Beyond the Early Modern OCR Project
or, What are We Going to do With Ourselves Now?
eMOP Info

**eMOP Website**
- [emop.tamu.edu/](emop.tamu.edu/)
- **TCDL 2015 – Beyond eMOP**
  - [http://emop.tamu.edu/presentations/#TCDL15](http://emop.tamu.edu/presentations/#TCDL15)
- eMOP Workflows
  - [emop.tamu.edu/workflows](emop.tamu.edu/workflows)
- Mellon Grant Proposal
  - [idhmc.tamu.edu/projects/Mellon/eMOPPublic.pdf](idhmc.tamu.edu/projects/Mellon/eMOPPublic.pdf)

**More eMOP**
- **Facebook**
  - The Early Modern OCR Project
- **Twitter**
  - #emop
  - @IDHMC_Nexus
  - @mandellc
  - @matt_christy
  - @EMGrumbach
The Early Modern OCR Project (eMOP) is an Andrew W. Mellon Foundation funded grant project running out of the Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University, to

- develop and test tools and techniques to apply Optical Character Recognition (OCR) to early modern English documents
- from the hand press period, roughly 1475-1800.

eMOP aims to improve the visibility of early modern texts by making their contents fully searchable. The current paradigm of searching special collections for early modern materials by either metadata alone or “dirty” OCR is insufficient for scholarly research.
The Numbers

Page Images

- Early English Books online (Proquest) EEBO: ~125,000 documents, ~13 million pages images (1475-1700)

- Eighteenth Century Collections Online (Gale Cengage) ECCO: ~182,000 documents, ~32 million page images (1700-1800)

- Total: >300,000 documents & 45 million page images.

Ground Truth

- Text Creation Partnership TCP: ~46,000 double-keyed hand transcribed documents
  - 44,000 EEBO
  - 2,200 ECCO
The Early Modern Data Set

EEBO

- 80,000 have No transcriptions
- 45,000 Typed by Hand
- 125,000 Total Documents

ECCO

- 179,800* have Gale’s OCR
- 182,000 Total Documents

Owned by ProQuest

1500 - Fonts Imported

1640 - Spelling regularizes Fonts still Imported

1700

1720 - Primarily Caslon and Caslon-like Fonts

1755 - Spelling Standardized

1790 -

Owned by Cengage Learning

1800
- **PRImA** (Pattern Recognition & Image Analysis Research) Lab at the University of Salford, Manchester, UK
- **SEASR** (Software Environment for the Advancement of Scholarly Research) at the University of Illinois, Urbana-Champaign
- **PSI** (Perception, Sensing, and Instrumentation) Lab at Texas A&M University
- **The Academy** for Advanced Telecommunications and Learning Technologies at Texas A&M University
  - The **Brazos** High Performance Computing Cluster (HPCC)
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
- over/under-inking, bleedthrough
- Old, low-quality, small tiff files
- Noise, skew, warp,
Results

- **ECCO**
  - Avg. Scores
    - 309,328 pages
    - 86% correct
    - 57% correct on 1st pages
  - Comparison to Prime Recognition’s OCR
    - 93% claim
    - (still running analysis)

- **EEBO**
  - Avg. Scores
    - 1,475,026 pages
    - 68% correct
  - No previous OCR to compare to

Using Google’s Tesseract open-source OCR engine with eMOP created training
eMOP Outcomes - Github

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach
Outcomes – Github Franken+

1. Windows based tool that uses a MySQL DB.
2. Developed for eMOP by IDHMC Graduate student worker Bryan Tarpley.
3. Designed to be easily used by eMOP Undergraduate student workers.
4. Takes Aletheia's output files as input.
5. Outputs the same box files and TIFF images that Tesseract's first stage of native training.

Early-MODERN-OCR/GITHUB.IO/FRANKENPLUS/
Outcomes – Github emop-dashboard

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach

[early-modern-ocr.github.io/emop-dashboard/]
Outcomes – Github hOCR deNoising

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach

[Link to Github repository: early-modern-ocr.github.io/hOCR-De-Noising/]
Outcomes – Github hOCR deNoising

- Code Repo
  - Tesseract Training
  - ImprintDB
  - TCP EEBO Phase 1
  - EEBO OCR
- Collection Evaluation
- Outreach
Outcomes – Github page evaluator

- Code Repo
  - Tesseract Training
  - ImprintDB
  - TCP EEBO Phase 1
  - EEBO OCR
  - Collection Evaluation
  - Outreach

3. Check token profile:

[link]

Determine how correctable a page’s OCR results are by examining the text. The score is based on the ratio of words that fit the correctable profile to the total number of words.

Correctable Profile:
1. Clean tokens:
   - remove leading and trailing punctuation
   - remaining token must have at least 3 letters
2. Spell check tokens > 1 character
3. Check token profile:
   - contain at most 2 non-alphabetic characters, and
   - at least 1 alphabetic character, and
   - have a length of at least 3, and
   - do not contain 4 or more repeated characters in a run
4. Also consider length of tokens compared to average for the page
Outcomes – Github page corrector

1. Preliminary cleanup
   - remove punctuation from begin/end of tokens & empty lines
   - combine hyphenated tokens at end of lines
   - retain cleaned & original tokens as “suggestions”

2. Apply common transformations and period specific dictionary lookups to gather suggestions for words.
   - transformation rules: rn->m; c->e; 1->l; e

3. Use context checking on a sliding window of 3 words, and their suggested changes, to find the best context matches in our(sanitized, period-specific) Google 3-gram dataset

Sir, he was silent so long, that I thought she was not going to answer at all; but at that I thought he Was
Outcomes – Github page corrector

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach

Sir, she was silent so long, that I thought she was not going to answer at all; but at that I thought she was

ContextMatch: though ice was (matchCount: 121, volCount: 120)
ContextMatch: though ike was (matchCount: 65, volCount: 59)
ContextMatch: though she was (matchCount: 556,763, volCount: 364,965)
ContextMatch: though the was (matchCount: 197, volCount: 196)
ContextMatch: thought ice was (matchCount: 45, volCount: 45)
ContextMatch: thought ike was (matchCount: 112, volCount: 108)
ContextMatch: thought she was (matchCount: 549,531, volCount: 325,822)
ContextMatch: thought the was (matchCount: 91, volCount: 91)
Outcomes – Github juxta-CL

- Juxta-CL(command line)
  - created for eMOP
  - based on JuxtaCommons tool ([juxtacommons.org/](https://juxtacommons.org/))
  - several different comparison algorithms to choose from and other options
    - Levenshtein / Jaro-Winkler / Juxta
    - ignore: punctuation, caps, hyphens

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach

• Powered by the eMOP DB
• Collection processing is managed via the online Dashboard [emop-dashboard.tamu.edu](http://emop-dashboard.tamu.edu)
• Run by emop-controller.py

**Code Repo**

• Tesseract Training
• ImprintDB
• TCP EEBO Phase 1
• EEBO OCR
• Collection Evaluation
• Outreach
Outcomes - Tesseract Training

- Font Training
  - 3 different types
    - Roman, Italic, Blackletter
  - 12 different printers
  - 1564-1764
  - 40 different typeface combinations
  - Even more combined typefaces training files

- Dictionaries
  - from multiple sources with alternate spellings

- More
  - Franken+ training files
  - common OCR error file

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach

early-modern-ocr.github.io/TesseractTraining/
Outcomes – DB of EM Printers

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach

- Parsing the imprint lines of all EEBO & ECCO docs to create the ImprintDB
- Gathering:
  - Printed By
  - Printed For
  - Sold by
  - Location ("at the Rose and Crown in St. Paul's Church-Yard", "at the signe of the Traytors head")
  - Place (London, Cambridge...)
  - Dates
- Those docs with ESTC numbers will be available via a public database
Outcomes – Fulltext Search of TCP Phase 1

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach

Phase 1 of TCP EEBO hand transcriptions will be available for fulltext searching in 18thConnect over 25,000 documents
Outcomes – Editable EEBO OCR

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach

Can already use TypeWright in 18thConnect to edit ECCO docs (without a subscription)

Will soon be able to edit EEBO docs in TypeWright

TCP hand transcriptions
eMOP OCR transcriptions

When doc is fully corrected, users can request text and/or XML versions of corrected transcriptions for scholarly use.

First time EEBO transcriptions will be available for over 80,000 docs.
## Outcomes – Collection Evaluation

<table>
<thead>
<tr>
<th>Page Number</th>
<th>Juxta</th>
<th>RETAS</th>
<th>SEASER corr</th>
<th>Est. Pg. Quality</th>
<th>Correction Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.692657</td>
<td>-</td>
<td>0.666667</td>
<td>0.342105</td>
<td>47</td>
</tr>
<tr>
<td>2</td>
<td>0.848329</td>
<td>-</td>
<td>0.92126</td>
<td>0.61165</td>
<td>183</td>
</tr>
<tr>
<td>3</td>
<td>0.869606</td>
<td>-</td>
<td>0.932773</td>
<td>0.567302</td>
<td>152</td>
</tr>
<tr>
<td>4</td>
<td>0.850645</td>
<td>-</td>
<td>0.979522</td>
<td>0.696602</td>
<td>372</td>
</tr>
<tr>
<td>5</td>
<td>0.925557</td>
<td>-</td>
<td>0.960916</td>
<td>0.706044</td>
<td>345</td>
</tr>
<tr>
<td>6</td>
<td>0.916622</td>
<td>-</td>
<td>0.978947</td>
<td>0.719072</td>
<td>379</td>
</tr>
<tr>
<td>7</td>
<td>0.0</td>
<td>-</td>
<td>0.985965</td>
<td>0.741425</td>
<td>374</td>
</tr>
<tr>
<td>8</td>
<td>0.911385</td>
<td>-</td>
<td>0.969388</td>
<td>0.744125</td>
<td>351</td>
</tr>
<tr>
<td>9</td>
<td>0.386674</td>
<td>-</td>
<td>0.967857</td>
<td>0.63515</td>
<td>392</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
<td>-</td>
<td>0.900415</td>
<td>0.627168</td>
<td>312</td>
</tr>
<tr>
<td>11</td>
<td>0.837313</td>
<td>-</td>
<td>0.966418</td>
<td>0.703804</td>
<td>339</td>
</tr>
<tr>
<td>12</td>
<td>0.920927</td>
<td>-</td>
<td>0.969811</td>
<td>0.694595</td>
<td>345</td>
</tr>
<tr>
<td>13</td>
<td>0.936136</td>
<td>-</td>
<td>0.968971</td>
<td>0.743094</td>
<td>355</td>
</tr>
<tr>
<td>14</td>
<td>0.915051</td>
<td>-</td>
<td>0.9631</td>
<td>0.696</td>
<td>340</td>
</tr>
<tr>
<td>15</td>
<td>0.92058</td>
<td>-</td>
<td>0.964962</td>
<td>0.757225</td>
<td>340</td>
</tr>
<tr>
<td>16</td>
<td>0.902174</td>
<td>-</td>
<td>0.967972</td>
<td>0.721485</td>
<td>351</td>
</tr>
<tr>
<td>17</td>
<td>0.89213</td>
<td>-</td>
<td>0.960317</td>
<td>0.720238</td>
<td>312</td>
</tr>
<tr>
<td>18</td>
<td>0.814873</td>
<td>-</td>
<td>0.904255</td>
<td>0.643939</td>
<td>118</td>
</tr>
</tbody>
</table>

- **Code Repo**
- **Tesseract Training**
- **ImprintDB**
- **TCP EEBO Phase 1**
- **EEBO OCR**
- **Collection Evaluation**
- **Outreach**

- We've collected over 6 TB of data
- Our post-processing algorithms collect data:
  - amount of skew
  - amount of noise
  - number and location of text columns
  - page quality
  - correctability
  - corrections made

- Saved to the eMOP DB
- First time this has been done on these collections
- Pre-process and re-OCR pages
- Several iterations will identify bad page images
Outcomes – Outreach

- 2013
  - ASECS (Amer. Soc. of Eighteenth Century Scholars)
  - KIAS (Kule Institute for Advanced Studies): Around the World Symposium
  - DocEng (ACM Symposium on Document Engineering)
- 2014
  - TxDHC
  - TCDL
  - DH (5 papers & a poster)
  - SAA (pre-conference workshop on OCR)
  - DHCS
- 2015
  - AAAI: (Association for the Advancement of Artificial Intelligence)
  - TAMU Big Data Workshop
  - Downstream from the Digital Humanities
  - Penn St. OCR Workshop
  - IDHMC CE Classes
  - TxDHC
  - TCDL
  - DHSI (July)

- Code Repo
- Tesseract Training
- ImprintDB
- TCP EEBO Phase 1
- EEBO OCR
- Collection Evaluation
- Outreach
The Future of eMOP

- We want eMOP to be viable long-term
- Continue to be developed/supported
  - open-source repos
  - outreach
- Looking for partners

from Kill Bill 2, 2004, Miramax
Future - Partners

- **Hathi Trust**
  - Talking about possible “next-step” grant application

- **Notre Dame Libraries**
  - Helping to recreate eMOP workflow on their systems to OCR *Cobbett’s Complete Collection of State Trials*
  - Swapping labor
  - Opportunity to test the robustness of our workflow and tools as a whole unit

- **Penn St. Libraries**
  - Held a workshop on OCR’ing with open source tools
  - Opportunity to test the robustness of our workflow and tools as discrete parts
Future - Partners

- **UT Austin**
  - Sub-awardee on an NEH grant to OCR Primeros Libros collection [primeroslibros.org/]
  - $$ for continued development of workflow and improved hardware
  - Opportunity to add another OCR engine and further test the robustness of the workflow on other document sets
    - step towards a voting algorithm

- **Adam Matthew Digital**
  - Going to OCR some of their EM collection to see how we do; maybe more if good results
  - Opportunity to test robustness on similar collection; establish a relationship with another industry leader
    - possibly acquire more data to ingest into ARC nodes and TypeWright
Future - Partners

- **Austin Fanzine Project**
  - Small, local project; personally interesting
  - Opportunity to test tools and workflow on a different collection and whole other type of documents (not that dissimilar)

- **Hathi Trust Research Center**
  - Talking about possibility of using our algorithms on their imprint data to expand and share Imprint DB
  - Opportunity to acquire more EM publishing data and make available in one place
Call to Libraries

- **Use the Code**
  - Give us feedback

- **Develop the Code**
  - Create branches in Git
  - Commit improvements/changes

- **Contact Us**
  - Consultation
  - Partnership
  - for Labor or Funds
The end

For eMOP questions please contact us at:

mchristy@tamu.edu
egrumbac@tamu.edu
mandell@tamu.edu