*
  http://idhmc.tamu.edu/arcgrant/
- *18thConnect*
  http://www.18thconnect.org/