IDHMC
TEXAS A&M UNIVERSITY
Initiative for Digital Humanities, Media, and Culture

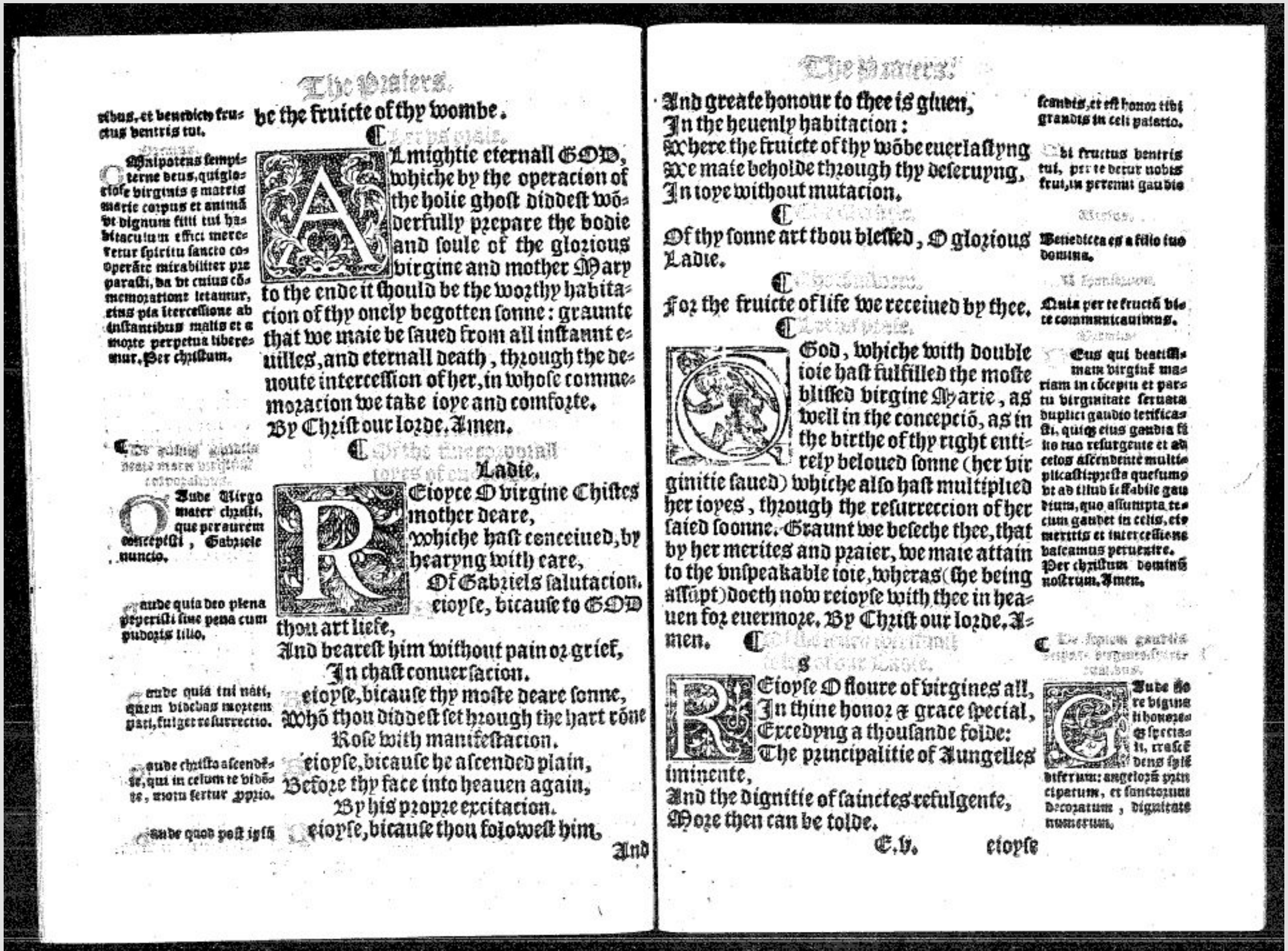
Early Modern OCR Project (eMOP) at Texas A&M University: Using Aletheia to Train Tesseract

Katayoun Torabi, PhD Student in English
Bryan Tarpley, PhD Student in English
Dr. Jessica Durgen, Post-Doctoral Fellow (not present)

The Problem:

- The IDHMC is tasked with performing OCR on over 45,000,000 page images of historical documents which were scanned using poor resolution scanners with problems such as idiosyncratic fonts and bleed-through.
- In-house, we have at our disposal 2 programmers and ~ 5 part time students in the English department and the Digital Humanities. We are also privileged to work with various collaborators, including PRImA.
- We may only work with open-source solutions (no ABBY FineReader).
- The time frame for our project is approximately 2 years.

Historical Documents



tibus, et benedictus fructus uentris tui.

Oremus.

Omnipotens sempiterne deus, qui gloriose virginis & matris Marie corpus et animã ut dignum filii tui habitaculum effici mereretur spiritu sancto cooperãte mirabiliter preparasti, da ut cuius cõmemoratione letamur, eius pia intercessione ab instantibus malis et a morte perpetua liberemur. Per christum.

be the fruicte of thy wombe.

¶ Let vs praie.

Almightie eternall GOD, whiche by the operacion of the holie ghost diddest wonderfully prepare the bodie and soule of the glorious virgine and mother Mary to the ende it should be the worthy habitacion of thy onely begotten sonne: graunte that we maie be saued from all instaunt euilles, and eternall death, through the deuoute intercession of her, in whose commemoracion we take ioye and comforte. By Christ our lorde. Amen.

¶ De quinque gaudiis beate marie virginis corporalibus.

¶ Of the fiue corporall ioyes of our Ladie.

Gaude Uirgo mater christi, que per aurem concepisti, Gabriele nuncio.

Reioyce O virgine Chistes mother deare,
whiche hast conceiued, by hearyng with care,
Of Gabriels salutacion.

Gaude quia deo plena peperisti sine pena cum pudoris lilio.

Reioyse, bicause to GOD thou art liefe,
And bearest him without pain or grief,
In chast conuersacion.

Gaude quia tui nati, quem videbas mortem pati, fulget resurrectio.

Reioyse, bicause thy moste deare sonne,
Whõ thou diddest se through the hart rõne
Rose with manifestacion.

Gaude christo ascendẽte, qui in celum te vidẽte, motu fertur proprio.

Reioyse, bicause he ascended plain,
Before thy face into heauen again,
By his propre excitacion.

Gaude quod post ipsũ

Reioyse, bicause thou folowest him,
And

scandis, et est honor tibi grandis in celi palatio.

And greate honour to thee is giuen,
In the heuenly habitacion:

Ubi fructus uentris tui, per te detur nobis frui, in perenni gaudio.

Where the fruicte of thy wõbe euerlastyng
We maie beholde through thy deseruyng,
In ioye without mutacion.

Uersus.

¶ The Uerse.

Benedicta es a filio tuo domina.

Of thy sonne art thou blessed, O glorious Ladie.

Responsum.

¶ The Answere.

Quia per te fructũ vite communicauimus.

For the fruicte of life we receiued by thee.

Oremus.

¶ Let vs praie.

Deus qui beatissimam virginẽ mariam in cõceptu et partu virginitate seruata duplici gaudio letificasti, quique eius gaudia filio tuo resurgente et ad celos ascendente multiplicasti: presta quesumus ut ad illud ineffabile gaudium, quo assumpta tecum gaudet in celis, eius meritis et intercessione valeamus peruenire. Per christum dominũ nostrum. Amen.

God, whiche with double ioie hast fulfilled the moste blissed virgine Marie, as well in the concepciõ, as in the birthe of thy right entirely beloued sonne (her virginitie saued) whiche also hast multiplied her ioyes, through the resurreccion of her saied soonne. Graunt we beseche thee, that by her merites and praier, we maie attain to the vnspeakable ioie, wheras (she being assũpt) doeth now reioyse with thee in heauen for euermore. By Christ our lorde. Amen.

¶ De septem gaudiis spiritualibus.

¶ Of the spirituall ioyes of our Ladie.

Gaude flore virginali, honore speciali, transcendens splendiferum: angelorum principatum, et sanctorum decoratum, dignitate numerum.

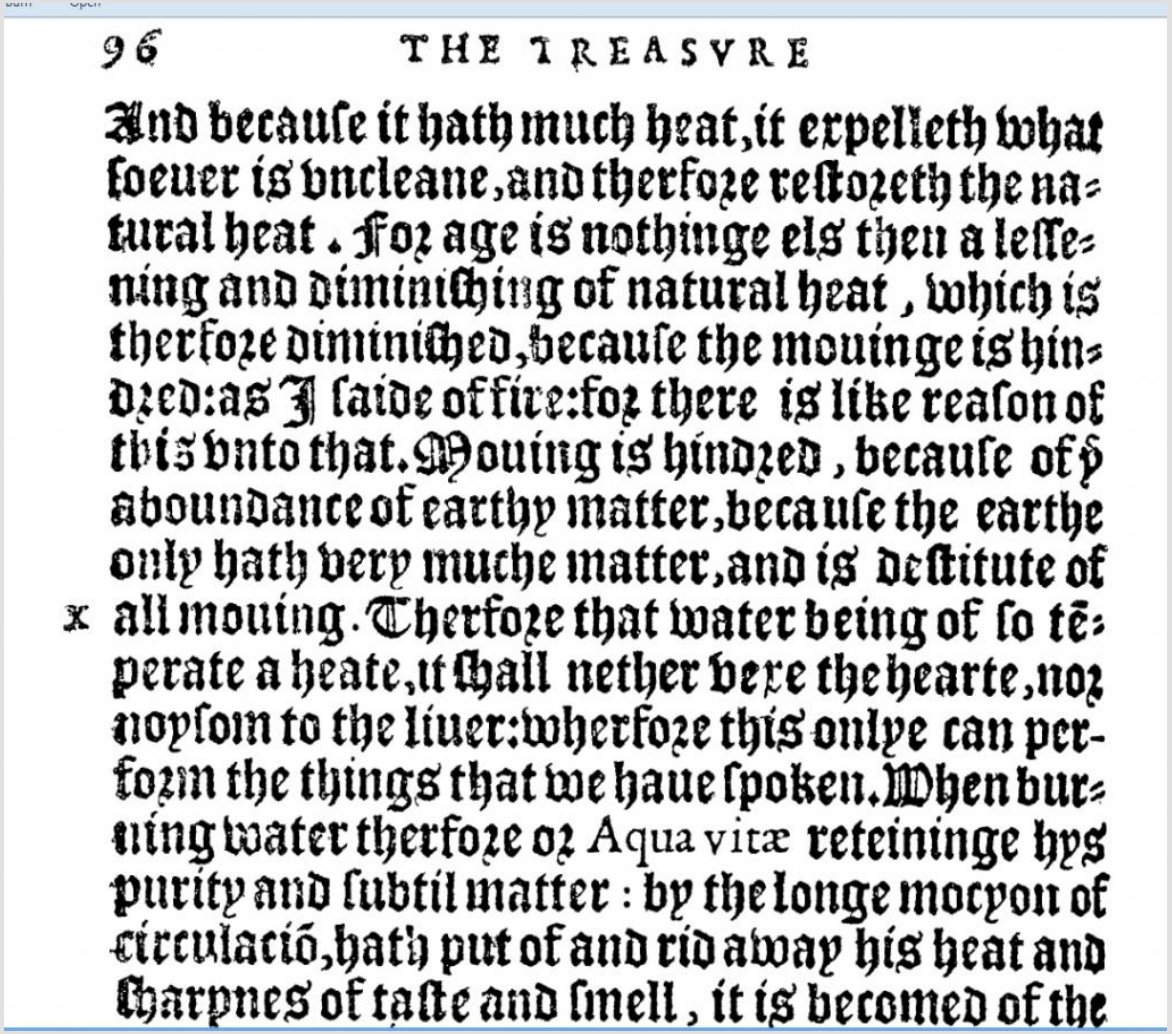
Reioyse O floure of virgines all,
In thine honor & grace special,
Excedyng a thousande folde:
The principalitie of Aungelles iminente,
And the dignitie of sainctes refulgente,
More then can be tolde.

E.v. Reioyse

Ligatures

Italics

Blackletter

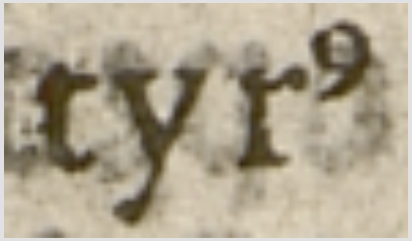


96 THE TREASVRE

And becauſe it hath much heat, it expelleth what ſoeuer is vncleane, and therfore reſtoreth the natural heat. For age is nothinge els then a leſſening and diminiſhing of natural heat, which is therfore diminiſhed, becauſe the mouinge is hindred: as I ſaide of fire: for there is like reaſon of this vnto that. Mouing is hindred, becauſe of ꝑ aboundance of earthy matter, becauſe the earthe only hath very muche matter, and is deſtitute of
x all mouing. Therfore that water being of ſo tẽperate a heate, it ſhall nether vexe the hearte, nor noyſom to the liuer: wherfore this onlye can perform the things that we haue ſpoken. When burninge water therfore or Aqua vitæ reteininge hys purity and ſubtil matter: by the longe mocyon of circulaciõ, hath put of and rid away his heat and ſharpnes of taſte and ſmell, it is becomed of the

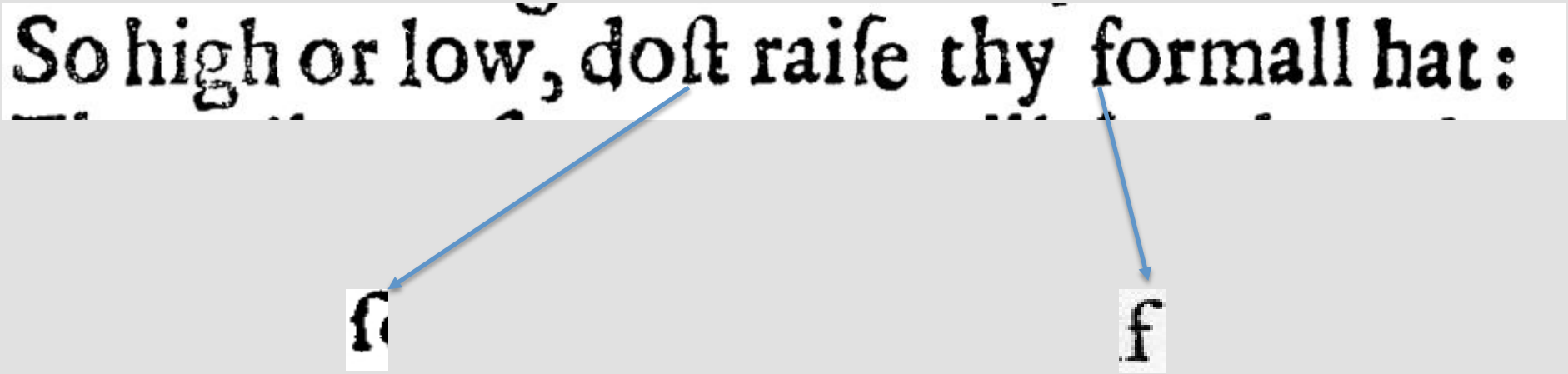
Other Special Characters

- Other special characters encountered are letters with suspension marks, rotunda r’s, and letters with superscripts



Long ‘s’ in the word ‘greatnesse’

Long ‘s’ and ‘f’



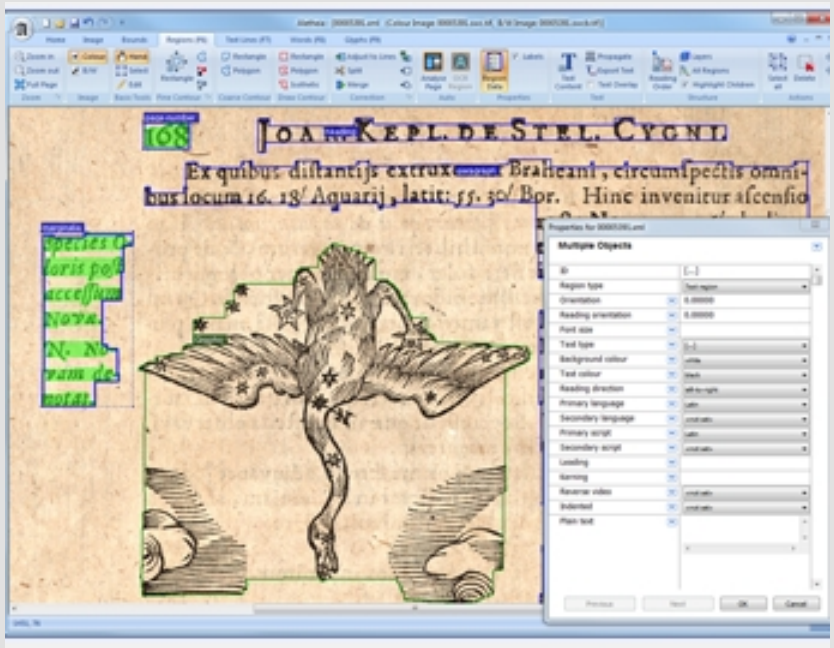
‘s’ with a half cross-bar

‘f’ with a full cross-bar

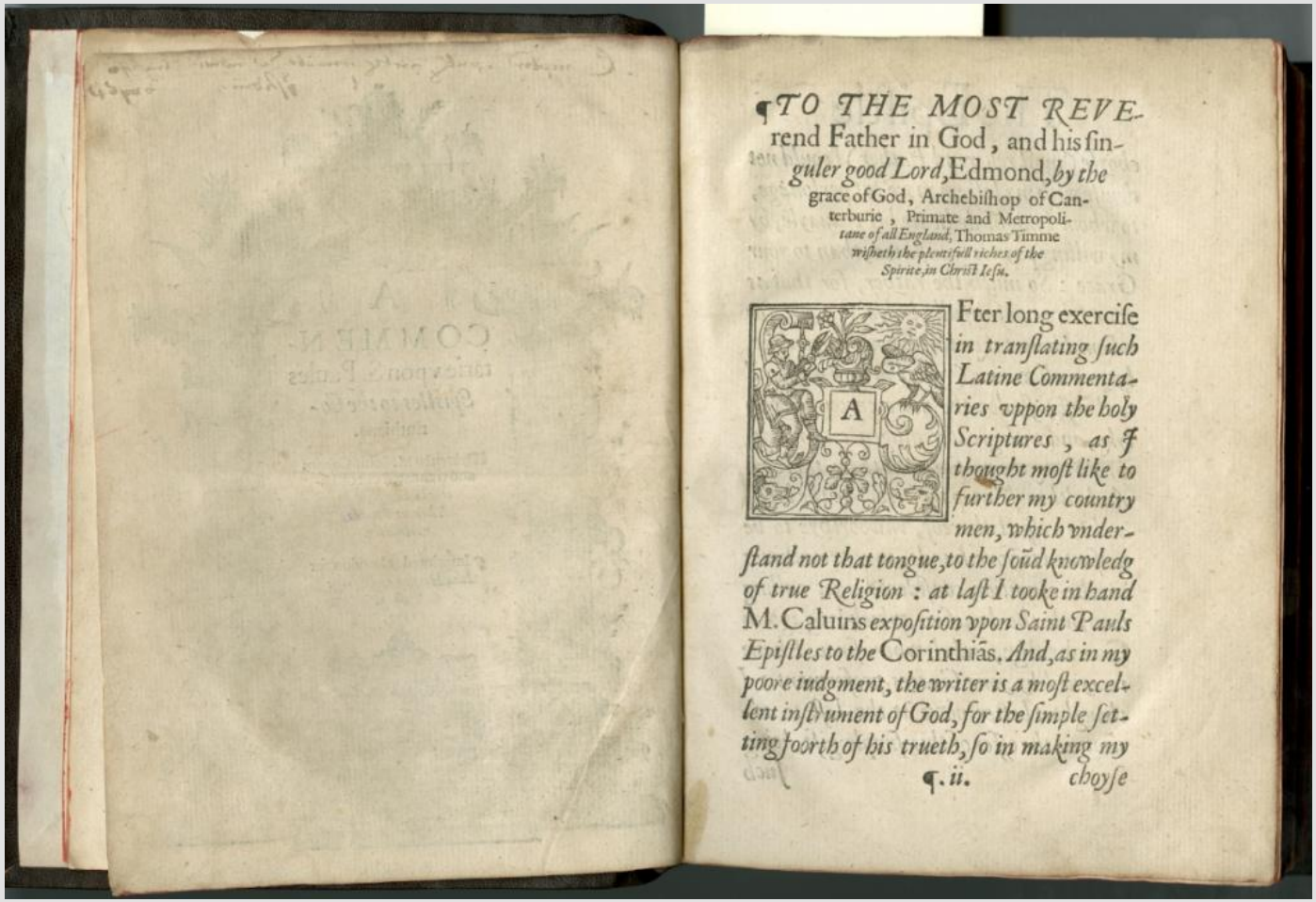
PRImA, the developers of Aletheia Desktop

University of Salford
A Greater Manchester University

PRImA
Pattern Recognition & Image Analysis Research



TIFF Page Image

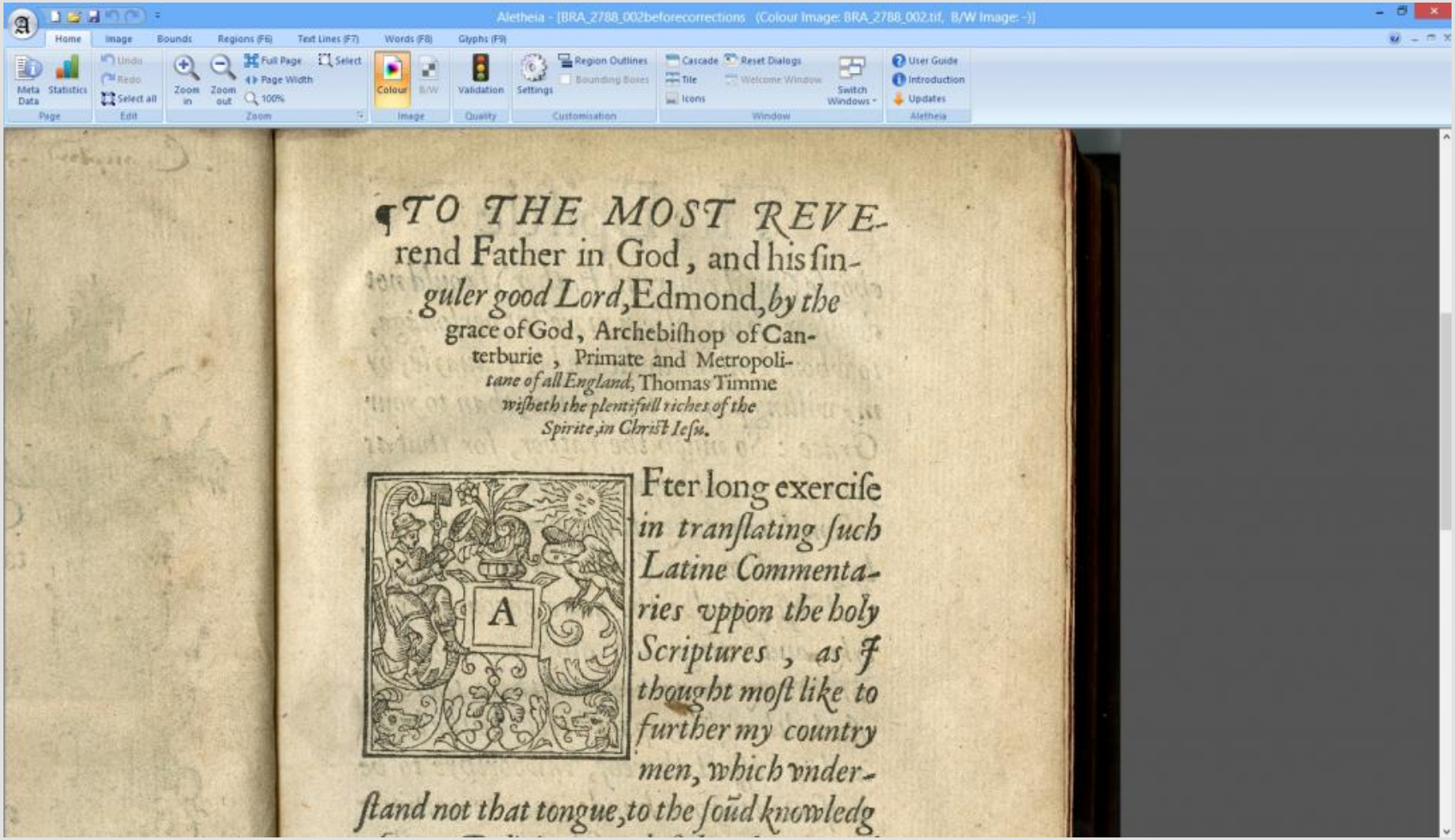


¶TO THE MOST REVE-
rend Father in God, and his ſin-
guler good Lord, Edmond, by the
grace of God, Archebiſhop of Can-
terburie, Primate and Metropoli-
tane of all England, Thomas Timme
wiſheth the plentifull riches of the
Spirite in Chriſt Ieſu.

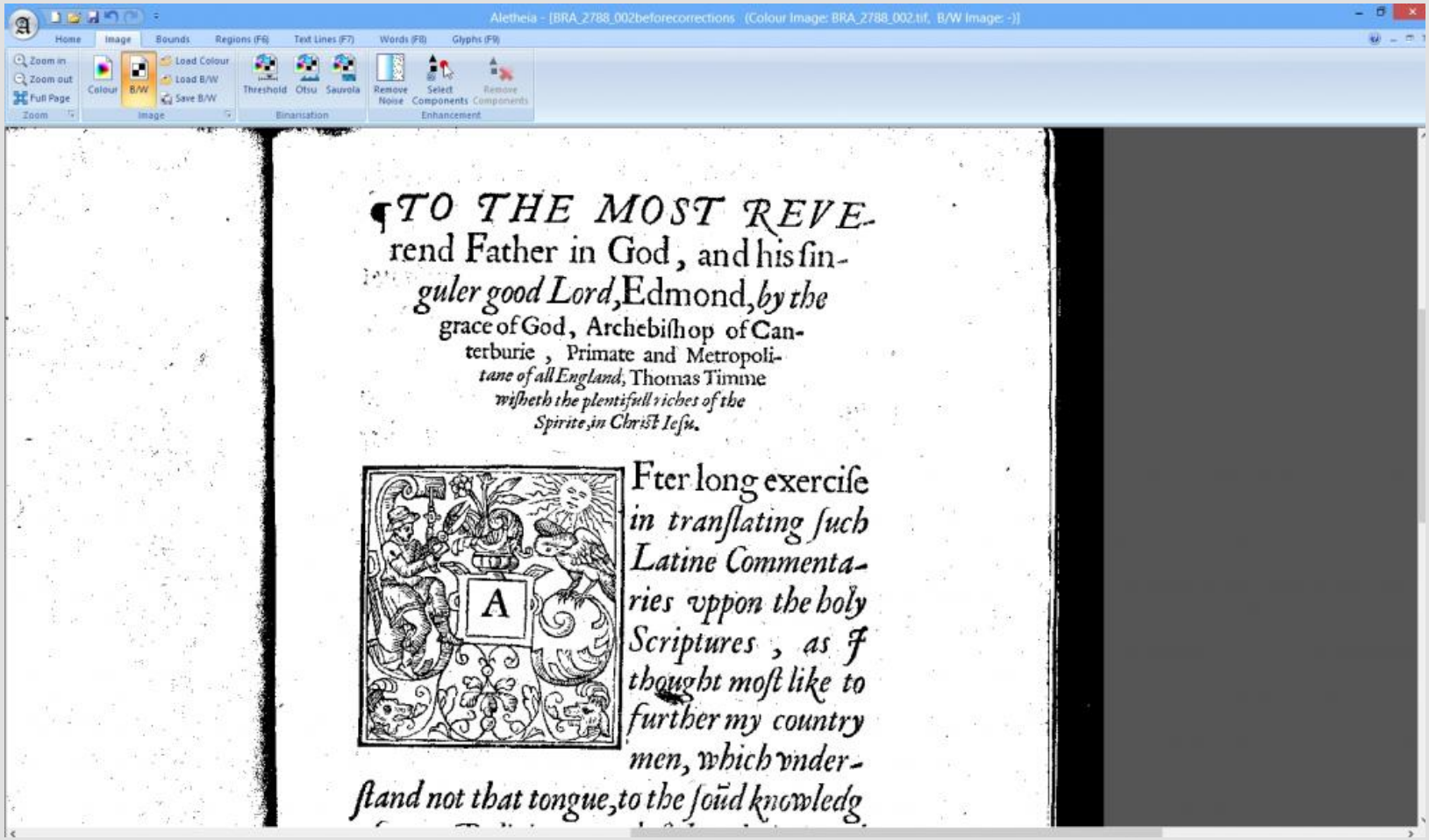
AFter long exerciſe in tranſlating ſuch Latine Commenta-ries vppon the holy Scriptures, as I thought moſt like to further my country men, which vnder-ſtand not that tongue, to the ſoũd knowledg of true Religion: at laſt I tooke in hand M. Caluins expoſition vpon Saint Pauls Epiſtles to the Corinthiãs. And, as in my poore iudgment, the writer is a moſt excel-lent inſtrument of God, for the ſimple ſet-ting foorth of his trueth, ſo in making my

¶. ii. choyſe

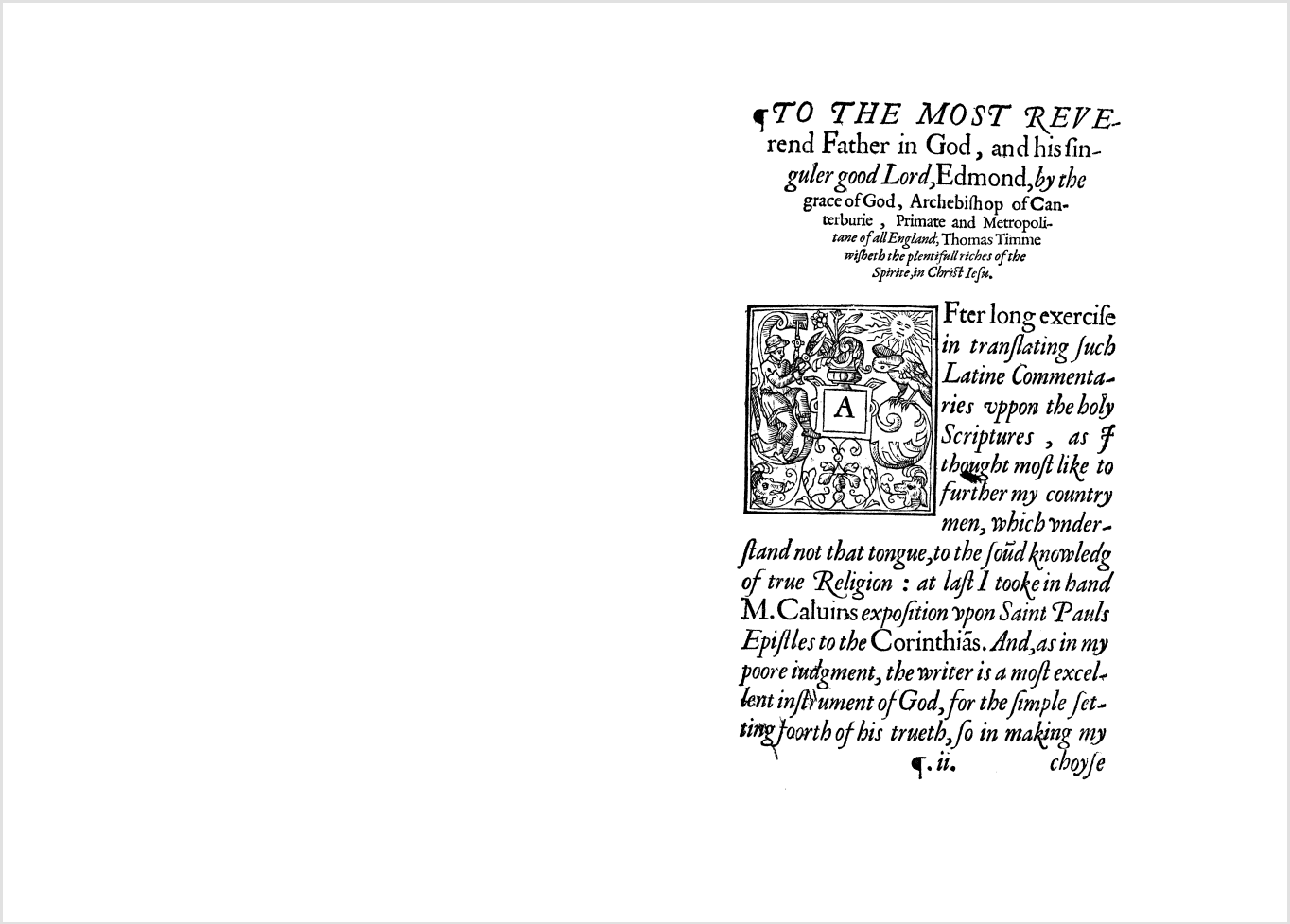
Step 1: Upload the TIFF image in Aletheia



Step 2: Binarize the image



Step 3: Remove noise

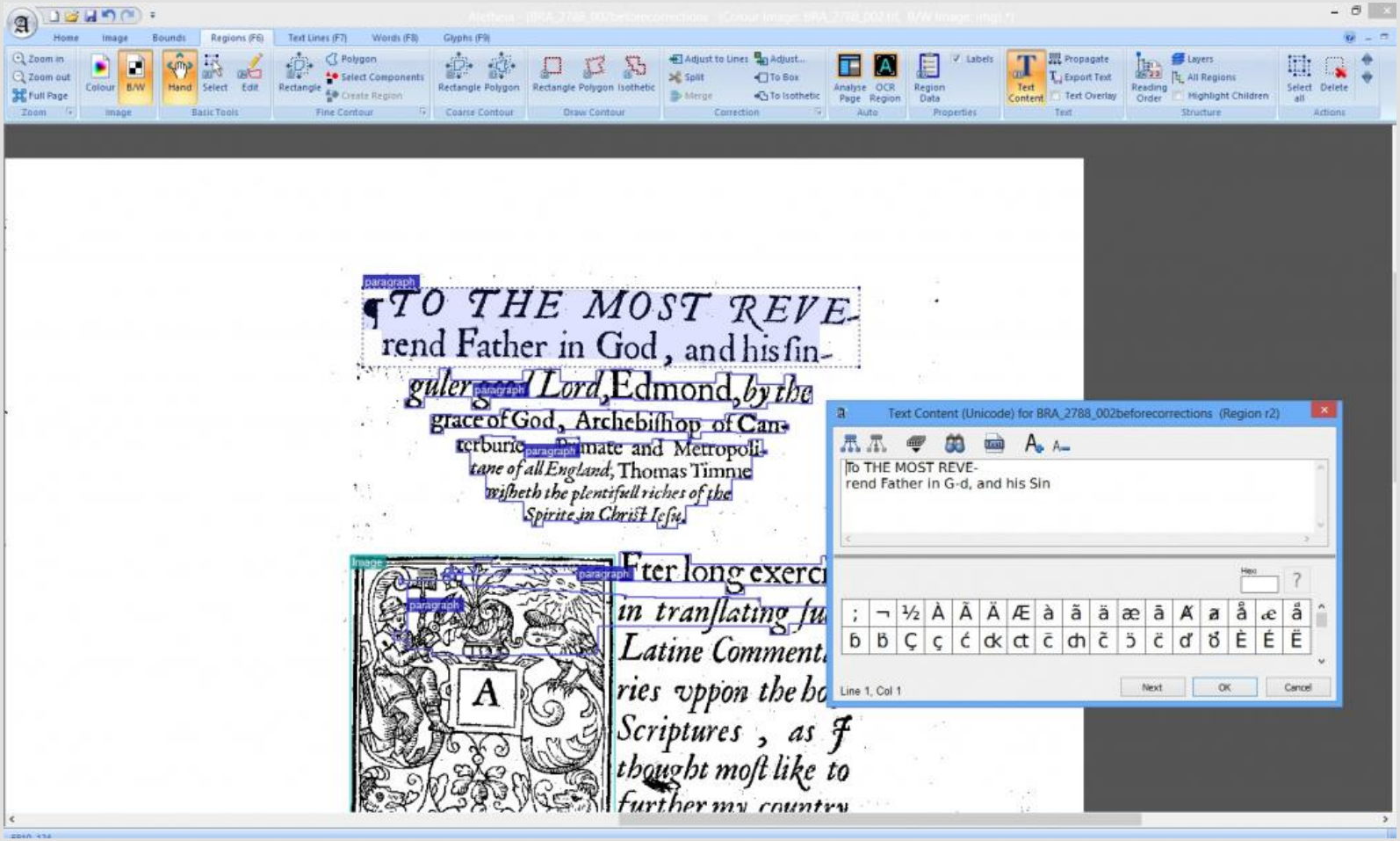


¶ *TO THE MOST REVE-*
rend Father in God, and his ſin-
guler good Lord, Edmond, *by the*
grace of God, Archebiſhop of Can-
terburie, Primate and Metropoli-
tane of all England, Thomas Timme
wiſheth the plentifull riches of the
Spirite, in Chriſt Ieſu.

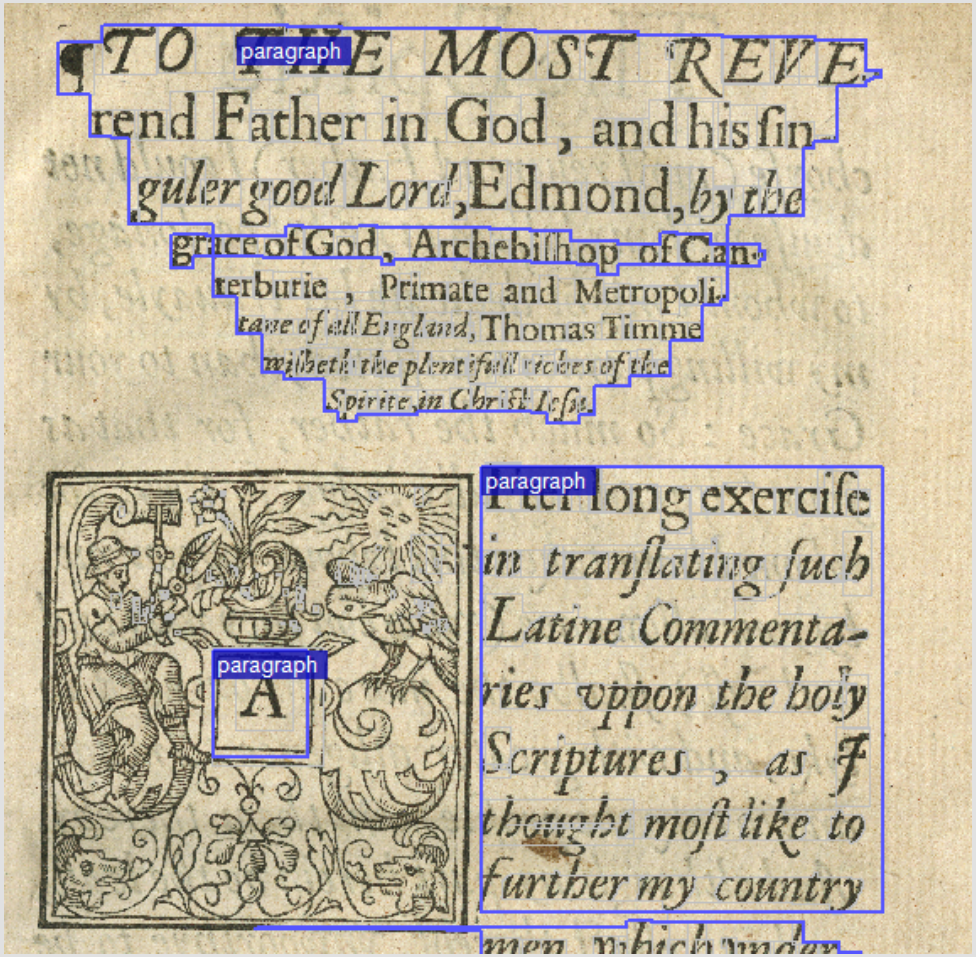
AFter long exerciſe *in tranſlating ſuch Latine Commenta-ries vppon the holy Scriptures, as I thought moſt like to further my country men, which vnder-ſtand not that tongue, to the ſoũd knowledg of true Religion: at laſt I tooke in hand* M. Caluins *expoſition vpon Saint Pauls Epiſtles to the* Corinthiãs. *And, as in my poore iudgment, the writer is a moſt excel-lent inſtrument of God, for the ſimple ſet-ting foorth of his trueth, ſo in making my*

¶. ii. *choyſe*

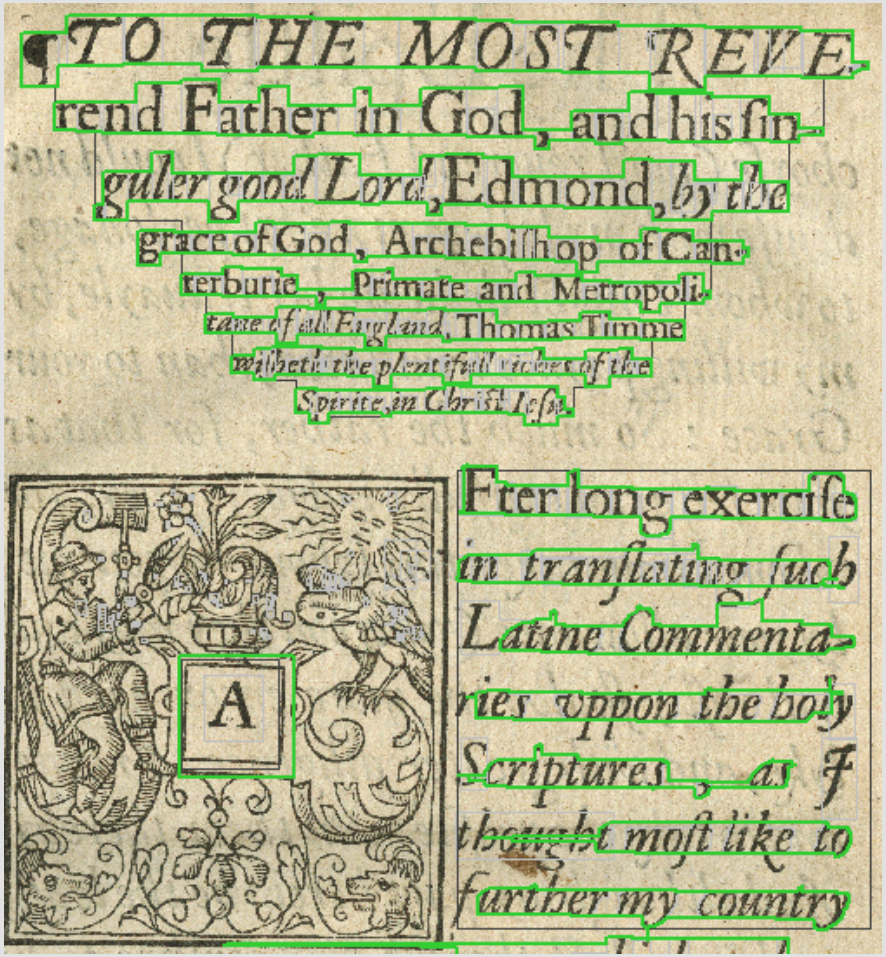
Step 4: Running the automatic segmentation tool in Aletheia



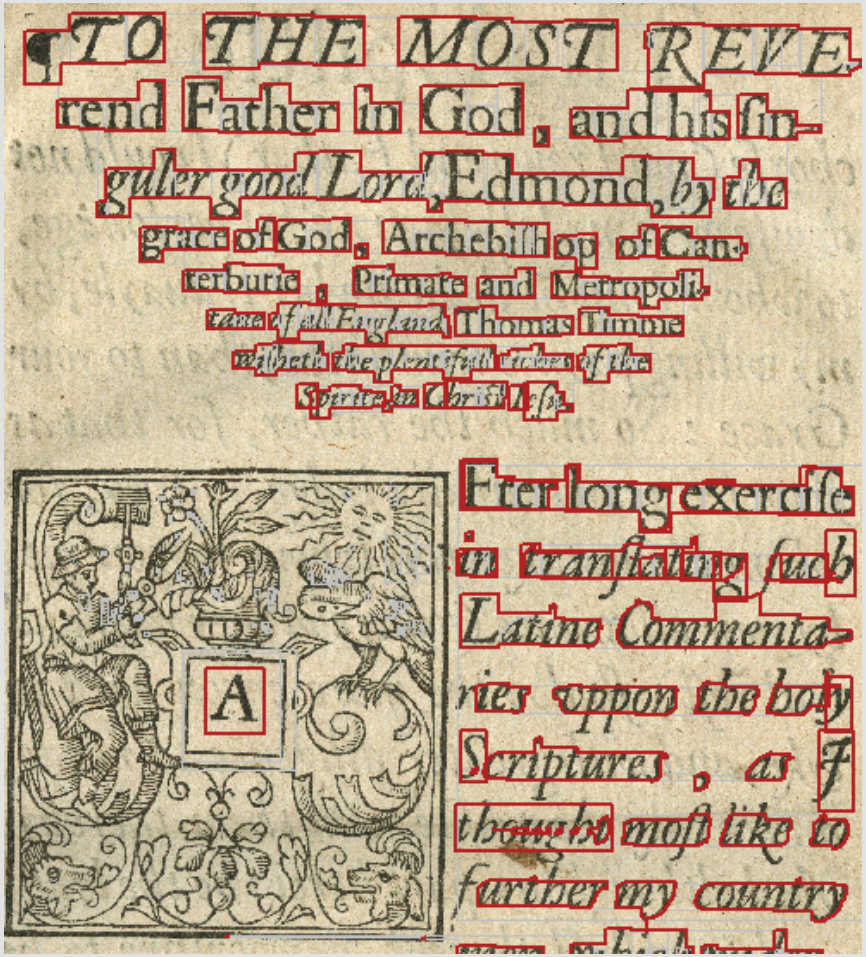
Regions


paragraph
TO THE MOST REVE
rend Father in God, and his ſin
guler good Lord, Edmond, by the
grace of God, Archebiſhop of Can-
terburie, Primate and Metropoli-
tane of all England, Thomas Timme
wiſheth the plentifull riches of the
Spirite in Chriſt Ieſu.
paragraph
paragraph
A
paragraph
long exerciſe
in tranſlating ſuch
Latine Commenta-
ries vppon the holy
Scriptures, as I
thought moſt like to
further my country

Lines


TO THE MOST REVE-
rend Father in God, and his ſin-
guler good Lord, Edmond, by the
grace of God, Archebiſhop of Can-
terburie, Primate and Metropoli-
tane of all England, Thomas Timme
wiſheth the plentifull riches of the
Spirite, in Chriſt Ieſu.
A
Fter long exerciſe
in tranſlating ſuch
Latine Commenta-
ries vppon the holy
Scriptures, as I
thought moſt like to
further my country

Words

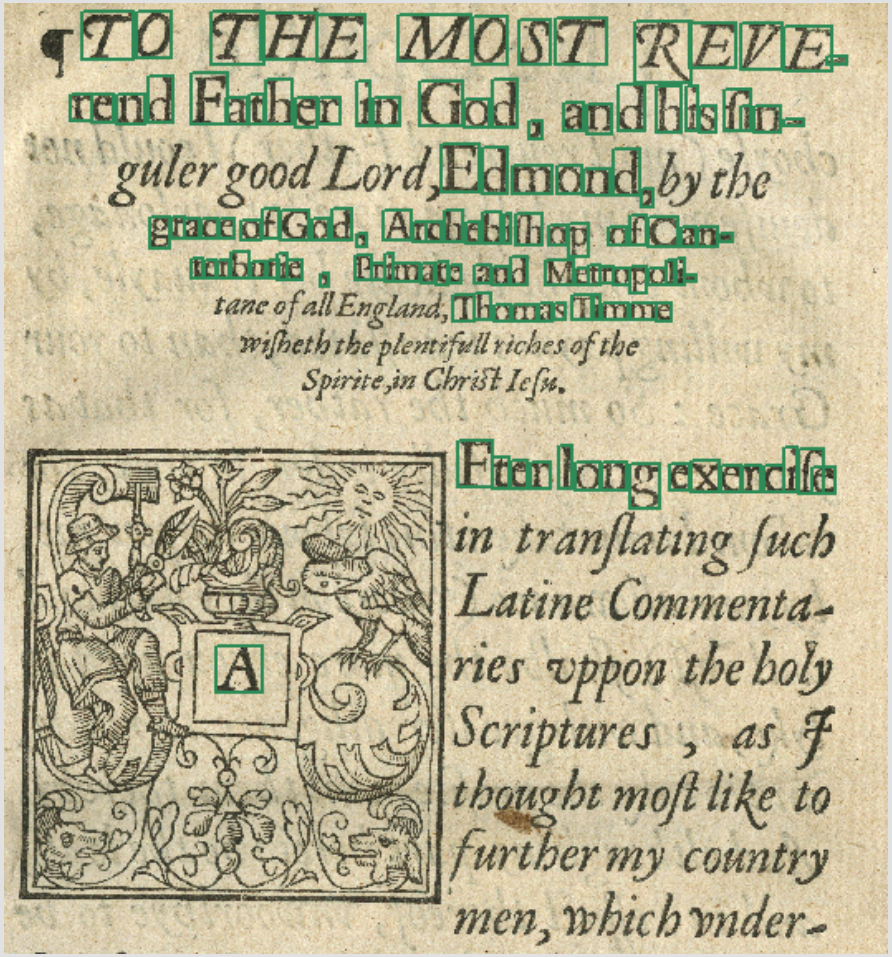


TO THE MOST REVE-
rend Father in God, and his fin-
guler good Lord, Edmond, by the
grace of God, Archebishop of Can-
terburie, Primate and Metropoli-
tane of all England, Thomas Timme
wisheth the plentifull riches of the
Spirite in Christ Jesu.

A

Fter long exercise
in translating such
Latine Commenta-
ries vppon the holy
Scriptures, as I
thought most like to
further my country

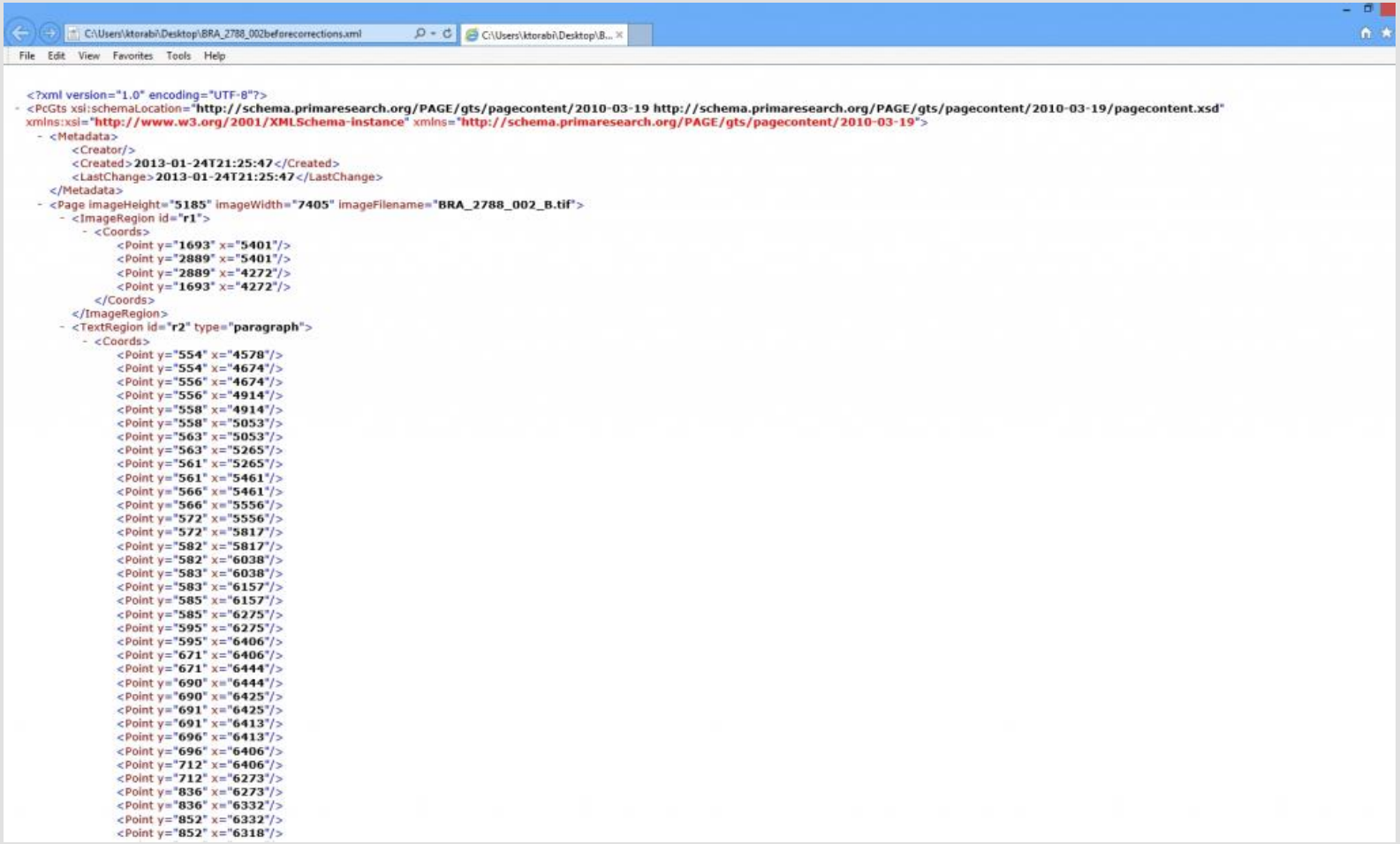
Glyphs



¶ TO THE MOST REVE-
rend Father in God, and his ſin-
guler good Lord, Edmond, by the
grace of God, Archebiſhop of Can-
torburie, Primate and Metropoli-
tane of all England, Thomas Timme
wiſheth the plentifull riches of the
Spirite, in Chriſt Ieſu.

A Ter long exercise
in tranſlating ſuch
Latine Commenta-
ries vppon the holy
Scriptures, as I
thought moſt like to
further my country
men, which vnder-

The segmentation tool generates text for each character and a corresponding XML file

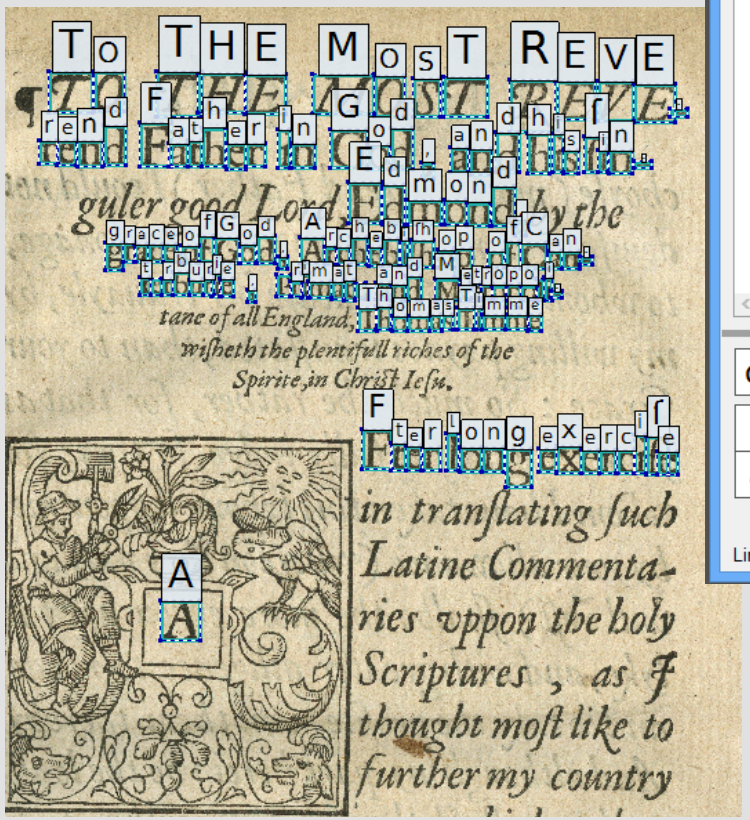


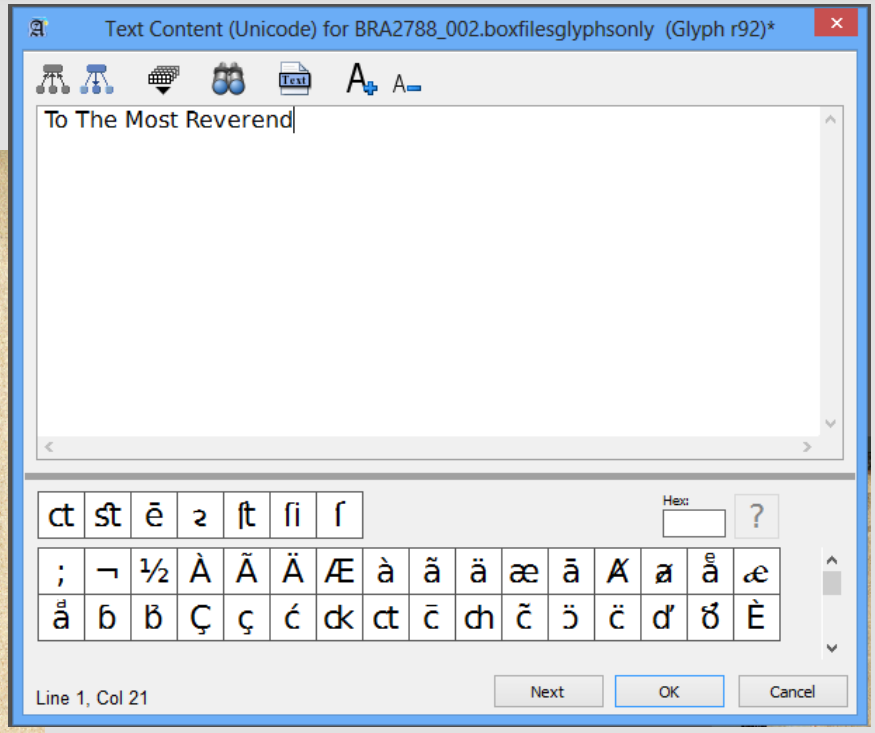
C:\Users\ktorabi\Desktop\BRA_2788_002beforecorrections.xml

File Edit View Favorites Tools Help

```
<?xml version="1.0" encoding="UTF-8"?>
<PcGts xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2010-03-19 http://schema.primaresearch.org/PAGE/gts/pagecontent/2010-03-19/pagecontent.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2010-03-19">
  <Metadata>
    <Creator/>
    <Created>2013-01-24T21:25:47</Created>
    <LastChange>2013-01-24T21:25:47</LastChange>
  </Metadata>
  <Page imageHeight="5185" imageWidth="7405" imageFilename="BRA_2788_002_B.tif">
    <ImageRegion id="r1">
      <Coords>
        <Point y="1693" x="5401"/>
        <Point y="2889" x="5401"/>
        <Point y="2889" x="4272"/>
        <Point y="1693" x="4272"/>
      </Coords>
    </ImageRegion>
    <TextRegion id="r2" type="paragraph">
      <Coords>
        <Point y="554" x="4578"/>
        <Point y="554" x="4674"/>
        <Point y="556" x="4674"/>
        <Point y="556" x="4914"/>
        <Point y="558" x="4914"/>
        <Point y="558" x="5053"/>
        <Point y="563" x="5053"/>
        <Point y="563" x="5265"/>
        <Point y="561" x="5265"/>
        <Point y="561" x="5461"/>
        <Point y="566" x="5461"/>
        <Point y="566" x="5556"/>
        <Point y="572" x="5556"/>
        <Point y="572" x="5817"/>
        <Point y="582" x="5817"/>
        <Point y="582" x="6038"/>
        <Point y="583" x="6038"/>
        <Point y="583" x="6157"/>
        <Point y="585" x="6157"/>
        <Point y="585" x="6275"/>
        <Point y="595" x="6275"/>
        <Point y="595" x="6406"/>
        <Point y="671" x="6406"/>
        <Point y="671" x="6444"/>
        <Point y="690" x="6444"/>
        <Point y="690" x="6425"/>
        <Point y="691" x="6425"/>
        <Point y="691" x="6413"/>
        <Point y="696" x="6413"/>
        <Point y="696" x="6406"/>
        <Point y="712" x="6406"/>
        <Point y="712" x="6273"/>
        <Point y="836" x="6273"/>
        <Point y="836" x="6332"/>
        <Point y="852" x="6332"/>
        <Point y="852" x="6318"/>
```

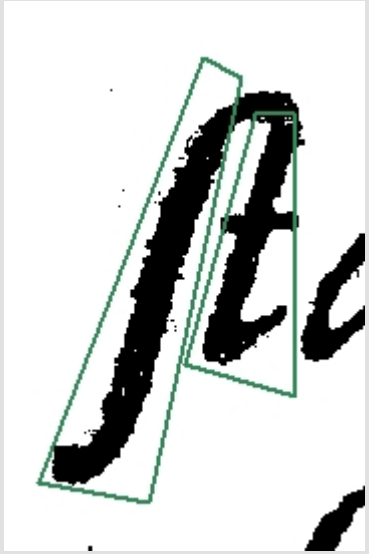
Text


To THE MOST REVE
F
G
E
guler good Lord
tane of all England
wiſheth the plentifull riches of the
Spirite in Chriſt Ieſu.
F
in tranſlating ſuch
Latine Commenta-
ries vppon the holy
Scriptures , as I
thought moſt like to
further my country


Text Content (Unicode) for BRA2788_002.boxfilesglyphsonly (Glyph r92)*
To The Most Reverend
Hex:
Line 1, Col 21
Next
OK
Cancel

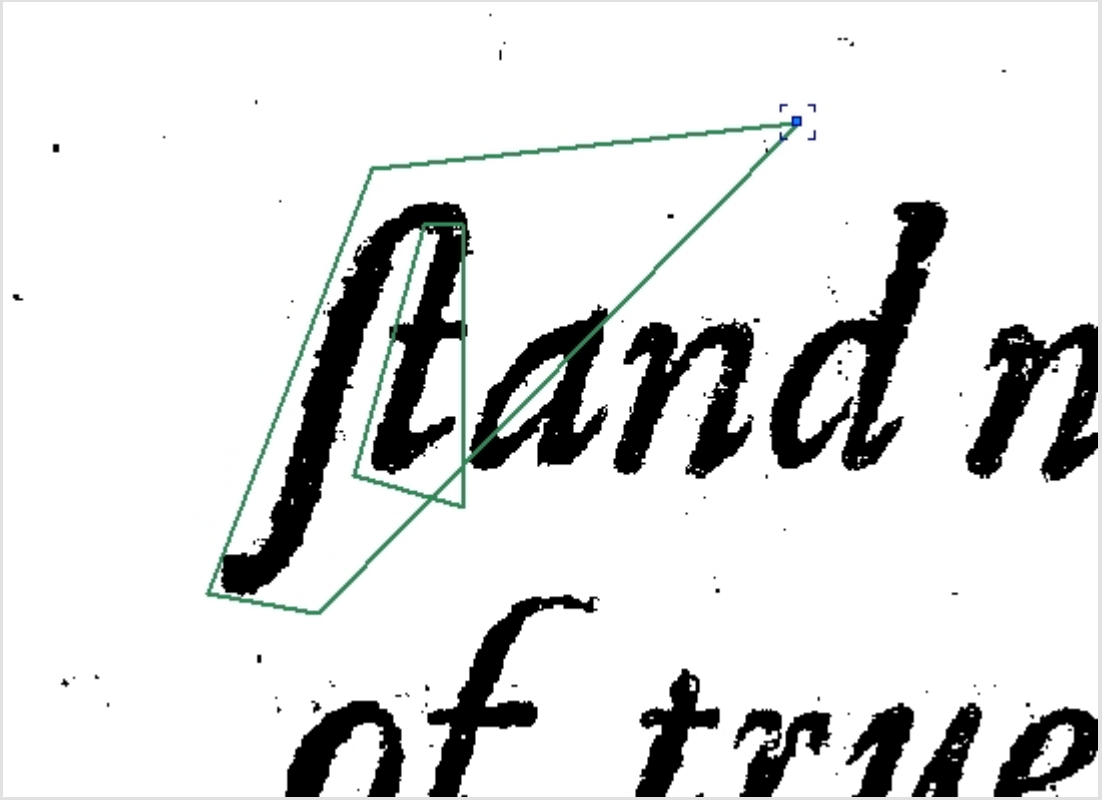
Misreads

- Ligature boundaries are sometimes confused



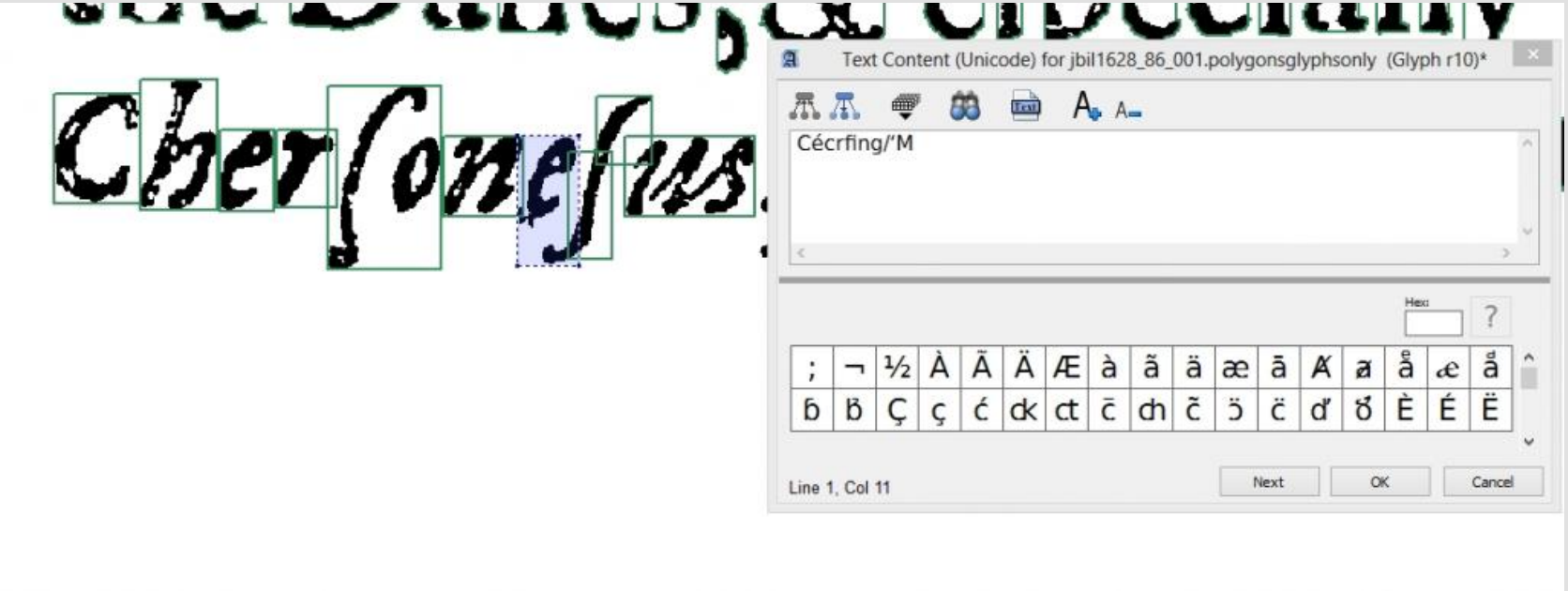
Step 5: Glyph Correction

- The parameters of a glyph may be corrected using the edit button to adjust lines



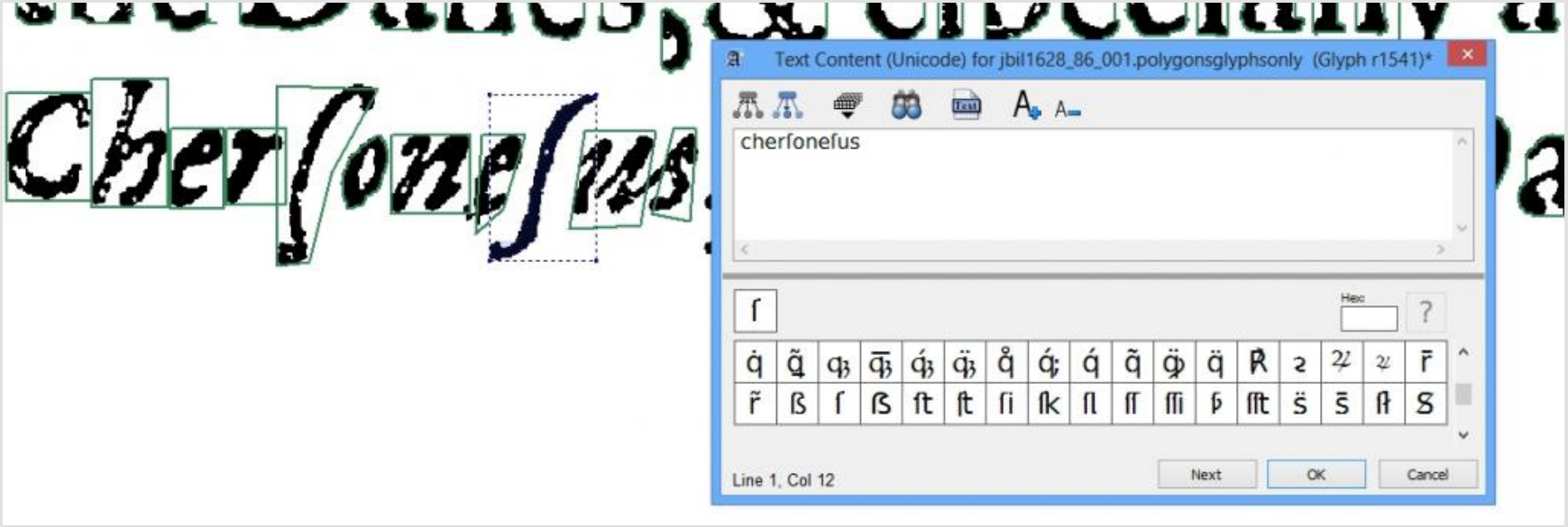
Step 6: Text Correction

- There is difficulty reading letters that are distorted, obscured by noise or ink bleed-through, or if they are especially faint.



Step 6: Text Correction

- Text misreads must be corrected by typing in the correct text in the Text Content Box for each corresponding character:



Font Sets

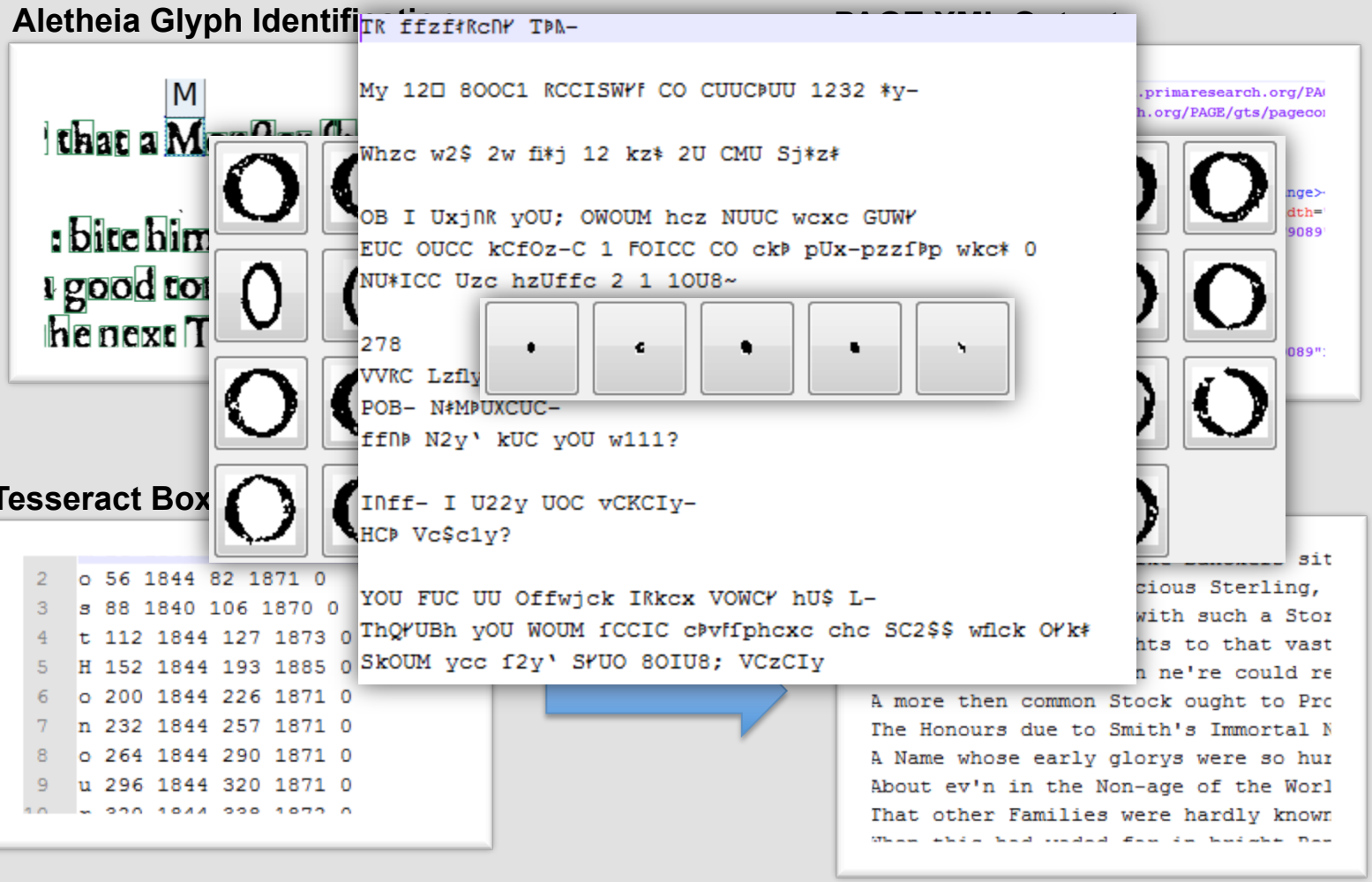
To date the Aletheia team at the IDHMC has produced fourteen font training sets

Font Name	Publisher	Publication Year	Images Processed
JDAY1559_001	John Daye	1559	15
MFLE1633_94_017	Miles Flesher	1633	11
TCOT1632_82_001	Thomas Cotes	1632	11
IJAGG_guyot_83_99	François Guyot	Unknown	13
BRA_2788_002	Pierre Haultin	Unknown	8
cbil1692_110_002	Charles Bill	1692	6
cbil1693_116_001	Charles Bill	1693	10
cbil1702_116_1_001	Charles Bill	1702	14
hhil1686_82_1_001	Henry Hills	1686	15
jbill1628_86_001	John Bill	1628	15
IJAGG_guyot_83_15_B italics	François Guyot	Unknown	12
BRA_2788_004_B italics	Pierre Haultin	Unknown	5
MFLE1633_94_017_B italics	Miles Flesher	1633	11
TCOT1632_82_001_B italics	Thomas Cotes	1632	14

Training Tesseract using Aletheia

- Once the font training libraries were in place, we started to train Tesseract by running each prepared font set through the Tesseract engine so that it would learn to read and recognize the characters of those fonts.

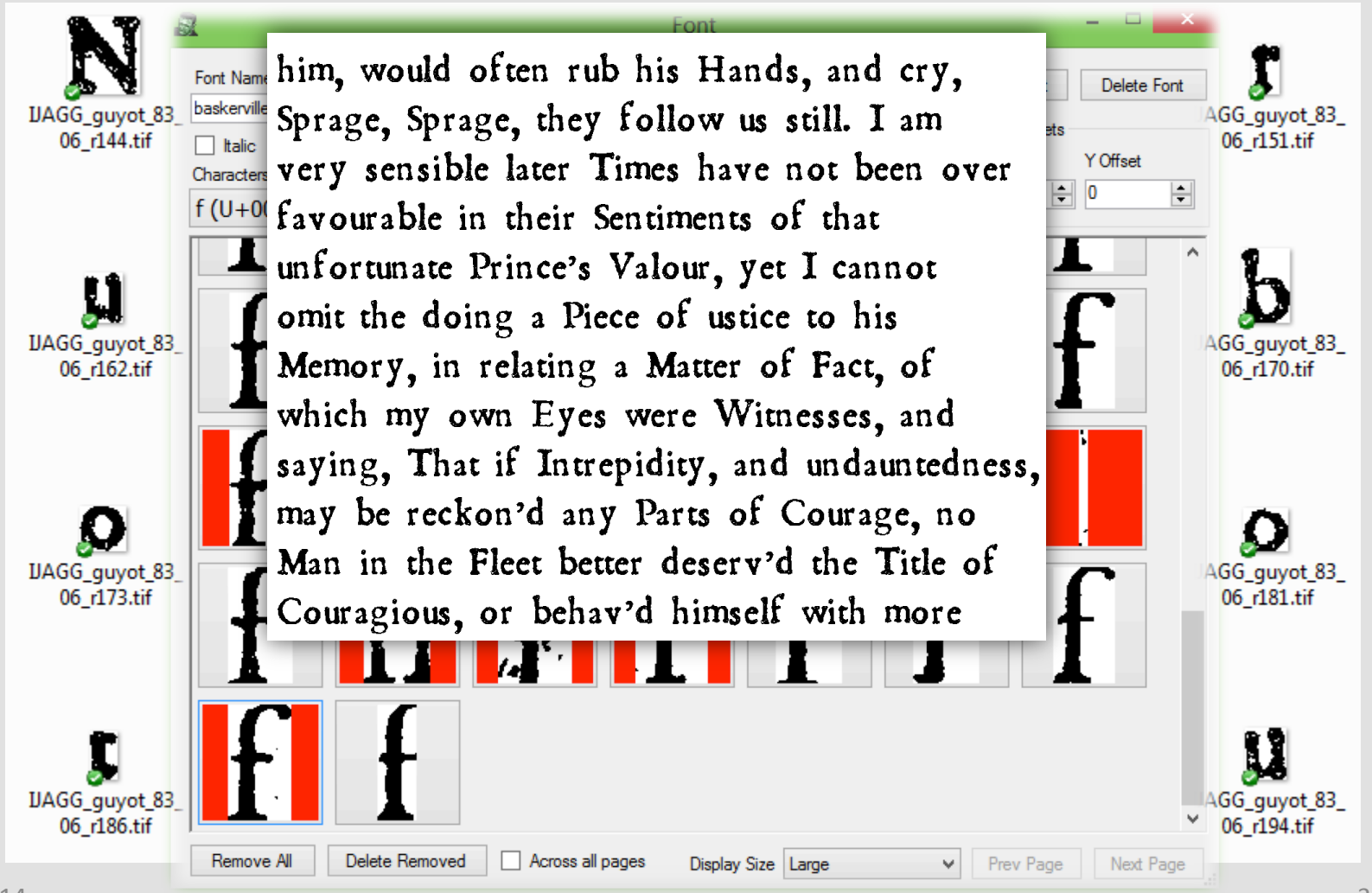
Aletheia and Tesseract



Roadblocks:

- How do we "feed" Tesseract only pristine instances of each letter for training purposes?
- How do we ensure that each glyph has been identified correctly and consistently?
- How do we eliminate the misidentification of pixel artifacts as characters?

The Solution: Franken+



Promising Results

TEXAS A&M UNIVERSITY

Before Franken+

```
OB I UxjnR yOU; OWOUM hcz NUUC wcxc GUWY
EUC OUCC kCfOz-C 1 FOICC CO ckÞ pUx-pzzfÞp wkc* 0
NU*ICC Uzc hzUffc 2 1 1OU8~

278
VVRC Lzfly TC hcc LOxch YOUÞIC *y 2
POB- N*MÞUXCUC-
ffnÞ N2y' kUC yOU w111?

Inff- I U22y UOC vCKCIy-
HCÞ Vc$c1y?

YOU FUC UU Offwjck IRkcx VOWCY hU$ L-
ThQYUBh yOU WOUM fCCIC cÞvffphcxc chc SC2$$ wflck OYk*
SkOUM ycc f2y' SYUO 8OIU8; VCzCIy
```

After Franken+

```
That when thou meet'af& one, with enqtziring eyes
Doi't fearch, and like a needy broker ptize

The Glke, and gold he weares, and to that rate
Sohigh or low, doG: raife thy formall hat:

That wilt confort none, Untill thou have knowne
What lands hee hath in hope, or of his owne,

As though all thy companions Ihould make thee
gointntes,and marry thy deare company.

Why HaouEd'R thou that dofk Uot onely apptove;
But in ranke itchie luf?, deGre, andlove

The nakednefe and barrenneHe to enioy,
```

Thank You!

- Kathy Torabi: torabik@tamu.edu
- Bryan Tarpley: bptarpley@tamu.edu
- The IDHMC: http://idhmc.tamu.edu
- The EMOP Project: http://emop.tamu.edu