EMOP

Early Modern OCR Project

eMOP guide to OCR

Flowcharting a Course Through Open-Source Waters

eMOP Info

eMOP Website

- emop.tamu.edu/

- TCDL 24x7 Presentation
  - emop.tamu.edu/news
- More information
  - emop.tamu.edu/TxDHC-flowcharts
- eMOP Workflows
  - emop.tamu.edu/workflows
- Mellon Grant Proposal
  - idhmc.tamu.edu/projects/Mellon/eMOPPublic.pdf

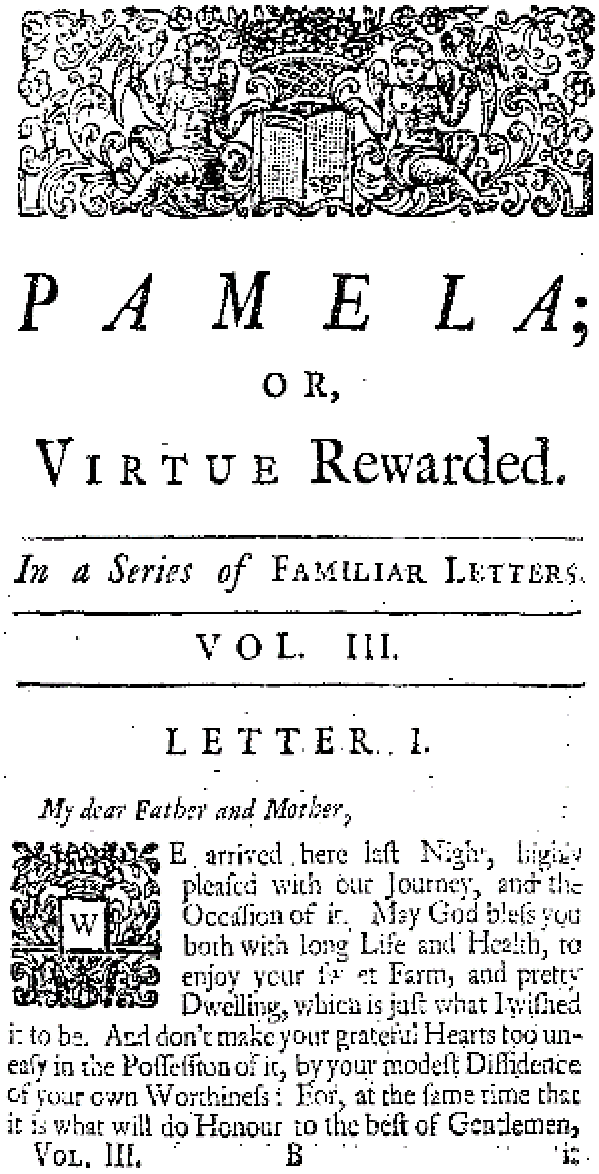
More eMOP

- **Facebook**
  - Early Modern OCR Project
- **Twitter**
  - #emop
  - @IDHMC_Nexus
  - @matt_christy

The Problems

Early Modern Printing

- Individual, hand-made typefaces
- Worn and broken type
- Poor quality equipment/paper
- Inconsistent line bases
- Unusual page layouts, decorative page elements,
- Special characters & ligatures
- Spelling variations
- Mixed typefaces and languages



PAMELA;

OR,

VIRTUE Rewarded.

In a Series of FAMILIAR LETTERS.

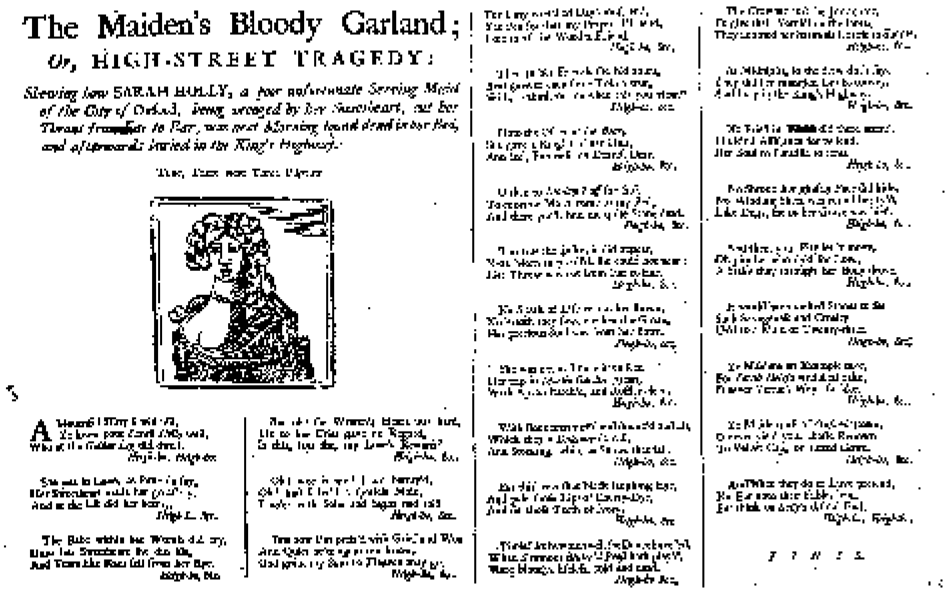
VOL. III.

LETTER I.

My dear Father and Mother,

WE arrived here laſt Night, highly pleaſed with our Journey, and the Occaſion of it. May God bleſs you both with long Life and Health, to enjoy your ſweet Farm, and pretty Dwelling, which is juſt what I wiſhed it to be. And don't make your grateful Hearts too uneaſy in the Poſſeſſion of it, by your modeſt Diffidence of your own Worthineſs: For, at the ſame time that it is what will do Honour to the beſt of Gentlemen,

VOL. III. B it



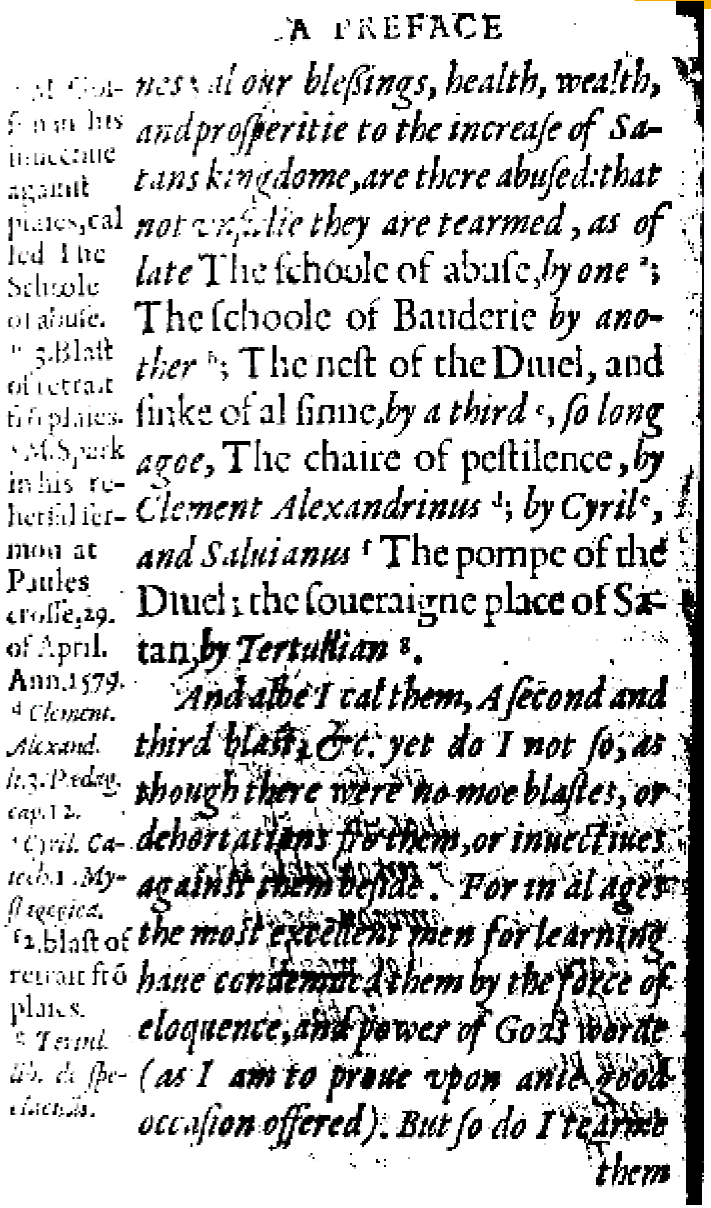
The Maiden's Bloody Garland;
Or, HIGH-STREET TRAGEDY:

Shewing how SARAH HOLLY, *a poor unfortunate Serving Maid of the City of Oxford, being wronged by her Sweetheart, cut her Throat from Ear to Ear, was next Morning found dead in her Bed, and afterwards buried in the King's Highway.*

Tune, There was Three [illegible]

[illegible]

FINIS.



A PREFACE

nes al our bleſsings, health, wealth, and proſperitie to the increaſe of Satans kingdome, are there abuſed: that not vniuſtlie 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 frō 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] St. Goſſon in his inuectiue againſt plaies, called The Schoole of abuſe.

[b] 3.Blaſt of retrait frō plaies.

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

[d] Clement. Alexand. lib.3.Pedag. cap.12.

[e] Cyril. Catech.1. Myſtagogica.

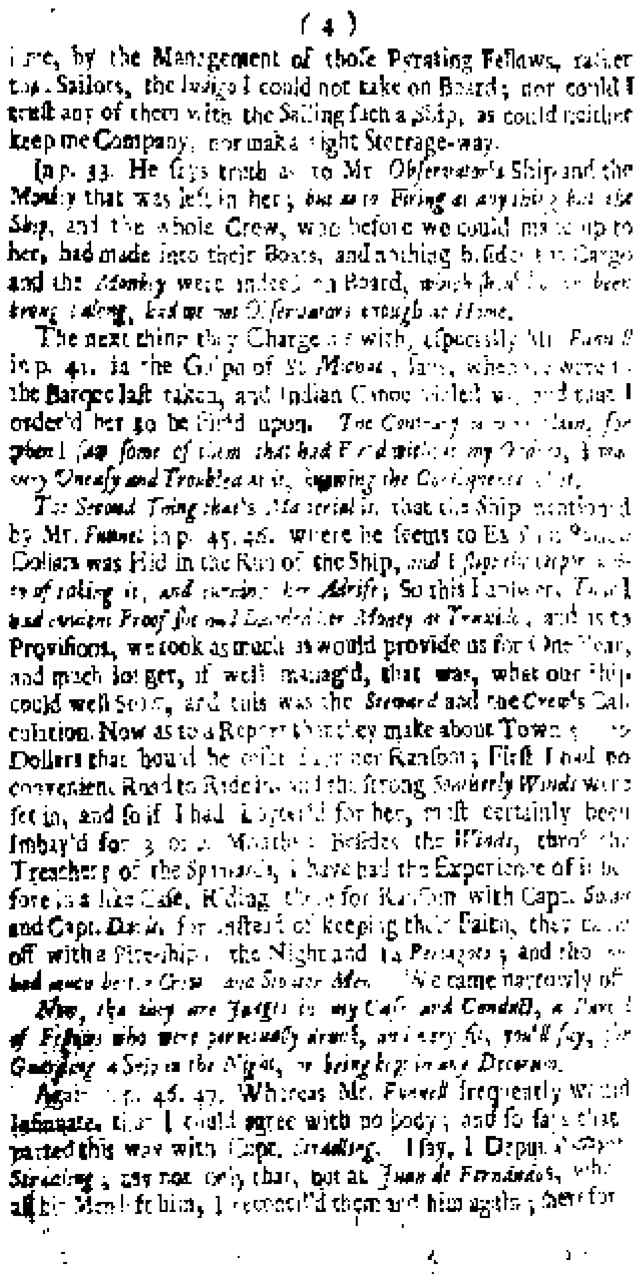
[f] 2.blaſt of retrait frō plaies.

[g] Tertul. lib. de ſpectaculis.

The Problems

Document/Image Quality

- Torn and damaged pages
- Noise introduced to images of pages
- Skewed pages
- Warped pages
- Missing pages
- Inverted pages
- Incorrect metadata
- Extremely low quality TIFFs (~50K)



(4)

[illegible] me, 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 right Steerage-way.

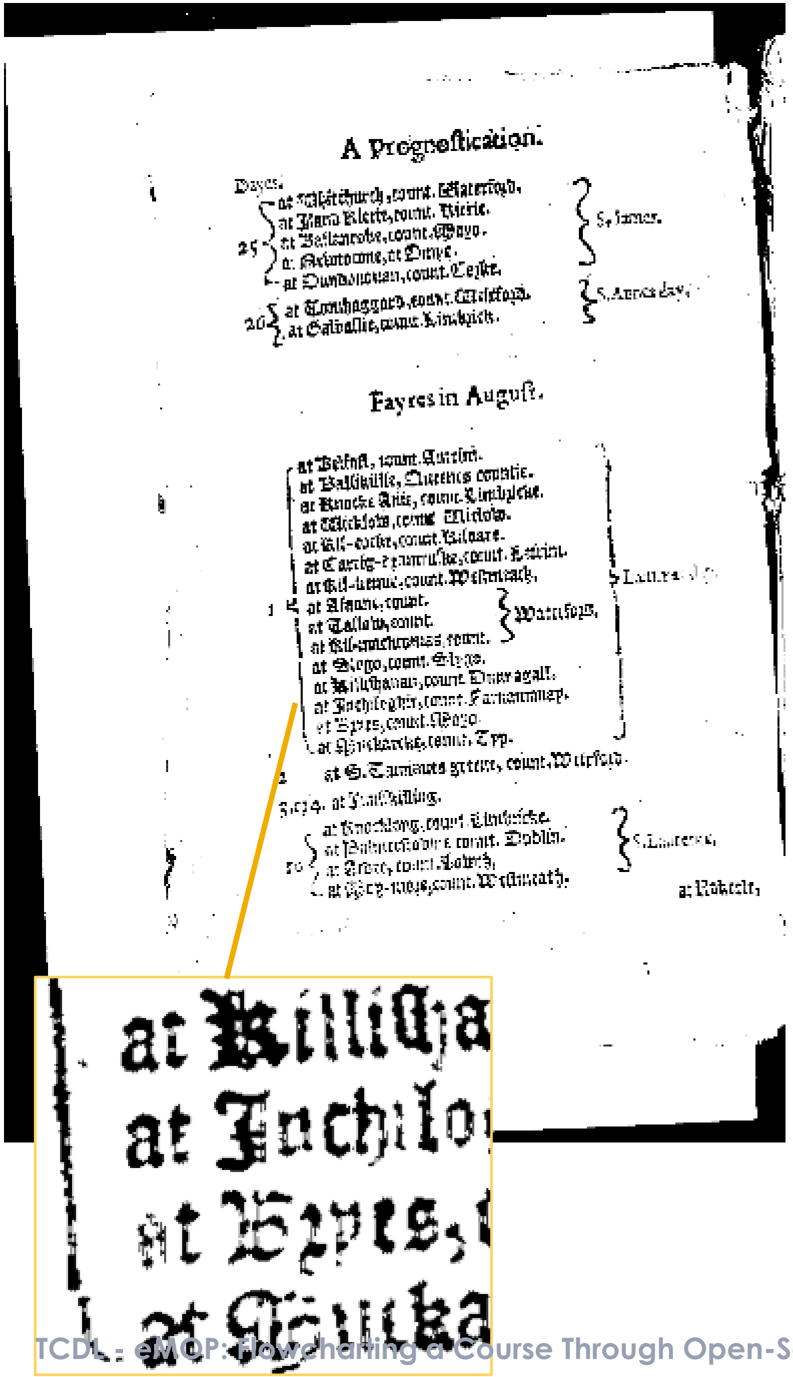
[In p. 33. He ſays truth as to Mr. Obſervator's Ship and the Money that was left in her; but as to Firing at any thing but the Ship, and the whole Crew, who before we could make up to her, had made into their Boats, and nothing beſides the Cargo and the Money were indeed on Board, which [illegible] been 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 [illegible] the Bargue laſt taken, and Indian Canoe [illegible] us, and that I order'd her to be Fir'd upon. The Contrary [illegible], 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 [illegible].

The Second Thing that's Material is, that the Ship mentioned by Mr. Funnel in p. 45. 46. where he ſeems to [illegible] Dollars was Hid in the Run of the Ship, and I [illegible] of taking it, and [illegible] her Adrift; So this I anſwer, [illegible] had [illegible] Proof [illegible] 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 [illegible] they make about Town [illegible] Dollars that should be [illegible] her Ranſom; Firſt I had no convenient Road to Ride in, and the ſtrong Southerly Winds were ſet in, and ſo if I had [illegible] 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 [illegible] for Ranſom with Capt. Swan and Capt. Davis, for inſtead of keeping their Faith, they came off with a Fire-ſhip [illegible] the Night and 14 Periagoes; and alſo [illegible] had more [illegible] Crew and Stouter Men. We came narrowly off.

Now, [illegible] are Judges in my Caſe and Conduct, a [illegible] 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 [illegible] p. 45. 47. Whereas Mr. Funnell frequently would Inſinuate, that I could agree with no body; and ſo ſays that parted this way with Capt. Stradling. I ſay, I [illegible] Stradling, and not only that, but at Juan de Fernandos, when all his Men left him, I reconcil'd them and him again; ſo for



A Prognoſtication.

Dayes.

25 at [illegible]church, count. Waterford. / at [illegible], count. Kerrie. / at Ballenrobe, count. Mayo. / at [illegible]towne, at Drinke. / at Dundonnan, count. Corke. — S. James.

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

Fayres in Auguſt.

1 at Belfaſt, count. Antrim. / at Ballikille, Queenes countie. / at Knocke Anie, count. Limbricke. / at Wicklow, count. Wicklow. / at Kil-cocke, count. Kildare. / at Carrig-[illegible], count. Leitrim. / at Kil-[illegible], count. Weſtmeath. / at Aſhane, count. / at Tallow, count. / at Kil-macthomas, count. — Waterford. / at Sligo, count. Sligo. / at Killibanan, count. Donnegall. / at Inchilogher, count. Farmanagh. / at Eyres, count. Mayo. / at Gowrane, count. Typ. — Lammas day.

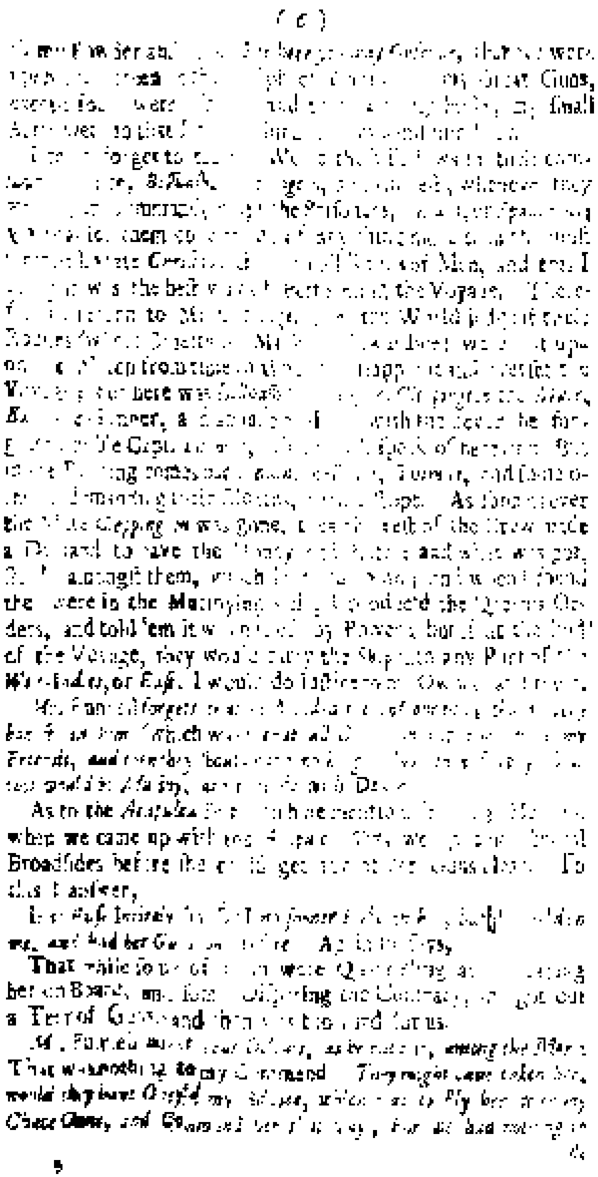
2 at S. [illegible] greene, count. Wexford.

3. 4. at [illegible]killing.

10 at Knocklong, count. Limbricke. / at [illegible], count. Dublin. / at Arde, count. Louth. / at [illegible], count. Weſtmeath. — S. Laurence.

at Rakeele,

at Killiha
at Inchilo
at Eyres,
at Gowka



(6)

[illegible] were [illegible] that we were [illegible] Great Guns, [illegible] were [illegible] small Arms [illegible] to that [illegible]

[illegible] forget to [illegible] We [illegible] was in [illegible] they [illegible] the Priſoners, [illegible] them [illegible] Conſideration [illegible] of Men, and that I [illegible] was the beſt [illegible] the Voyage. [illegible] return to [illegible] World [illegible] [illegible] from thence [illegible] happ[illegible] Voyage [illegible] here was [illegible] the Stars, [illegible], a [illegible] with the [illegible] he [illegible] the Capt. [illegible] of [illegible] [illegible] ing [illegible] Turner, and [illegible] [illegible] their [illegible] Capt. As ſoon as ever the [illegible] Clipping [illegible] was gone, [illegible] of the Crew made a [illegible] to have the [illegible] and what was got, [illegible] amongſt them, [illegible] the [illegible] in the Mutinying [illegible] the Queens Orders, and told 'em it w[illegible] by Powers, but [illegible] of the Voyage, they would carry the Ship [illegible] any Port of the Weſt-Indies, or Eaſt. I would do [illegible]

Mr. Funnell forgets [illegible]

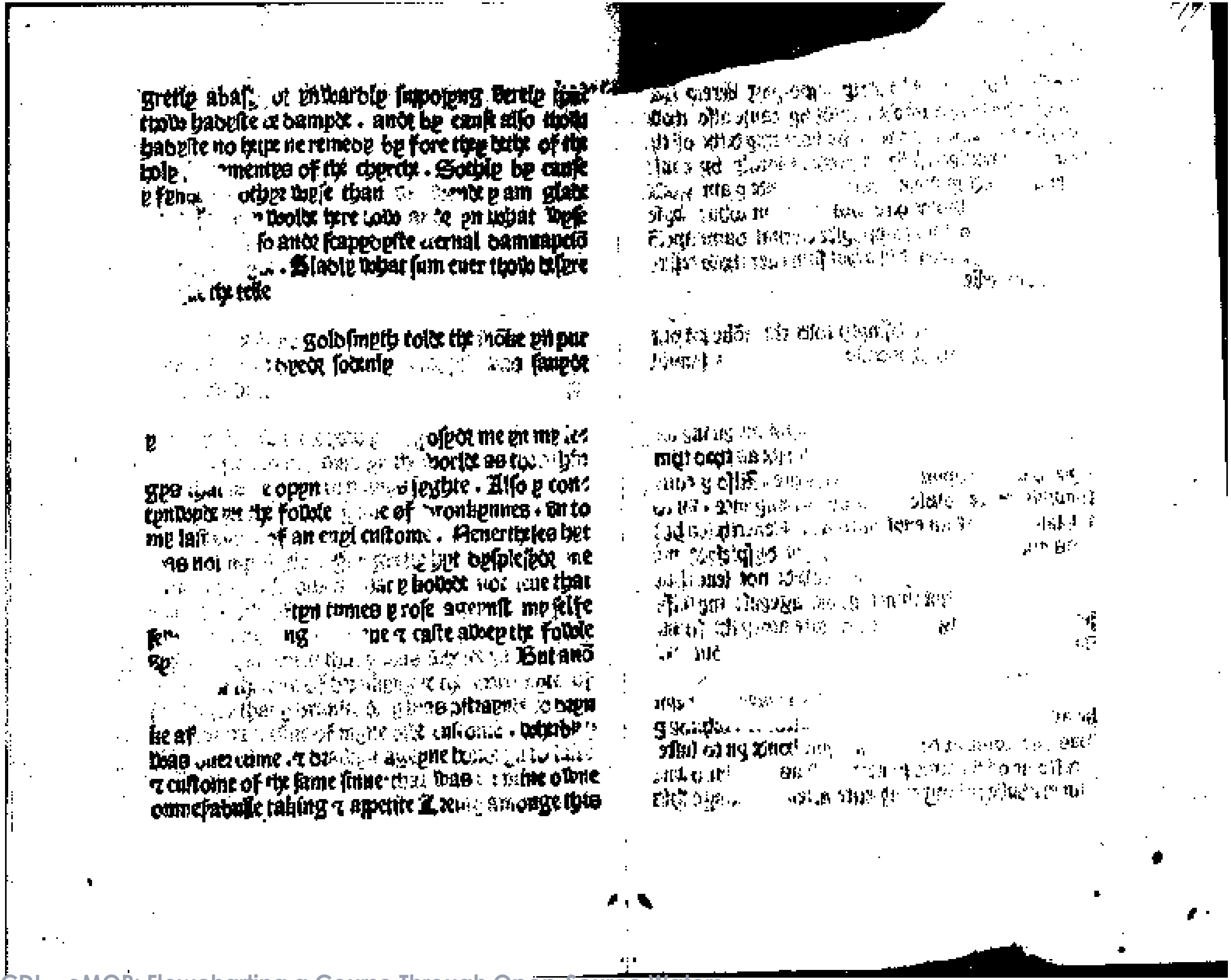
[illegible] Friends, and [illegible] Deck.

As to the *Arabella* [illegible] when we came up with [illegible] Broadſides before [illegible] get [illegible] To this I anſwer,

[illegible]

That while ſo [illegible] were Quarrelling [illegible] her on Board, [illegible] the Contrary, [illegible] got out a Tier of Guns and [illegible] us.

Mr. Funnell [illegible] among the [illegible] That was nothing to my Command. *They might [illegible] taken her, would they have Obeyed* my [illegible] *Fly her* [illegible] *Chace Guns, and* [illegible] *had* [illegible]



gretly abaſſ[illegible] vnſuarble ſuppoſyng verily that thow haddeſt be dampned . and be cauſe alſo thow haddeſt no hope ne remedye by fore thy dethe of the holy [illegible]mentes of the chirche . Wherfore be cauſe y ſeno[illegible] other wyſe than [illegible] y am glade [illegible] wolde here [illegible] what thow ſo and ſuppoſyſt eternal dampnacion [illegible] . Sayde what ſum euer thow before [illegible] the teſte

[illegible] goldſmyth tolde the [illegible] yet pur [illegible] foundg [illegible] ſayde

y [illegible] oſſed me yn my [illegible] the worlde as [illegible] gre[illegible] open [illegible] leyſure . Alſo y contynewyd [illegible] the foule [illegible] of [illegible]wordynes . vn to my laſt [illegible] of an euyl cuſtome . Neuertheles yet as now [illegible] gretly hit dyſpleſed me [illegible] y wolde [illegible] that [illegible] tymes y roſe [illegible] my ſelfe [illegible] ng [illegible] caſte awey the foule [illegible] But and [illegible] ſtraunge [illegible] lie af[illegible] of [illegible] cuſtome . wherfore was [illegible] awey [illegible] cuſtome of the ſame ſynne that was [illegible] owne comfortable takyng [illegible] appetite [illegible] amonge this

[illegible]

Wrangling Data

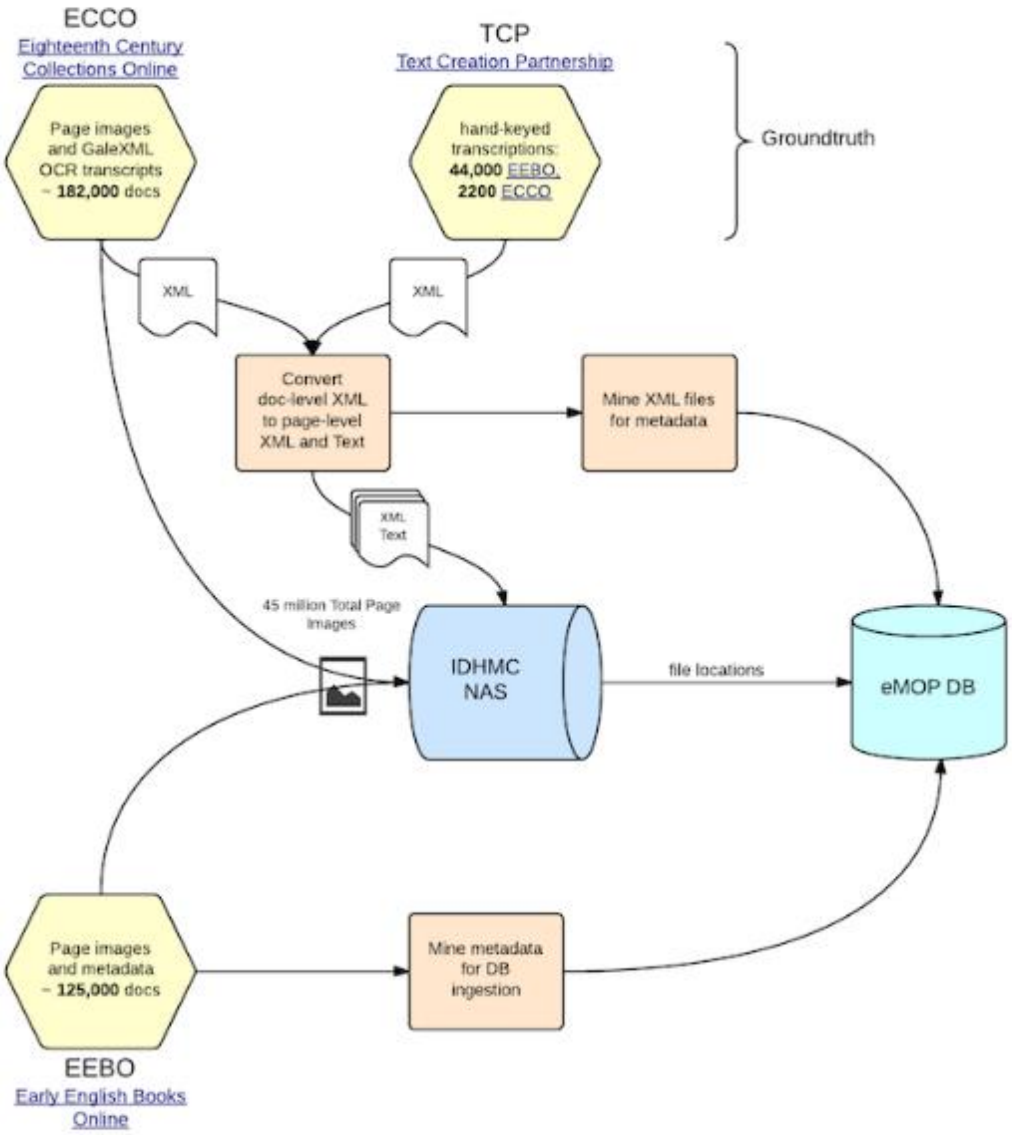
The Numbers

- **EEBO**: ~125,000 documents, ~13 million pages images (1475-1700)
- **ECCO**: ~182,000 documents, ~32 million page images (1700-1800)
- **TCP**: ~46,000 double-keyed hand transcriptions (44,000 EEBO, 2,200 ECCO) - Groundtruth
- **Total**: >300,000 documents & 45 million page images.

The Data

- **ECCO page images** (1 pg/ image)
- **ECCO original OCR results** (doc-level XML files)
- **ECCO TCP transcriptions** (doc-level XML and text files)
- **EEBO page images** (2 pgs/ image)
- **EEBO TCP transcriptions** (doc-level XML and text files)

Wrangling Data



ECCO

Eighteenth Century Collections Online

Page images and GaleXML OCR transcripts ~ **182,000** docs

TCP

Text Creation Partnership

hand-keyed transcriptions: **44,000** EEBO, **2200** ECCO

Groundtruth

XML

XML

Convert doc-level XML to page-level XML and Text

Mine XML files for metadata

XML Text

45 million Total Page Images

IDHMC NAS

file locations

eMOP DB

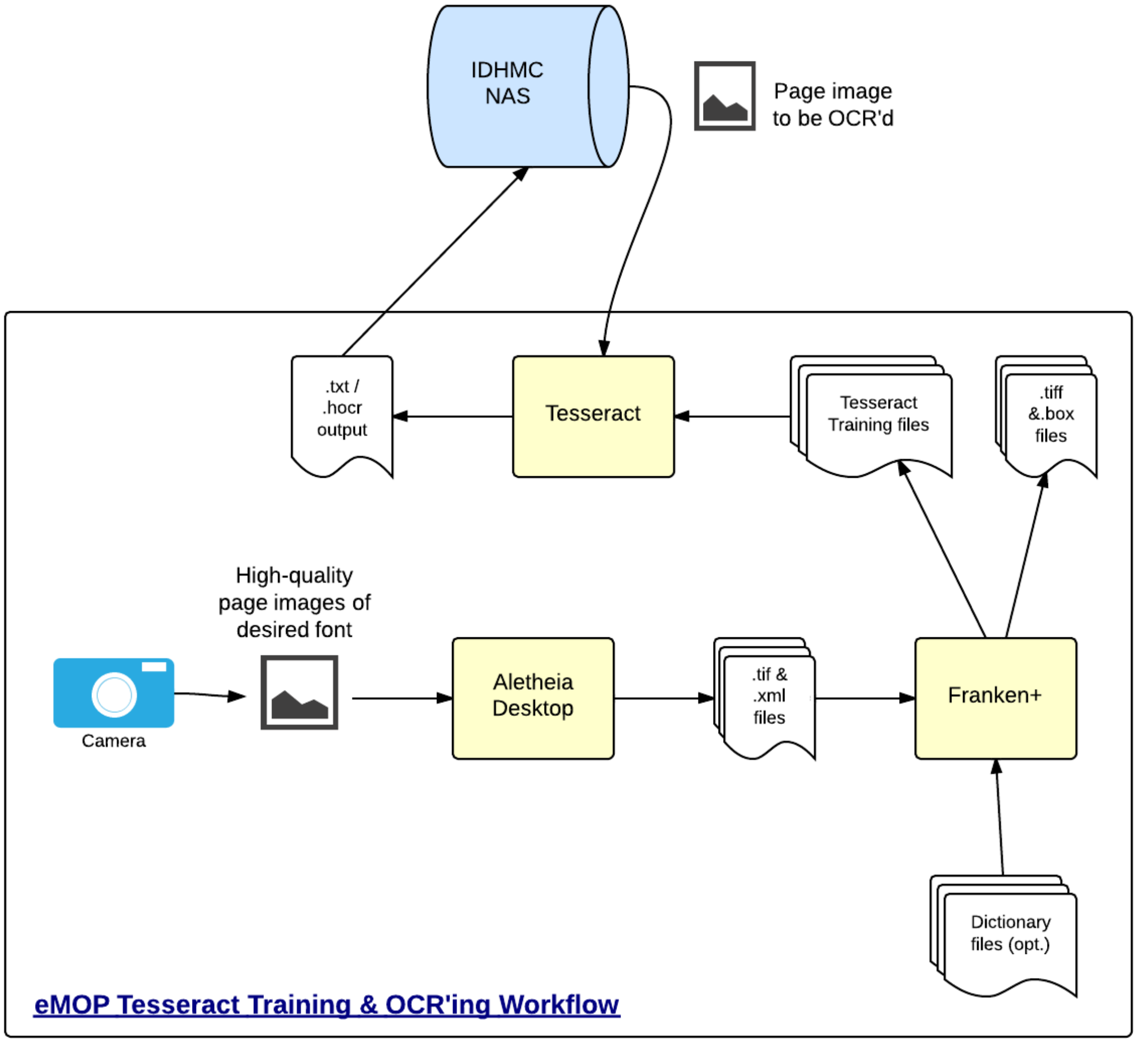
Page images and metadata ~ **125,000** docs

Mine metadata for DB ingestion

EEBO

Early English Books Online

Training Tesseract

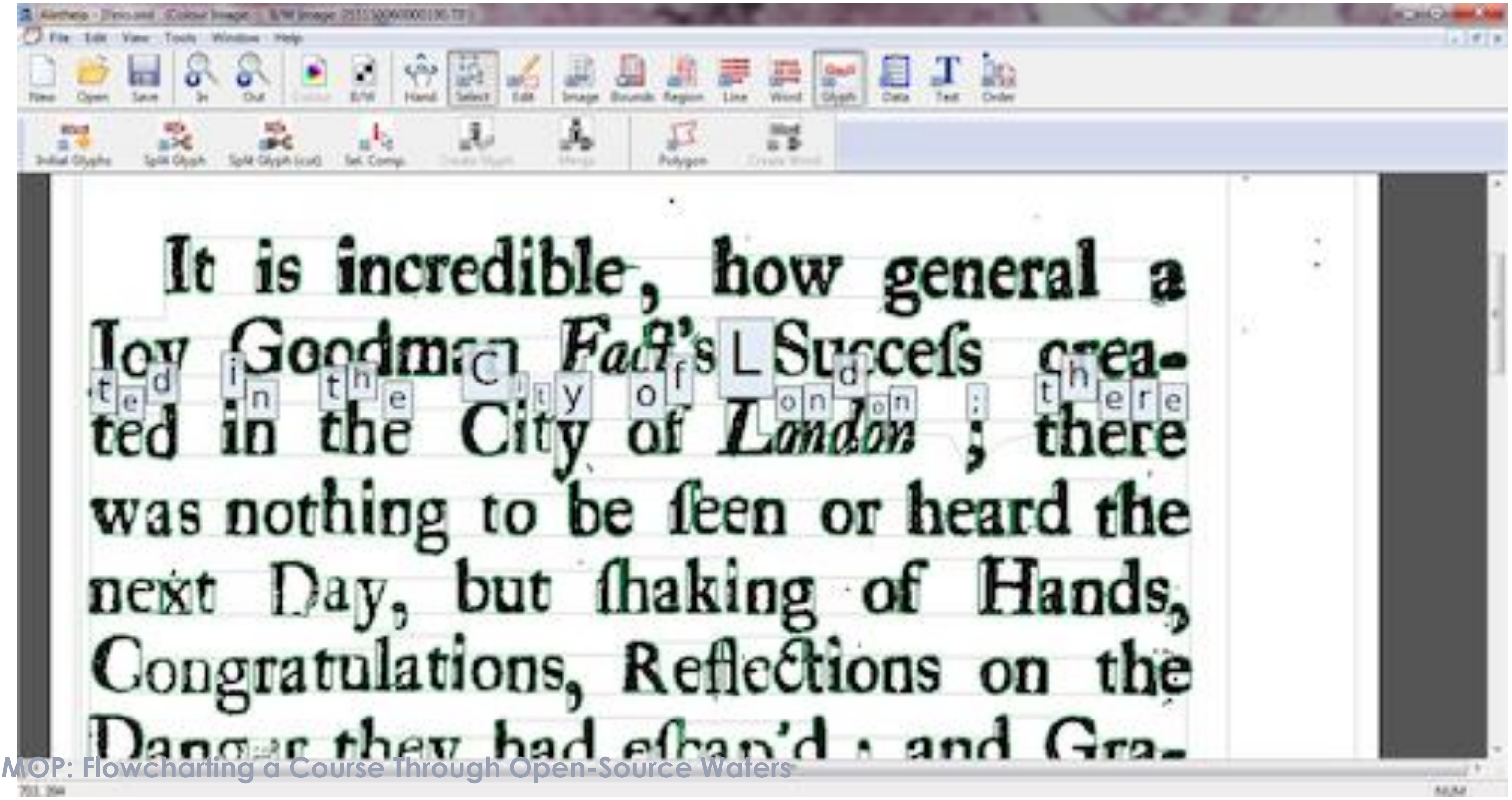


eMOP Tesseract Training & OCR'ing Workflow

Training Tesseract

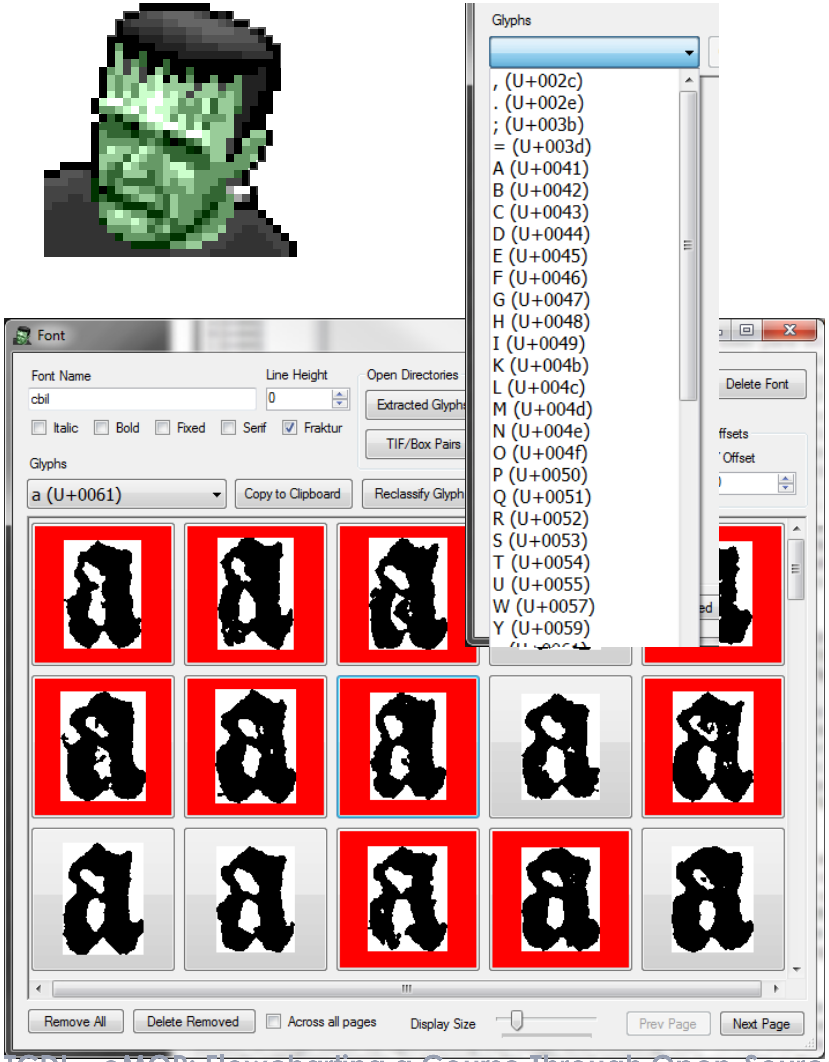
Aletheia

Created by PRImA Research Labs. A team of undergraduates uses Aletheia to identify each glyph on the page images, and ensure that the correct Unicode value is assigned to each. Aletheia outputs an XML file containing all identified glyphs on a page with their corresponding coordinates and Unicode values.



It is incredible, how general a
Joy Goodman Fact's Succeſs crea-
ted in the City of London ; there
was nothing to be ſeen or heard the
next Day, but ſhaking of Hands,
Congratulations, Reflections on the
Danger they had eſcap'd ; and Gra-

Training Tesseract



Font Name: cbil

Line Height: 0

☐ Italic ☐ Bold ☐ Fixed ☐ Serif ☑ Fraktur

Glyphs: a (U+0061)

Glyphs

, (U+002c)
. (U+002e)
; (U+003b)
= (U+003d)
A (U+0041)
B (U+0042)
C (U+0043)
D (U+0044)
E (U+0045)
F (U+0046)
G (U+0047)
H (U+0048)
I (U+0049)
K (U+004b)
L (U+004c)
M (U+004d)
N (U+004e)
O (U+004f)
P (U+0050)
Q (U+0051)
R (U+0052)
S (U+0053)
T (U+0054)
U (U+0055)
W (U+0057)
Y (U+0059)

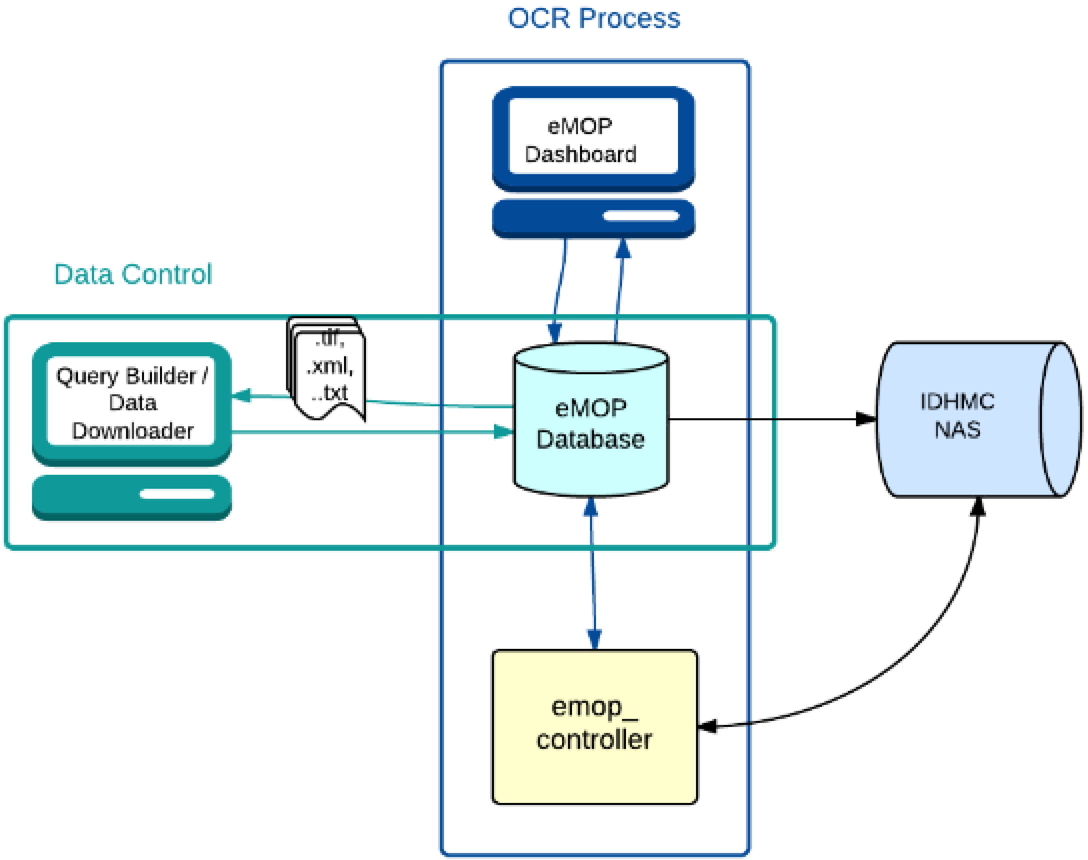
☐ Across all pages

Franken+

1. Takes Aletheia's output files as input.
2. Groups all glyphs with the same Unicode values into one window for comparison.
3. Mistakenly coded glyphs are easily identified and re-coded.
4. A user can quickly compare all exemplars of a glyph and choose just the best subset, if desired.
5. Uses all selected glyphs to create a Franken-page image (TIFF) using a selected text as a base.
6. Outputs the same box files and TIFF images that Tesseract's first stage of native training.
7. Also allows users to complete Tesseract training using newly created box/TIFF file pairs, and add optional dictionary and other files.
8. Outputs a .traineddata file used by Tesseract when OCRing page images.

Control Tools

- **Query Builder**: Users can build queries to find specific documents, or sets of documents, in the eMOP DB. Document sets can be labeled or grouped with an identifier.
- **Data Downloader**: Available via the QB. Allows identified items to be downloaded: page images, OCR results, associated groundtruth.
- **eMOP Dashboard**: Users select documents from the database, individually or with the identifier created in the QB, for OCRing with specific typeface training. The Dashboard displays results with Juxta/RETAS scores when groundtruth is available.

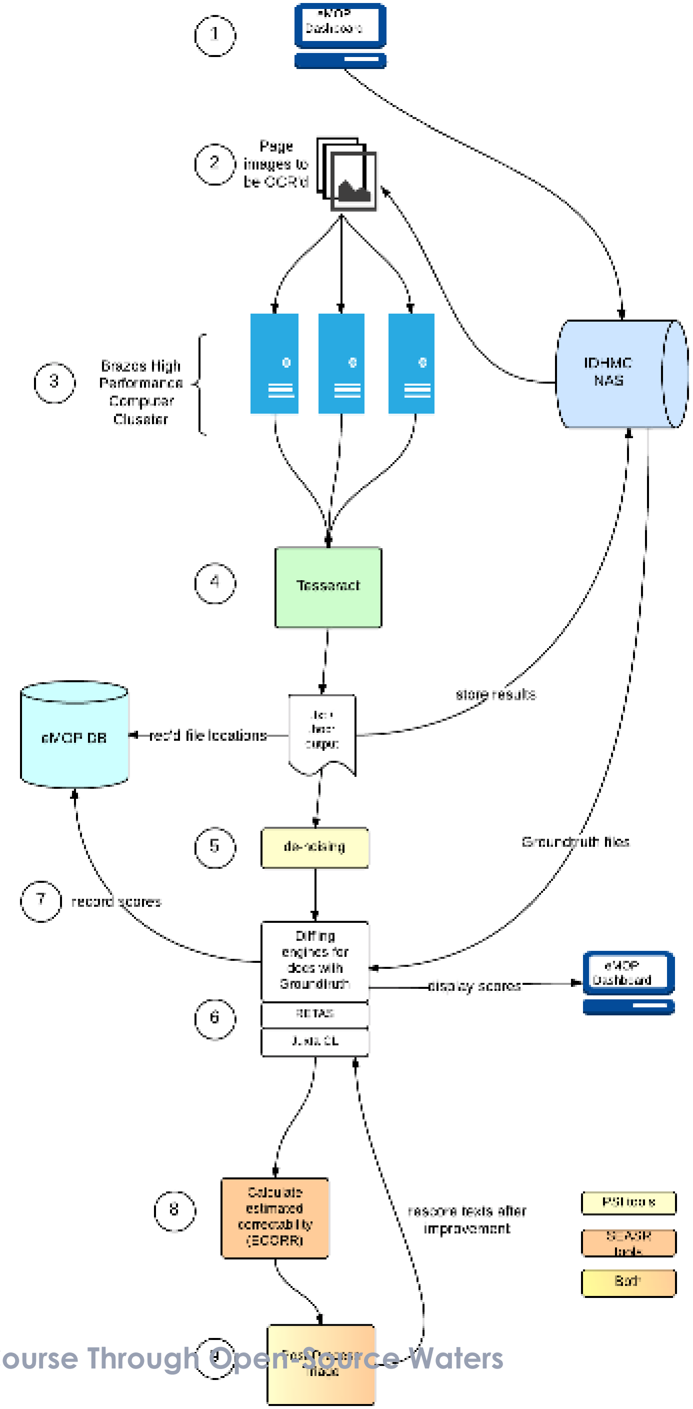


emop_controller

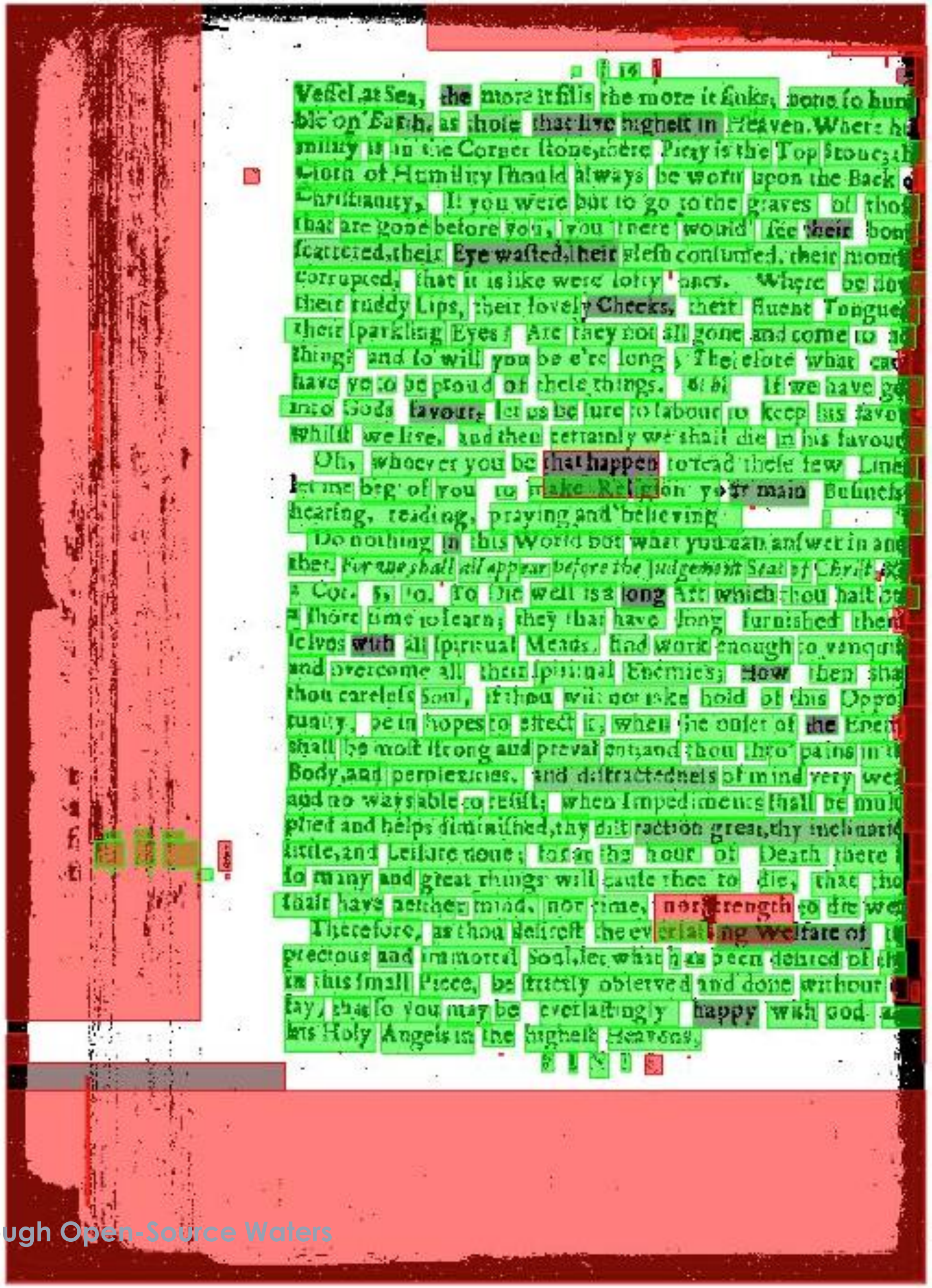
A java program that runs on the Brazos High Performance Computing Cluster and ensures maximum utilization of the 128 processors available for our use by scheduling various functions and processes of the controller separately.

1. The selected documents are marked in DB as scheduled and the documents are queued for processing.
2. A cron job continuously checks the scheduled queue and OCRs any unscheduled pages.
3. Each page is OCRd with Tesseract.
4. After OCRing each page's hOCR output file is de-noised.
5. hOCR pages that have matching Groundtruth are scored using Juxta and RETAS algorithms.
6. File paths and scores are written to the eMOP DB.
7. hOCR documents are examined again and given an estimated correctability (ECORR) score.
8. hOCR page results are sent to post-processing triage.

emop_controller



De-noising



[16]

Veſſel at Sea, the more it fills the more it ſinks, none ſo hum
ble on Earth, as thoſe that live higheſt in Heaven. Where hu
mility is in the Corner ſtone, there Piety is the Top ſtone; th
Cloth of Humility ſhould always be worn upon the Back o
Chriſtianity. If you were but to go to the graves of thoſ
that are gone before you, you there would ſee their bon
ſcattered, their Eye waſted, their Fleſh conſumed, their mou
corrupted, that it is like were lofty Faces. Where be no
their ruddy Lips, their lovely Cheeks, their fluent Tongue
their ſparkling Eyes? Are they not all gone and come to n
thing? and ſo will you be e're long; Therefore what ca
have ye to be proud of theſe things. If we have ge
into Gods favour, let us be ſure to labour to keep his favo
whilſt we live, and then certainly we ſhall die in his favou

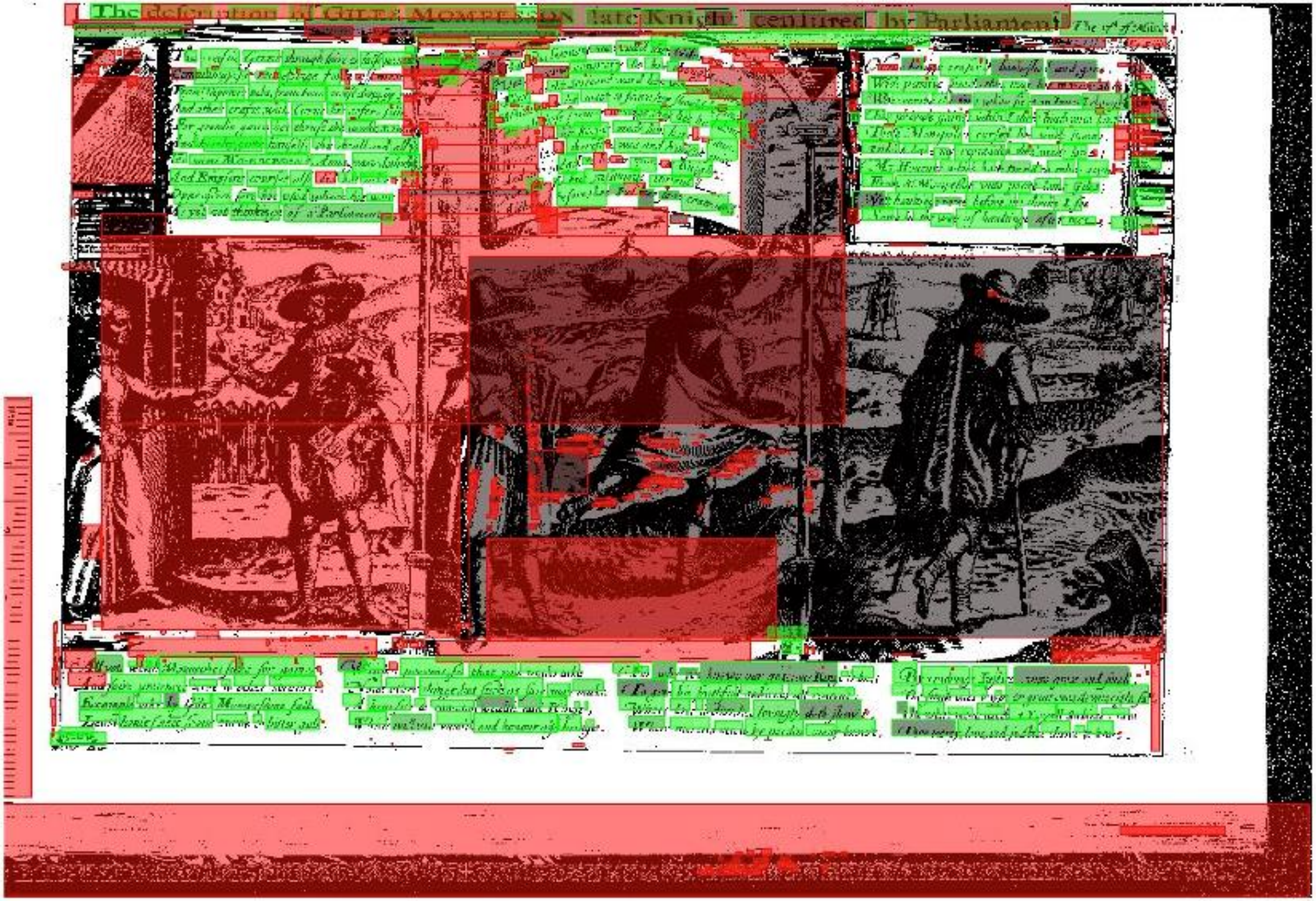
Oh, whoever you be that happen to read theſe few Line
let me beg of you to make Religion your main Buſineſs
hearing, reading, praying and believing.

Do nothing in this World but what you can anſwer in an
other. *For we ſhall all appear before the Judgement Seat of Chriſt*,
2 Cor. 5. 10. To Die well is a long Art which thou haſt bu
a ſhort time to learn; they that have long furniſhed them
ſelves with all ſpiritual Means, find work enough to vanqui
and overcome all their ſpiritual Enemies; How then ſha
thou careleſs Soul, if thou wilt not take hold of this Oppo
tunity, be in hopes to effect it, when the onſet of the Enem
ſhall be moſt ſtrong and prevalent, and thou thro' pains in t
Body, and perplexities, and diſtractedneſs of mind very we
and no ways able to reſiſt; when Impediments ſhall be mul
plied and helps diminiſhed, thy diſtraction great, thy inclinatio
little, and Leiſure none; for in the hour of Death there
ſo many and great things will cauſe thee to die, that tho
ſhalt have neither mind, nor time, nor ſtrength to die we

Therefore, as thou deſireſt the everlaſting Welfare of th
precious and immortal Soul, let what has been deliver'd of th
in this ſmall Piece, be ſtrictly obſerved and done without
lay, that ſo you may be everlaſtingly happy with God a
his Holy Angels in the higheſt Heavens.

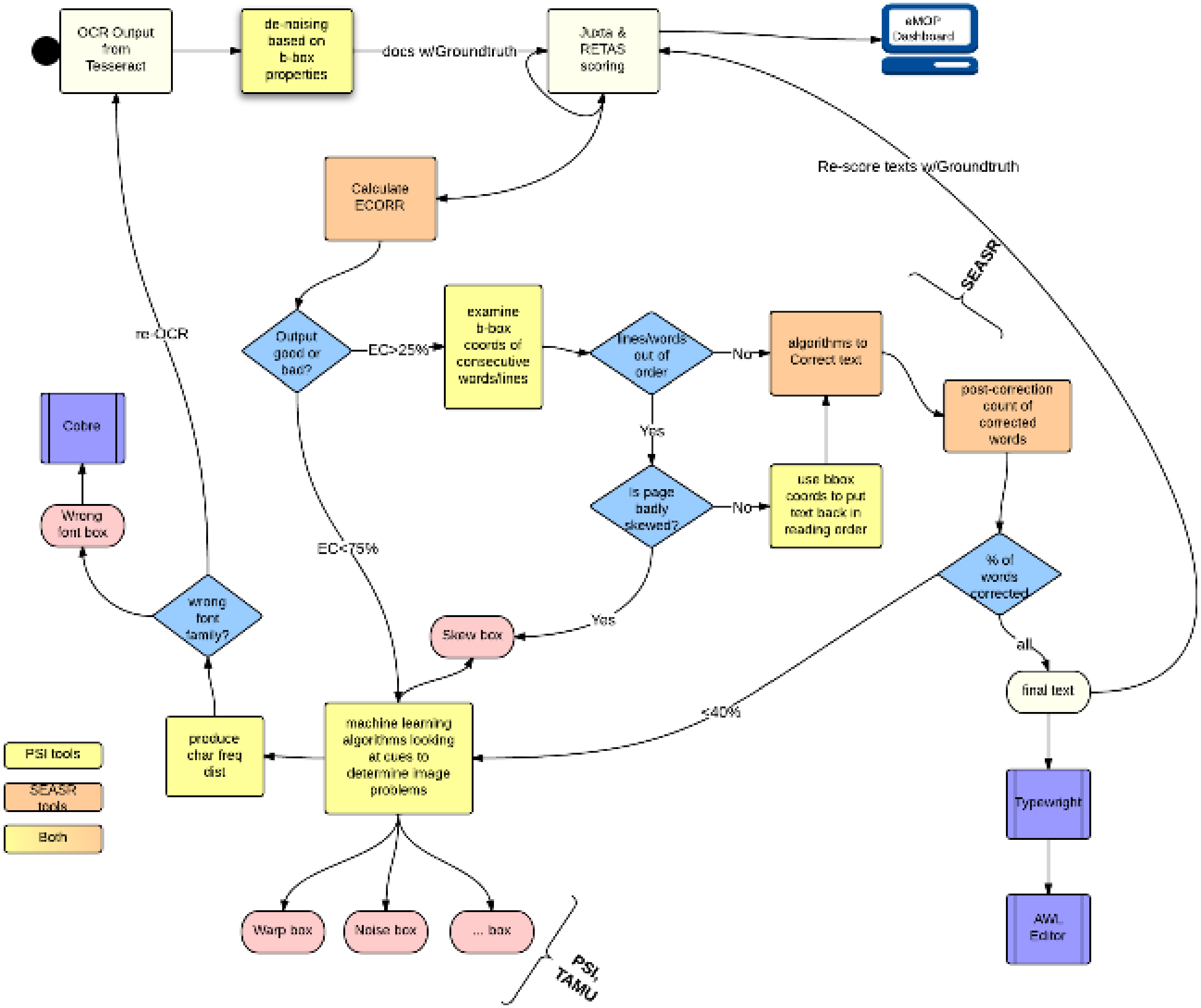
FINIS.

De-noising

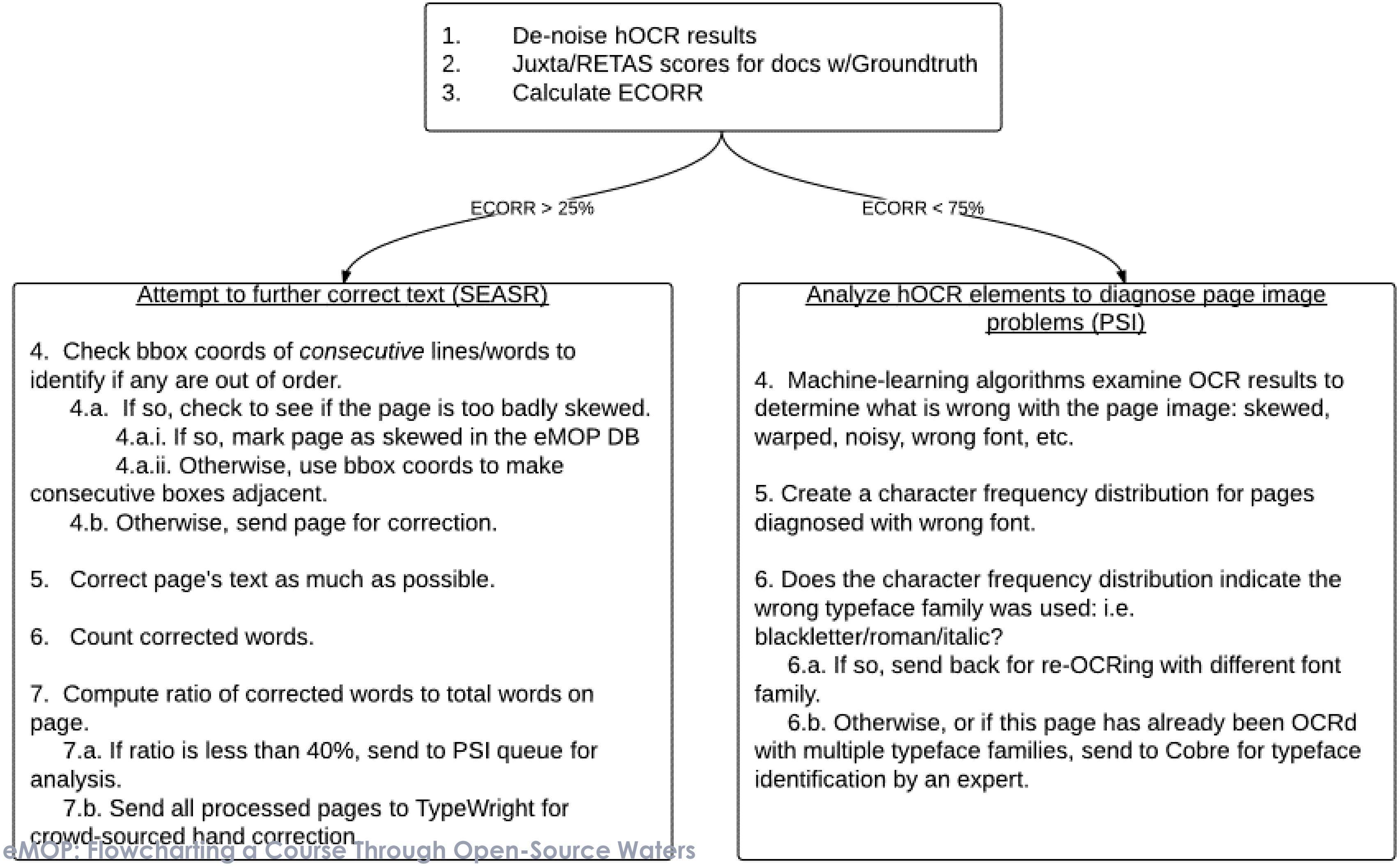


The description of Giles Mompesson late Knight censured by Parliament

Post-Processing Triage



Post-Processing Triage



1. De-noise hOCR results
2. Juxta/RETAS scores for docs w/Groundtruth
3. Calculate ECORR

ECORR > 25%

ECORR < 75%

Attempt to further correct text (SEASR)

4. Check bbox coords of *consecutive* lines/words to identify if any are out of order.

4.a. If so, check to see if the page is too badly skewed.

4.a.i. If so, mark page as skewed in the eMOP DB

4.a.ii. Otherwise, use bbox coords to make consecutive boxes adjacent.

4.b. Otherwise, send page for correction.

5. Correct page's text as much as possible.

6. Count corrected words.

7. Compute ratio of corrected words to total words on page.

7.a. If ratio is less than 40%, send to PSI queue for analysis.

7.b. Send all processed pages to TypeWright for crowd-sourced hand correction.

Analyze hOCR elements to diagnose page image problems (PSI)

4. Machine-learning algorithms examine OCR results to determine what is wrong with the page image: skewed, warped, noisy, wrong font, etc.

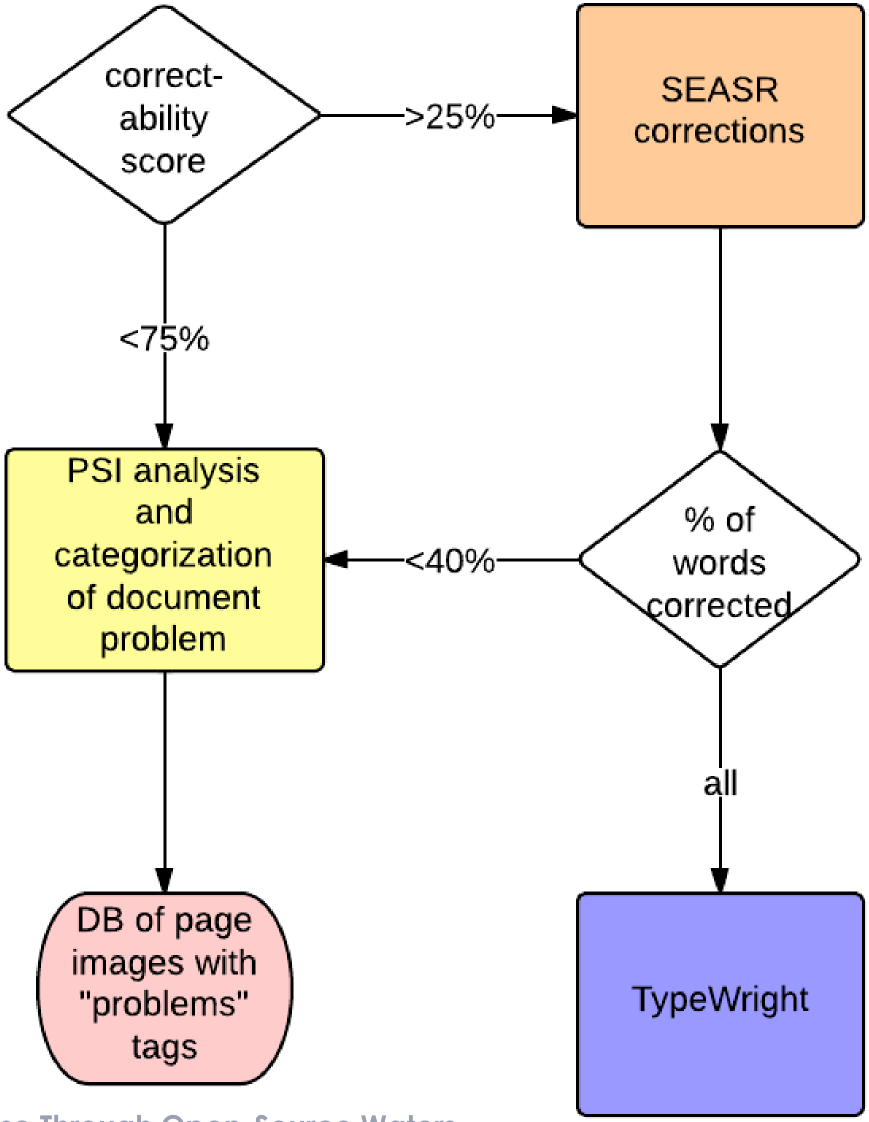
5. Create a character frequency distribution for pages diagnosed with wrong font.

6. Does the character frequency distribution indicate the wrong typeface family was used: i.e. blackletter/roman/italic?

6.a. If so, send back for re-OCRing with different font family.

6.b. Otherwise, or if this page has already been OCRd with multiple typeface families, send to Cobre for typeface identification by an expert.

Post-Processing Triage


correct-ability score

>25%

SEASR corrections

<75%

PSI analysis and categorization of document problem

<40%

% of words corrected

all

DB of page images with "problems" tags

TypeWright

TypeWright

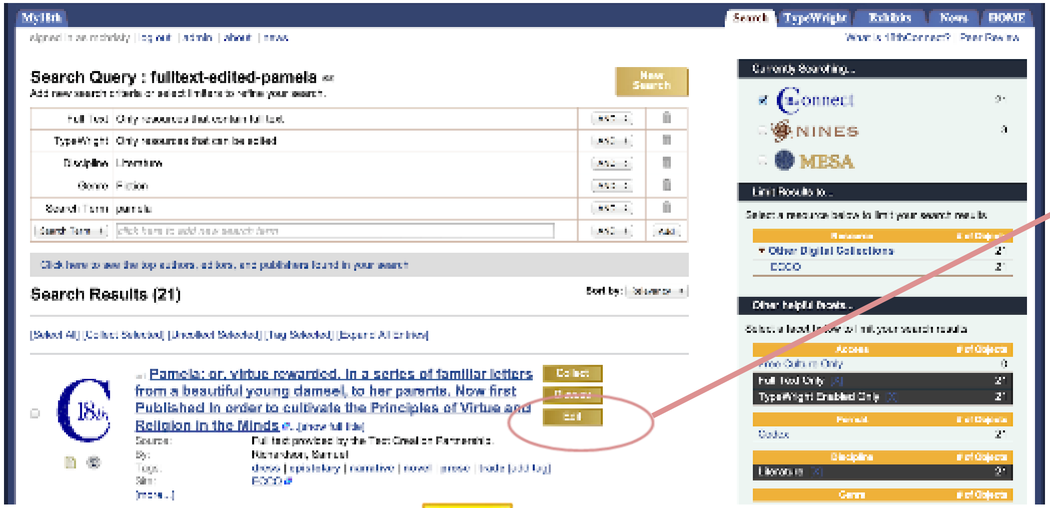
- www.18thconnect.org/typewright/
- Available through 18thconnect.org

  An aggregator of 18th Century metadata & content

This will be the first time that EEBO documents are available as searchable text, and that they can be corrected by scholars for their own use.

- Crowd-sourced correction tool for OCR.
- Currently available for ECCO collection, based on original ECCO OCR results.
- Will be ingested with eMOP OCR for ECCO and EEBO by year-end.
- Users can correct a whole document and then receive the text or a lightly encoded TEI XML version for use in a digital edition or whatever.

TypeWright



Search Query : fulltext-edited-pamela

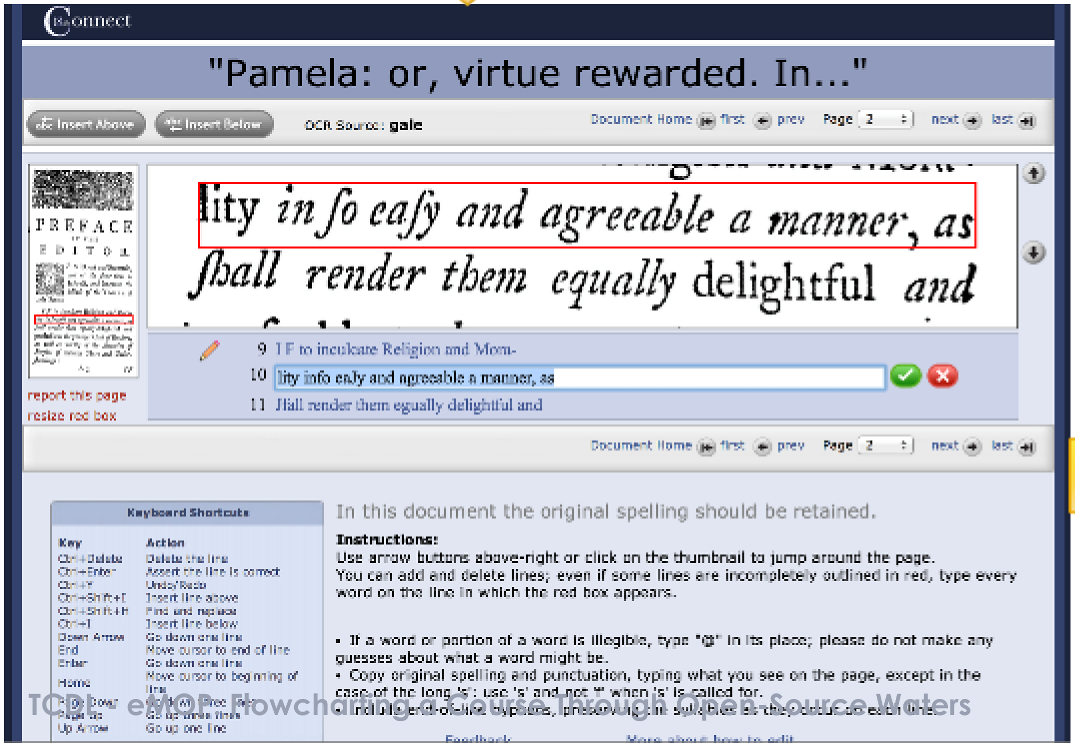
Full Text	Only resources that contain full text	
TypeWright	Only resources that can be edited	
Discipline	Literature	
Genre	Fiction	
Search Term	pamela	

Search Results (21)

Pamela: or, virtue rewarded. In a series of familiar letters from a beautiful young damsel, to her parents. Now first Published In order to cultivate the Principles of Virtue and Religion in the Minds

Source: Full text provided by the Text Creation Partnership
By: Richardson, Samuel

Edit button identifies "TypeWright" enabled Texts.



"Pamela: or, virtue rewarded. In..."

OCR Source: gale

9 I F to inculcate Religion and Mora-

10 lity info eafy and agreeable a manner, as

11 Ihall render them equally delightful and

In this document the original spelling should be retained.

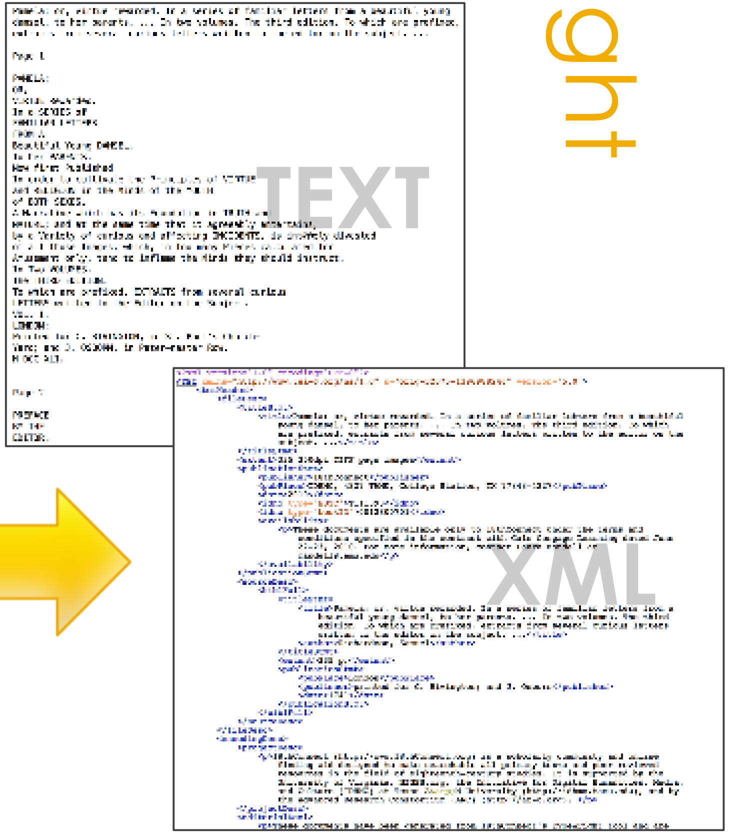
Instructions:
Use arrow buttons above-right or click on the thumbnail to jump around the page.
You can add and delete lines; even if some lines are incompletely outlined in red, type every word on the line in which the red box appears.

- If a word or portion of a word is illegible, type "@" in its place; please do not make any guesses about what a word might be.
- Copy original spelling and punctuation, typing what you see on the page, except in the case of the long 's': use 's' and not 'ſ' when 's' is called for.
- Include end-of-line hyphens, preserving the syllables as they occur on each line.

Keyboard Shortcuts

Key	Action
Ctrl+Delete	Delete the line
Ctrl+Enter	Assert the line is correct
Ctrl+Y	Undo/Redo
Ctrl+Shift+I	Insert line above
Ctrl+Shift+H	Find and replace
Ctrl+I	Insert line below
Down Arrow	Go down one line
End	Move cursor to end of line
Enter	Go down one line
Home	Move cursor to beginning of line
Page Down	Go down three lines
Page Up	Go up three lines
Up Arrow	Go up one line

18thConnect
Insert Above
Insert Below
Document Home
first
prev
Page 2
next
last
report this page
resize red box
Feedback
More about how to edit



TEXT
XML

Tags DB

- All page images that don't produce OCR text that meets our standards will be tagged in the eMOP DB with tags that indicate what is wrong with the page image.
  - skew, warp, noise, wrong font, etc.
- Tags can then be used to later apply the appropriate pre-processing techniques and then re-OCR the page to produce better results.
- **This will be the first time that any sort of comprehensive analysis has been done on the page images of these collections.**
- We'll end up (ideally) with a list of page images that simply need to be re-scanned.

The end

For eMOP questions please contact me at :

mchristy@tamu.edu