eMOP Workflow

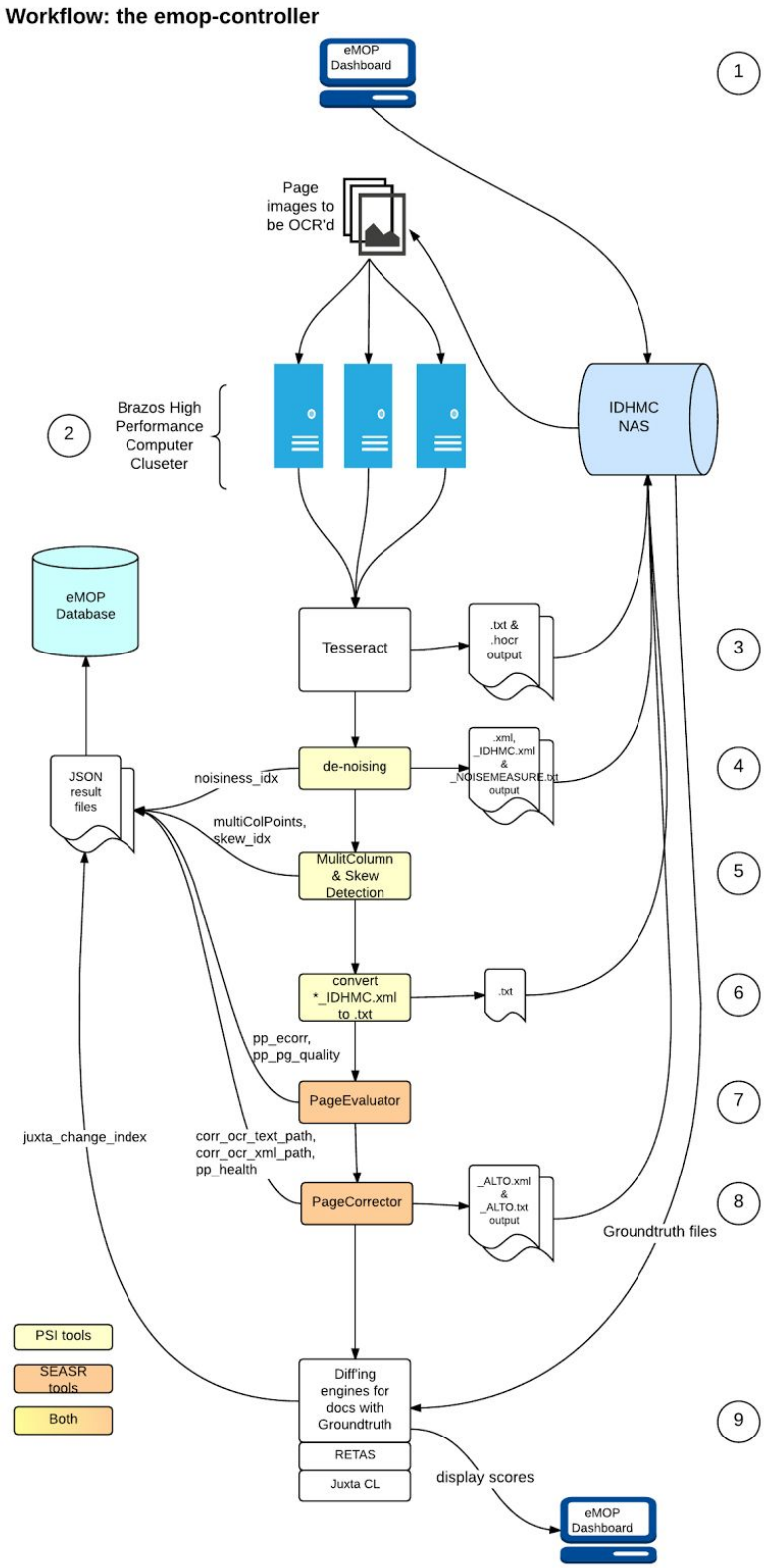
Design Description

1. eMOP Workflow

This section describes the current OCR process workflow at TAMU based on the work completed for the Early Modern OCR Project (eMOP)[1]. As seen in Figure 1, the eMOP workflow is a simple pipeline of various software components, which turns page images into their text and XML equivalents. The workflow is embodied in the emop-controller, which is a Python framework used to interact with the eMOP DB, a Network Access Storage (NAS) system, and software components written in several different languages.

[1] http://emop.tamu.edu/

Figure 1: The emop-controller, manager of the eMOP workflow



1.1. eMOP Controller

As pictured in Figure 1, the emop-controller works as follows:

① An authorized user uses the eMOP Dashboard web application to select documents from the collection to be OCRd. That opens a dialogue box which allows the user to select the OCR engine and, where applicable, which training set to use, while OCRing.
The Dashboard then marks the associated pages for the selected documents as “Not Started” in the eMOP DB’s JOB_QUEUE table.

② The **Dashboard** also servers as the point of contact between the **emop-controller** and the eMOP DB via an API. The emop-controller queries the Dashboard for information pertaining to all the selected documents’ pages (image file location, groundtruth info, current job status). The Dashboard returns information from the eMOP DB in the form of a JSON response, which the emop-controller writes as a set of input files to a temporary location on the NAS.
The **scheduler** splits pages into jobs of an equal number of pages for each available processor (128 dedicated processors for the IDHMC queue and a variable number of processors for the background queue) on the Brazos High Performance Computing Cluster (HPCC or Brazos). These jobs are then assigned to a processor queue for processing, where the emop-controller is called for each page.
Finally, all assigned pages have their status updated to “Processing” in a JSON formatted response file to be written to the eMOP DB when all pages assigned to a processor are finished.
For eMOP, parallelization was done on a page-level basis. Each available processor was utilized to run the emop-controller on a single page at a time to completion (when possible) on the 128 processors available to us at all times on the IDHMC queue, and many more on unused processors of background queues.

③ The TIF page images are OCRd with **Tesseract** using the training specified by the user in the Dashboard at job submission. Text and hOCR files are produced and saved on the NAS in with a path like <eMOPhome>/<batch#>/<emop#>/<files>.
hOCR is Tesseract’s proprietary XML-like output. It is actually HTML with extra “attributes” added as semi-colon separated values in the @title attribute of associated block tags. Tesseract partitions the hOCR output

into nested page, area, paragraph, line and work blocks. Each block contains bounding box (bbox) coordinates in it's @title attribute.[2]
The job status for each page is then updated to "Pending Postprocessing" in the JSON response file.
Tesseract is written in C++.
The current released version of Tesseract is 3.02. eMOP is using version 3.03 in order to take advantage of that version's ability to add confidence scores to each word in the hOCR output.
eMOP's hOCR files are renamed to have an .xml extension.
For eMOP all training includes a unicharambigs file which is used to convert special characters (ſ,) and ligatures (ft, œ) to their modern or multi-character equivalents (s, r, st, & oe respectively). We made this decision to improve searchability of the texts.

④ The eMOP **de-noising** algorithm analyzes the hOCR output in order to attempt to remove "noise words"—page noise and images that Tesseract identifies as words. The algorithm looks at **bbox coordinates** to identify words whose position and size indicate they are not part of the page's text block.
The de-noising algorithm is run on every page that is OCRd and takes the page's hOCR files as input. It produces one new file and updates another. The new output file is an xml file with a "_IDHMC" suffix added to the page number used as the filename. Both of these files have had all "noise words" removed. The updated file is the original input hOCR file (with an .xml extension) with additional values added to the @title string:

- pred: a value of 0 or 1 indicating that the word is likely valid or noise, respectively—based on the noiseConf value and a defualut cuttoff of 50%.
- noiseConf: measure of confidence of noise, the current default causes any word with a noiseConf value greater than 50% to be removed from the *_IDHMC produced files.[3]

The new file is written to the NAS in the same folder with the Tesseract output for that document.
The pages job status is updated to "Postprocessing" in the JSON response file. The de-noising algorithm also produces an overall measure of noise for the page, which is written to the JSON response file.
The de-noising algorithm, created by the PSI Lab at TAMU, is written in Python and requires the beautifulsoup, numpy, and scipy modules.

[2] See Appendix A.1 for a sample hOCR file.
[3] See Appendix A.2 for a sample de-noised hOCR file.

⑤ The multiple column and skew detection (**MultiColSkew**) algorithm utilizes the pages bounding boxes to analyze its geometry and identify when multiple columns are present in a page image, and their locations. It also identifies and measures the amount of skew present. These values are then written to each pages' JSON response file.
MultiColSkew, created by the PSI lab, is written in Python and requires the numpy module.

⑥ A final algorithm creates a new text version of the output with the "noise words" removed—i.e. based on the newly created *_IDHMC.txt files. That file is also appended with an "_IDHMC" suffix.
The new file is written to the NAS in the same folder with the Tesseract output for that document.

⑦ The **page-evaluator** is the first step in the page correction process. It evaluates the **text** produced by Tesseract (after de-noising) for each page to determine whether it fits the profile of expected from a "normal" page of text. After some cleaning of the text (removing leading punctuation) It looks at things like

- the number of words (tokens) on the page,
- the average length of those words,
- the occurrence of a continuous string of repeated characters,
- the length of each word compared to the page average,
- the interspersion of alphabetic and numeric characters, and punctuation in a word, and
- how many words can be found in a dictionary.

The page-evaluator creates a score for Estimated Correctability (ECORR) and Estimated Page Quality (the ECORR divided by the number of words on the page). These values are then added to each pages' JSON response.
The page-evaluator, created by SEASR at the University of Illinois, was written in scala and then converted to java for the the eMOP controller.

⑧ Pages are then passed to the **page-corrector** to undergo correction based on early modern (EM) dictionaries and an DB of google 3-grams collected from EM documents. The dictionaries include alternate and abbreviated spellings with special characters and ligatures converted to modern, multi-character equivalents. There are multiple English language dictionaries and a French and Latin dictionary as well.
The page-corrector takes as input a de-noised hOCR file containing the de-noising confidence measures.
In short, the page-corrector:

1. Starts with the first three words on the page,

2. looks up each word in the dictionaries for a match,
3. makes character substitutions for each word looking for other possible dictionary matches, then
4. uses all possible matches for each word to look for matches in the google 3-gram DB. Matching 3-grams are weighted based on the number of uses in the original texts. Words that matched in the dictionary without substitution are given more weight. All of this is used to determine the correct "matching" 3-gram.
5. The corrector then gets the next 3-word window, consisting of two words from the previous window and the next word in reading order.
6. Repeat from 2 till done.

When the page-corrector is complete for each page, it creates an ALTO XML and a text file containing all corrections. The ALTO XML[4] file also contains word confidence measures in a @WC attribute. Any word that is changed by the corrector contains one or more <ALTERNATIVE> sub-tags, the last of which is the original version of the word from the hOCR input.

The page-corrector, created by SEASR at the University of Illinois, was written in scala and then converted to java for the the eMOP controller.

⑨ Pages that have groundtruth equivalents available are then scored using **Juxta-CL**, using one of three character distance measurement algorithms (Levenshtein, Jaro-Winkler, and Juxta). The juxta score is then written to the JSON response file for each page. Each page's job status is updated to "Done".

The JSON input file from ② above contains a flag about whether groundtruth is available for each page as well the file path information for any groundtruth files.

Juxta-CL, created by Performant Software and based on JuxtaCommons, is written in java.

When a processor is finished with every page in it's job queue, the Dashboard writes all associated JSON response files to the eMOP DB. Document and page level results are viewable via the Dashboard.

1.2. Inputs/Outputs

At the lowest level of abstraction, the input for the overall eMOP workflow is page images. In the case of eMOP all of our input documents were low quality, small

[4] See Appendix A.3 for a sample ALTO XML file.

(avg ~ 40KB for ECCO and ~140KB for EEBO) TIF files. Every document was broken up into individual page TIFs (one page per image for ECCO and 2 pages per image for EEBO) by the provider. Tesseract is capable of handling a single TIF document with multiple pages (there are a handful of those in our collection as well).
A config.ini file residing in the eMOP home directory on Brazos is used to control the workflow by passing parameters to each component of the workflow.

```
$EMOP_HOME = /home/mchristy/emop/emop-controller-test
```

eMOP Dashboard [git]
Input: User selection of documents to be OCRd and training to be used.
Output: A set of temporary JSON files containing information obtained from a query of the eMOP DB for each page associated with the user-selected documents.
Location: `$EMOP_HOME/payload/input`
Configuration:

```
[dashboard]
api_version = 1
url_base = http://emop-dashboard.tamu.edu
```

Requirements: Ruby on Rails, Juxta web service

cluster scheduler [git]
Input: JobID(s)
Output: None
Configuration:

```
[scheduler]
max_jobs = 128
queue = idhmc
name = emop-controller
min_job_runtime = 300
max_job_runtime = 259200
avg_page_runtime = 480
logdir = /fdata/scratch/mchristy/emop-controller/logs
mem_per_cpu = 4000
cpus_per_task = 1
set_walltime = False
extra_args = ["--constraint", "core32"]
```

Requirements: Slurm (other schedulers are possible)
Dashboard interaction: None.

emop-controller [git]]
Input: TIF pages images.

Output: An hOCR (renamed with .xml extension) and a text file for each page, written to the IDHMC NAS. The filename is the page number.
Location: `/data/shared/text-xml/IDHMC-ocr/<batch#>/<emop#>/`
Configuration:

```
[controller]
payload_input_path =
/fdata/scratch/mchristy/emop-controller/payload/input
payload_output_path =
/fdata/scratch/mchristy/emop-controller/payload/output
ocr_root = /data/shared/text-xml/IDHMC-ocr
input_path_prefix = /dh
output_path_prefix = /dh
log_level = INFO
scheduler = slurm
skip_existing = True
```

Requirements: All of the following code packages.
Dashboard interaction: File paths of output files for each page, written to ocr_text_path and ocr_xml_path fields of page_results table. Job status for each page, written to job_status field of job_queue table.

Tesseract (v3.03) [git]
Input: Page images, training, DAWG files (dictionaries), unicharabmigs.
Output: hOCR / text files.
Configuration: None
Requirements: Leptonica
Dashboard interaction: None

De-nosing [git]
Input: hOCR files (<page#>.xml) with word-confidence levels included in the x_wconf field of the @title attribute.
Output: The original hOCR file (<page#>.xml) is updated to include a noiseConf measure for each word, and a pred field to indicate that the word falls above (1) or below (0) the default of 50%. A new XML file (<page#>_IDHMC.xml) is created that has all words with a pred value of 1 removed. A new text file (<page#>_IDHMC.txt is created from the associated xml file.
Location: `/data/shared/text-xml/IDHMC-ocr/<batch#>/<emop#>/`
Configuration: None
Requirements: beautifulsoup4, numpy, scipy
Dashboard interaction: An overall noise measure for the page, written to noisiness_idx field of page_results table.

MultiColumnSkew [git]
Input: De-noised <page#>_IDHMC.xml page file.
Output: None

Configuration:

```
[multi-column-skew]
enabled = True
```

Requirements: numpy

Dashboard interaction: Column coordinates and skew measure, written to multiColPoints and skew_idx fields of postproc_pages table.

page-evaluator [git]

Input: The original hOCR file (<page#>.xml) with updated noiseConf and pred fields added by de-noising algorithm.

Output: None.

Configuration:

```
[page-evaluator]
java_args = ["-Xms128M", "-Xmx128M"]
```

Requirements:

Dashboard interaction: Estimated correctability and page quality scores, written to the pp_ecorr and pp_pg_quality fields of postproc_pages table.

page-corrector [git]

Input: The original hOCR file (<page#>.xml) with updated noiseConf and pred fields added by de-noising algorithm.

Output: Creates two new files on the NAS: <page#>_ALTO.xml and <page#>_ALTO.txt. All words identified by the de-noiser with pred=0 are removed. The noiseConf value from the input XML is transferred to the ALTO XML output as the @emop:DNC attribute. A new @WC attribute is added to record the page-corrector's confidence that its contents are correct (0-100 value). One or more <ALTERNATIVE> sub-tags are included for every word which is corrected. The last <ALTERNATIVE> tag is the original word form from the input XML file.

If "save = True" in the config.ini, then a statistics file is created for each page and save to the NAS.

Location: `/data/shared/text-xml/IDHMC-ocr/<batch#>/<emop#>/`

Configuration:

```
[page-corrector]
java_args = ["-Xms2048M", "-Xmx2048M"]
alt_arg = 2
max_transforms = 20
noise_cutoff = 0.5
ctx_min_match =
ctx_min_vol =
dump = False
save = False
timeout = 300
```

Requirements:

Dashboard interaction: File paths of output files for each page, written to corr_ocr_text_path and corr_ocr_xml_path fields of page_results table. In addition, a string of corrector statistics is written to the pp_health field of the postproc_pages table. The string is a ";" separated list containing numbers for:

- total words
- ignored words (no attempt to process)
- correct words (processed and determined to be correct)
- corrected words
- unchanged words (processed and determined to be incorrect, but no correction available)

juxta-cl [git]

Input: Corrected text file (<page#>_ALTO.txt) and its associated groundtruth page file (the availability of, and a file path to, any groundtruth files are contained in the JSON file created by the eMOP Dashboard and stored in `$EMOP_HOME/payload/input.`

Output: None

Configuration:

```
[juxta-cl]
jx_algorithm = levenshtein
```

Requirements:

Dashboard interaction: The character level distance between the two pages is stored as a score between 0 & 1, written to the juxta_change_index field of the page_results table.

1.3. System Configuration

1.3.1. Brazos Cluster

As a stakeholder in the Brazos cluster, the IDHMC has full-time, uninterruptable access to 128 processors via the idhmc queue. We also have access to unused processors as they are available, via the background queue. However, background queue jobs can be interrupted at any time by higher priority queues.

The Brazos login server includes access to the IDHMC NAS for IO file storage.

1.3.1.1. Configuration Files

Upon logging in to the Brazos login server, I cd into the emop-controller directory and then load all required modules along with the emop module, which loads all software needed by the

emop-controller. In this same directory are several configuration files:

- config.ini: contains parameters to control the flow of the emop-controller and it's various components.
- emop.properties: contains location and login info for the google 3-gram DB.
- emop.slrm: is used by the scheduler to create job queues and call the emop-controller.

1.3.2. eMOP DB

The eMOP DB is quite large and resides on a dedicated database server accessible via the Brazos cluster. To minimize the potential for DB access to become a bottleneck in the workflow, database reading and writing is handle by the eMOP Dashboard for blocks of pages. Upon submission of a batch, by a Dashboard user, the eMOP DB is queried for all relevant data on the submitted pages. The batch is then split up into several queues to be assigned to available processors. While a job is processing, all output data is written to a JSON file. When a job has completed, it corresponding JSON files are sent back to the Dashboard where they are processed and written to the eMOP DB in a block.
All interaction with the eMOP DB is via the eMOP Dashboard through 4 available subcommands:

- **query:** Reads from dashboard - informational only - does not impact OCR workflow.
- **submit:** Reserves pages from dashboard (read+write). The pages returned by dashboard are modified to reflect they have been reserved for processing on cluster. This subcommand writes the returned data to JSON file and submits that as a job to the cluster scheduler.
- **run:** Runs the eMOP workflow on a compute node using JSON data as input and writing JSON data as output. The individual pieces (Tesseract, denoise, etc) also write their own files.
- **upload:** Typically executed after the "run" subcommand completes. This sends the job's JSON data back to dashboard to update dashboard on final status of OCR page(s). Data can aso be uploaded on-demand via this subcommand.

1.3.3. NAS

The IDHMC NAS is accessible from several of the IDHMC's servers as well as from the Brazos cluster. It serves as the IDHMC's and eMOP's

primary network storage device. It contains 42TB of disk space, about 25TB of which are currently (11/9/15) free. The NAS contains all of eMOP's page images, groundtruth files, and the result files of the entire eMOP workflow.

1.4. Github

All of the above described code is available open source via an Apache v2.0 licence at the eMOP Github page: https://github.com/Early-Modern-OCR.

1.5. Dashboard

1.5.1. API

The eMOP Dashboard also has an admin interface that provides an API into the eMOP DB via Ruby on Rails: http://emop-dashboard.tamu.edu/admin/dashboard. Documentation for using the API is availble at http://emop-dashboard.tamu.edu/apidoc. This is available to authenticated users only.

Appendix A

A.1 Original hOCR Output (eMOP work_id 32, page1)

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html lang="en" xml:lang="en" xmlns="http://www.w3.org/1999/xhtml">
    <head>
        <title> </title>
        <meta content="text/html;charset=utf-8" http-equiv="Content-Type" />
        <meta content="tesseract 3.03" name="ocr-system" />
        <meta content="ocr_page ocr_carea ocr_par ocr_line ocrx_word" name="ocr-capabilities" />
    </head>
    <body>
        <div class="ocr_page" id="page_1"
            title='image "/dh/data/eebo/e0031/40133/00001.000.001.tif"; bbox 0 0 2179 1842; ppageno
0; noisiness 0.1386'>
            <div class="ocr_carea" id="block_1_1" title="bbox 1211 0 2179 310">
                <p class="ocr_par" dir="ltr" id="par_1" title="bbox 1211 22 2179 310">
                    <span class="ocr_line" id="line_1"
                        title="bbox 1560 22 2179 127; baseline -0.0016155089 -33"><span
                            class="ocrx_word" dir="ltr" id="word_1" lang="SC8b-R8-D2b"
                            title="bbox 1560 24 1625 94; x_wconf 79">A</span>
                    </span>
                    <span class="ocr_line" id="line_2"
                        title="bbox 1211 132 1947 215; baseline 0.014945652 -10"><span
                            class="ocrx_word" dir="ltr" id="word_3" lang="SC8b-R8-D2b"
                            title="bbox 1211 132 1471 209; x_wconf 73"><em>Serious</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_4" lang="SC8b-R8-D2b"
                            title="bbox 1510 132 1947 215; x_wconf 72"><em>Exho.rtation</em></span>
                    </span>
                    <span class="ocr_line" id="line_3"
                        title="bbox 1438 253 2109 310; baseline 0.0074515648 -3"><span
                            class="ocrx_word" dir="ltr" id="word_5" lang="SC8b-R8-D2b"
                            title="bbox 1438 253 1552 308; x_wconf 84"><em>T()</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_6" lang="SC8b-R8-D2b"
                            title="bbox 1599 260 1645 309; x_wconf 89">A</span>
                        <span class="ocrx_word" dir="ltr" id="word_7" lang="SC8b-R8-D2b"
                            title="bbox 1666 258 1720 310; x_wconf 92">N</span>
                    </span>
                </p>
            </div>
            <div class="ocr_carea" id="block_2_2" title="bbox 1225 334 2013 599">
                <p class="ocr_par" dir="ltr" id="par_2" title="bbox 1225 334 2013 548">
                    <span class="ocr_line" id="line_4"
                        title="bbox 1225 334 2013 548; baseline -0.0038071066 -60"><span
                            class="ocrx_word" dir="ltr" id="word_11" lang="SC8b-R8-D2b"
                            title="bbox 1225 334 1563 533; x_wconf 79"><em>floly</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_12" lang="SC8b-R8-D2b"
                            title="bbox 1635 343 1930 492; x_wconf 69"><em>Life.</em></span>
                    </span>
                </p>
            </div>
            <div class="ocr_carea" id="block_3_3" title="bbox 1142 550 2003 716">
                <p class="ocr_par" dir="ltr" id="par_3" title="bbox 1142 625 2003 716">
                    <span class="ocr_line" id="line_5"
                        title="bbox 1142 625 2003 716; baseline 0.023228804 -19"><span
                            class="ocrx_word" dir="ltr" id="word_14" lang="SC8b-R8-D2b"
```

```
                    title="bbox 1142 629 1222 698; x_wconf 21"><em>.A</em></span>
                <span class="ocrx_word" dir="ltr" id="word_15" lang="SC8b-R8-D2b"
                    title="bbox 1241 625 1394 698; x_wconf 83"><em>Plea</em></span>
                <span class="ocrx_word" dir="ltr" id="word_16" lang="SC8b-R8-D2b"
                    title="bbox 1440 628 1679 701; x_wconf 74"><em>forthe</em></span>
                <span class="ocrx_word" dir="ltr" id="word_17" lang="SC8b-R8-D2b"
                    title="bbox 1720 636 2003 716; x_wconf 80"><em>absolute</em></span>
            </span>
        </p>
    </div>
    <div class="ocr_carea" id="block_4_4" title="bbox 1200 724 1939 1019">
        <p class="ocr_par" dir="ltr" id="par_4" title="bbox 1200 724 1939 1019">
            <span class="ocr_line" id="line_6"
                title="bbox 1222 724 1918 807; baseline 0.018678161 -22"><span
                    class="ocrx_word" dir="ltr" id="word_18" lang="SC8b-R8-D2b"
                    title="bbox 1222 724 1469 807; x_wconf 55"><em>flaïc(:efficy</em></span>
                <span class="ocrx_word" dir="ltr" id="word_19" lang="SC8b-R8-D2b"
                    title="bbox 1518 728 1610 786; x_wconf 25"><em>dof</em></span>
                <span class="ocrx_word" dir="ltr" id="word_20" lang="SC8b-R8-D2b"
                    title="bbox 1661 736 1918 797; x_wconf 82"><em>Inhcrcnt</em></span>
            </span>
            <span class="ocr_line" id="line_7"
                title="bbox 1277 807 1874 871; baseline 0.0050251256 -16"><span
                    class="ocrx_word" dir="ltr" id="word_21" lang="SC8b-R8-D2b"
                    title="bbox 1277 808 1643 871; x_wconf 40"
                    ><em>-Rjghteousness</em></span>
                <span class="ocrx_word" dir="ltr" id="word_22" lang="SC8b-R8-D2b"
                    title="bbox 1676 807 1722 859; x_wconf 67"><em>ilf</em></span>
                <span class="ocrx_word" dir="ltr" id="word_23" lang="SC8b-R8-D2b"
                    title="bbox 1748 817 1874 865; x_wconf 69">those</span>
            </span>
            <span class="ocr_line" id="line_8"
                title="bbox 1335 872 1823 926; baseline 0.020491803 -11"><span
                    class="ocrx_word" dir="ltr" id="word_24" lang="SC8b-R8-D2b"
                    title="bbox 1335 872 1426 917; x_wconf 87">that</span>
                <span class="ocrx_word" dir="ltr" id="word_25" lang="SC8b-R8-D2b"
                    title="bbox 1445 873 1554 926; x_wconf 77">hope</span>
                <span class="ocrx_word" dir="ltr" id="word_26" lang="SC8b-R8-D2b"
                    title="bbox 1570 887 1613 918; x_wconf 87">to</span>
                <span class="ocrx_word" dir="ltr" id="word_27" lang="SC8b-R8-D2b"
                    title="bbox 1631 873 1678 920; x_wconf 77"><em>be</em></span>
                <span class="ocrx_word" dir="ltr" id="word_28" lang="SC8b-R8-D2b"
                    title="bbox 1694 877 1823 924; x_wconf 78"><em>saved.</em></span>
            </span>
            <span class="ocr_line" id="line_9"
                title="bbox 1406 928 1454 958; baseline 0.083333333 -4"><span
                    class="ocrx_word" dir="ltr" id="word_29" lang="SC8b-R8-D2b"
                    title="bbox 1406 928 1454 958; x_wconf 61"><em>Z-</em></span>
            </span>
            <span class="ocr_line" id="line_10"
                title="bbox 1200 928 1939 1019; baseline 0.01082544 -19"><span
                    class="ocrx_word" dir="ltr" id="word_30" lang="SC8b-R8-D2b"
                    title="bbox 1200 958 1262 1014; x_wconf 89"><em>By</em></span>
                <span class="ocrx_word" dir="ltr" id="word_31" lang="SC8b-R8-D2b"
                    title="bbox 1287 928 1662 1019; x_wconf 48"
                    ><em>YT'o.JlWQdsworth-j</em></span>
                <span class="ocrx_word" dir="ltr" id="word_32" lang="SC8b-R8-D2b"
                    title="bbox 1683 963 1878 1009; x_wconf 74"><em>preacher</em></span>
                <span class="ocrx_word" dir="ltr" id="word_33" lang="SC8b-R8-D2b"
                    title="bbox 1893 979 1939 1011; x_wconf 89">to</span>
            </span>
        </p>
    </div>
    <div class="ocr_carea" id="block_5_5" title="bbox 1291 1013 1878 1128">
        <p class="ocr_par" dir="ltr" id="par_5" title="bbox 1291 1013 1878 1128">
```

```
                <span class="ocr_line" id="line_11"
                    title="bbox 1291 1013 1878 1074; baseline 0.0085178876 -19"><span
                        class="ocrx_word" dir="ltr" id="word_34" lang="SC8b-R8-D2b"
                        title="bbox 1291 1014 1359 1059; x_wconf 86"><em>che</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_35" lang="SC8b-R8-D2b"
                        title="bbox 1385 1013 1550 1057; x_wconf 38"><em>Glyn-reb</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_36" lang="SC8b-R8-D2b"
                        title="bbox 1565 1029 1604 1058; x_wconf 83">at</span>
                    <span class="ocrx_word" dir="ltr" id="word_37" lang="SC8b-R8-D2b"
                        title="bbox 1622 1018 1878 1074; x_wconf 67"><em>Newington-</em></span>
                </span>
                <span class="ocr_line" id="line_12"
                    title="bbox 1412 1071 1748 1128; baseline 0.020833333 -20"><span
                        class="ocrx_word" dir="ltr" id="word_38" lang="SC8b-R8-D2b"
                        title="bbox 1412 1073 1529 1110; x_wconf 81"><em>Butts</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_39" lang="SC8b-R8-D2b"
                        title="bbox 1542 1071 1582 1113; x_wconf 80"><em>in</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_40" lang="SC8b-R8-D2b"
                        title="bbox 1597 1076 1748 1128; x_wconf 73"><em>Surrey.</em></span>
                </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_6_6" title="bbox 1165 1153 1911 1170">
            <p class="ocr_par" dir="ltr" id="par_6" title="bbox 1165 1153 1911 1170">
                <span class="ocr_line" id="line_13"
                    title="bbox 1165 1153 1911 1170; baseline 0 672"> </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_7_7" title="bbox 1917 1166 2003 1176">
            <p class="ocr_par" dir="ltr" id="par_7" title="bbox 1917 1166 2003 1176">
                <span class="ocr_line" id="line_14"
                    title="bbox 1917 1166 2003 1176; baseline 0.023255814 -2"> </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_8_8" title="bbox 1155 1200 2003 1410">
            <p class="ocr_par" dir="ltr" id="par_8" title="bbox 1155 1200 2003 1410">
                <span class="ocr_line" id="line_15"
                    title="bbox 1432 1200 1705 1260; baseline -0.0036630037 -13"><span
                        class="ocrx_word" dir="ltr" id="word_43" lang="SC8b-R8-D2b"
                        title="bbox 1432 1200 1543 1247; x_wconf 81"><em>Heb.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_44" lang="SC8b-R8-D2b"
                        title="bbox 1562 1221 1577 1247; x_wconf 76">r</span>
                    <span class="ocrx_word" dir="ltr" id="word_45" lang="SC8b-R8-D2b"
                        title="bbox 1590 1218 1626 1245; x_wconf 88"><em>2.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_46" lang="SC8b-R8-D2b"
                        title="bbox 1640 1221 1654 1246; x_wconf 82">1</span>
                    <span class="ocrx_word" dir="ltr" id="word_47" lang="SC8b-R8-D2b"
                        title="bbox 1665 1224 1705 1260; x_wconf 65"><em>4.</em></span>
                </span>
                <span class="ocr_line" id="line_16"
                    title="bbox 1155 1259 2003 1322; baseline 0.014150943 -27"><span
                        class="ocrx_word" dir="ltr" id="word_48" lang="SC8b-R8-D2b"
                        title="bbox 1155 1259 1285 1298; x_wconf 74"><em>Follow</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_49" lang="SC8b-R8-D2b"
                        title="bbox 1293 1270 1400 1311; x_wconf 78"><em>peace</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_50" lang="SC8b-R8-D2b"
                        title="bbox 1413 1260 1510 1300; x_wconf 79"><em>with</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_51" lang="SC8b-R8-D2b"
                        title="bbox 1524 1278 1546 1299; x_wconf 77"><em>a</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_52" lang="SC8b-R8-D2b"
                        title="bbox 1562 1265 1576 1300; x_wconf 84"><em>l</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_53" lang="SC8b-R8-D2b"
                        title="bbox 1591 1277 1672 1301; x_wconf 77"><em>men</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_54" lang="SC8b-R8-D2b"
                        title="bbox 1685 1294 1698 1312; x_wconf 83">'</span>
```

```
                    <span class="ocrx_word" dir="ltr" id="word_55" lang="SC8b-R8-D2b"
                        title="bbox 1729 1264 1808 1305; x_wconf 75"><em>and</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_56" lang="SC8b-R8-D2b"
                        title="bbox 1833 1264 2003 1322; x_wconf 61"><em>holincss,</em></span>
                </span>
                <span class="ocr_line" id="line_17"
                    title="bbox 1213 1314 2002 1376; baseline 0.020278834 -26"><span
                        class="ocrx_word" dir="ltr" id="word_57" lang="SC8b-R8-D2b"
                        title="bbox 1213 1314 1376 1352; x_wconf 65"><em>vvithout</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_58" lang="SC8b-R8-D2b"
                        title="bbox 1405 1315 1533 1356; x_wconf 69"><em>which</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_59" lang="SC8b-R8-D2b"
                        title="bbox 1557 1333 1600 1359; x_wconf 78"><em>no</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_60" lang="SC8b-R8-D2b"
                        title="bbox 1621 1320 1823 1375; x_wconf 68"><em>m.-mstqall</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_61" lang="SC8b-R8-D2b"
                        title="bbox 1844 1324 1908 1376; x_wconf 70"><em>stc</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_62" lang="SC8b-R8-D2b"
                        title="bbox 1937 1325 2002 1366; x_wconf 72"><em>tht</em></span>
                </span>
                <span class="ocr_line" id="line_18"
                    title="bbox 1212 1370 1328 1410; baseline 0.0086206897 -2"><span
                        class="ocrx_word" dir="ltr" id="word_63" lang="SC8b-R8-D2b"
                        title="bbox 1212 1370 1328 1410; x_wconf 82"><em>Lord.</em></span>
                </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_9_9" title="bbox 1153 1460 2000 1466">
            <p class="ocr_par" dir="ltr" id="par_9" title="bbox 1153 1460 2000 1466">
                <span class="ocr_line" id="line_19"
                    title="bbox 1153 1460 2000 1466; baseline 0 376"> </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_10_10" title="bbox 1149 1486 2179 1716">
            <p class="ocr_par" dir="ltr" id="par_10" title="bbox 1149 1486 2179 1675">
                <span class="ocr_line" id="line_20"
                    title="bbox 1422 1486 2179 1566; baseline 0.0052840159 -24"><span
                        class="ocrx_word" dir="ltr" id="word_65" lang="SC8b-R8-D2b"
                        title="bbox 1422 1504 1460 1542; x_wconf 87"><em>L</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_66" lang="SC8b-R8-D2b"
                        title="bbox 1475 1509 1506 1546; x_wconf 89"><em>o</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_67" lang="SC8b-R8-D2b"
                        title="bbox 1518 1506 1568 1545; x_wconf 89"><em>N</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_68" lang="SC8b-R8-D2b"
                        title="bbox 1582 1507 1617 1544; x_wconf 89">D</span>
                    <span class="ocrx_word" dir="ltr" id="word_69" lang="SC8b-R8-D2b"
                        title="bbox 1633 1509 1665 1544; x_wconf 83"><em>o</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_70" lang="SC8b-R8-D2b"
                        title="bbox 1682 1510 1743 1556; x_wconf 87"><em>N'</em></span>
                </span>
                <span class="ocr_line" id="line_21"
                    title="bbox 1149 1554 2179 1615; baseline 0.0077669903 -15"><span
                        class="ocrx_word" dir="ltr" id="word_74" lang="SC8b-R8-D2b"
                        title="bbox 1149 1554 1315 1600; x_wconf 77"><em>Psinted</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_75" lang="SC8b-R8-D2b"
                        title="bbox 1330 1555 1384 1612; x_wconf 86"><em>by</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_76" lang="SC8b-R8-D2b"
                        title="bbox 1402 1564 1451 1601; x_wconf 81"><em>R-.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_77" lang="SC8b-R8-D2b"
                        title="bbox 1465 1563 1502 1601; x_wconf 79"><em>I.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_78" lang="SC8b-R8-D2b"
                        title="bbox 1525 1558 1590 1604; x_wconf 83"><em>for</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_79" lang="SC8b-R8-D2b"
                        title="bbox 1608 1564 1765 1603; x_wconf 73"><em>Andrew</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_80" lang="SC8b-R8-D2b"
```

```
                        title="bbox 1784 1563 1949 1614; x_wconf 75"><em>Kembc'</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_81" lang="SC8b-R8-D2b"
                        title="bbox 1962 1576 2003 1608; x_wconf 83"><em>at</em></span>
                </span>
                <span class="ocr_line" id="line_22"
                    title="bbox 1163 1614 2179 1675; baseline 0.018700787 -25"><span
                        class="ocrx_word" dir="ltr" id="word_83" lang="SC8b-R8-D2b"
                        title="bbox 1163 1614 1452 1672; x_wconf 74"
                        ><em>sr.Ma,-gare:s</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_84" lang="SC8b-R8-D2b"
                        title="bbox 1468 1615 1561 1660; x_wconf 79">Hill</span>
                    <span class="ocrx_word" dir="ltr" id="word_85" lang="SC8b-R8-D2b"
                        title="bbox 1577 1618 1617 1660; x_wconf 82"><em>iu</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_86" lang="SC8b-R8-D2b"
                        title="bbox 1635 1619 1987 1675; x_wconf 73"
                        ><em>Scm-hwark;And</em></span>
                </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_11_11" title="bbox 1195 1668 1953 1758">
            <p class="ocr_par" dir="ltr" id="par_11" title="bbox 1195 1668 1953 1758">
                <span class="ocr_line" id="line_23"
                    title="bbox 1195 1668 1953 1719; baseline 0.0092348285 -15"><span
                        class="ocrx_word" dir="ltr" id="word_88" lang="SC8b-R8-D2b"
                        title="bbox 1195 1682 1246 1705; x_wconf 79"><em>are</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_89" lang="SC8b-R8-D2b"
                        title="bbox 1258 1682 1293 1704; x_wconf 85">to</span>
                    <span class="ocrx_word" dir="ltr" id="word_90" lang="SC8b-R8-D2b"
                        title="bbox 1303 1668 1360 1704; x_wconf 83"><em>bee</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_91" lang="SC8b-R8-D2b"
                        title="bbox 1375 1668 1442 1705; x_wconf 81"><em>fold</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_92" lang="SC8b-R8-D2b"
                        title="bbox 1455 1673 1550 1707; x_wconf 87"><em>under</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_93" lang="SC8b-R8-D2b"
                        title="bbox 1565 1675 1610 1708; x_wconf 83"><em>St.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_94" lang="SC8b-R8-D2b"
                        title="bbox 1621 1675 1748 1719; x_wconf 64"><em>,M.:rga.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_95" lang="SC8b-R8-D2b"
                        title="bbox 1758 1683 1803 1709; x_wconf 71"><em>ers</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_96" lang="SC8b-R8-D2b"
                        title="bbox 1820 1676 1953 1714; x_wconf 70"><em>Church</em></span>
                </span>
                <span class="ocr_line" id="line_24"
                    title="bbox 1287 1718 1848 1758; baseline 0.016042781 -9"><span
                        class="ocrx_word" dir="ltr" id="word_97" lang="SC8b-R8-D2b"
                        title="bbox 1287 1727 1331 1750; x_wconf 72"><em>on</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_98" lang="SC8b-R8-D2b"
                        title="bbox 1346 1718 1616 1755; x_wconf 76"
                        ><em>New-Filhstreet</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_99" lang="SC8b-R8-D2b"
                        title="bbox 1628 1721 1711 1755; x_wconf 81">Hill.</span>
                    <span class="ocrx_word" dir="ltr" id="word_100" lang="SC8b-R8-D2b"
                        title="bbox 1744 1729 1848 1758; x_wconf 76"><em>166.-).</em></span>
                </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_12_12" title="bbox 0 0 2179 1842">
            <p class="ocr_par" dir="ltr" id="par_12" title="bbox 0 0 2179 1842">
                <span class="ocr_line" id="line_25" title="bbox 0 0 2179 1842; baseline 0 0"
                > </span>
            </p>
        </div>
    </div>
</body>
</html>
```

A.2 De-noised hOCR Output (eMOP work_id 32, page1)

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html lang="en" xml:lang="en" xmlns="http://www.w3.org/1999/xhtml">
    <head>
        <title> </title>
        <meta content="text/html;charset=utf-8" http-equiv="Content-Type" />
        <meta content="tesseract 3.03" name="ocr-system" />
        <meta content="ocr_page ocr_carea ocr_par ocr_line ocrx_word" name="ocr-capabilities" />
    </head>
    <body>
        <div class="ocr_page" id="page_1"
            title='image "/dh/data/eebo/e0031/40133/00001.000.001.tif"; bbox 0 0 2179 1842; ppageno
0; noisiness 0.1386'>
            <div class="ocr_carea" id="block_1_1" title="bbox 1211 0 2179 310">
                <p class="ocr_par" dir="ltr" id="par_1" title="bbox 1211 22 2179 310">
                    <span class="ocr_line" id="line_1"
                        title="bbox 1560 22 2179 127; baseline -0.0016155089 -33"><span
                            class="ocrx_word" dir="ltr" id="word_1" lang="SC8b-R8-D2b"
                            title="bbox 1560 24 1625 94; x_wconf 79; pred 1; noiseConf 0.0096"
                            >A</span>
                        <span class="ocrx_word" id="word_2" lang="SC8b-R8-D2b"
                            title="bbox 2121 22 2179 127; x_wconf 0; pred 0; noiseConf 0.9933"
                        ></span>
                    </span>
                    <span class="ocr_line" id="line_2"
                        title="bbox 1211 132 1947 215; baseline 0.014945652 -10"><span
                            class="ocrx_word" dir="ltr" id="word_3" lang="SC8b-R8-D2b"
                            title="bbox 1211 132 1471 209; x_wconf 73; pred 1; noiseConf 0.0048"
                                ><em>Serious</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_4" lang="SC8b-R8-D2b"
                            title="bbox 1510 132 1947 215; x_wconf 72; pred 1; noiseConf 0.0048"
                                ><em>Exho.rtation</em></span>
                    </span>
                    <span class="ocr_line" id="line_3"
                        title="bbox 1438 253 2109 310; baseline 0.0074515648 -3"><span
                            class="ocrx_word" dir="ltr" id="word_5" lang="SC8b-R8-D2b"
                            title="bbox 1438 253 1552 308; x_wconf 84; pred 1; noiseConf 0.0024"
                                ><em>T()</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_6" lang="SC8b-R8-D2b"
                            title="bbox 1599 260 1645 309; x_wconf 89; pred 1; noiseConf 0.0048"
                            >A</span>
                        <span class="ocrx_word" dir="ltr" id="word_7" lang="SC8b-R8-D2b"
                            title="bbox 1666 258 1720 310; x_wconf 92; pred 1; noiseConf 0.0083"
                            >N</span>
                        <span class="ocrx_word" dir="ltr" id="word_8" lang="SC8b-R8-D2b"
                            title="bbox 1834 302 1841 308; x_wconf 93; pred 0; noiseConf 0.9953"
                            >.</span>
                        <span class="ocrx_word" dir="ltr" id="word_9" lang="SC8b-R8-D2b"
                            title="bbox 1972 275 1983 282; x_wconf 69; pred 0; noiseConf 0.9839"
                            >'</span>
                        <span class="ocrx_word" dir="ltr" id="word_10" lang="SC8b-R8-D2b"
                            title="bbox 2106 291 2109 293; x_wconf 95; pred 0; noiseConf 0.9500"
                            >-</span>
                    </span>
```

```
    </p>
</div>
<div class="ocr_carea" id="block_2_2" title="bbox 1225 334 2013 599">
    <p class="ocr_par" dir="ltr" id="par_2" title="bbox 1225 334 2013 548">
        <span class="ocr_line" id="line_4"
            title="bbox 1225 334 2013 548; baseline -0.0038071066 -60"><span
                class="ocrx_word" dir="ltr" id="word_11" lang="SC8b-R8-D2b"
                title="bbox 1225 334 1563 533; x_wconf 79; pred 1; noiseConf 0.0242"
                    ><em>floly</em></span>
            <span class="ocrx_word" dir="ltr" id="word_12" lang="SC8b-R8-D2b"
                title="bbox 1635 343 1930 492; x_wconf 69; pred 1; noiseConf 0.0965"
                    ><em>Life.</em></span>
            <span class="ocrx_word" dir="ltr" id="word_13" lang="SC8b-R8-D2b"
                title="bbox 2008 397 2013 403; x_wconf 41; pred 0; noiseConf 0.9894"
                    ><em>-</em></span>
        </span>
    </p>
</div>
<div class="ocr_carea" id="block_3_3" title="bbox 1142 550 2003 716">
    <p class="ocr_par" dir="ltr" id="par_3" title="bbox 1142 625 2003 716">
        <span class="ocr_line" id="line_5"
            title="bbox 1142 625 2003 716; baseline 0.023228804 -19"><span
                class="ocrx_word" dir="ltr" id="word_14" lang="SC8b-R8-D2b"
                title="bbox 1142 629 1222 698; x_wconf 21; pred 1; noiseConf 0.0883"
                    ><em>.A</em></span>
            <span class="ocrx_word" dir="ltr" id="word_15" lang="SC8b-R8-D2b"
                title="bbox 1241 625 1394 698; x_wconf 83; pred 1; noiseConf 0.0041"
                    ><em>Plea</em></span>
            <span class="ocrx_word" dir="ltr" id="word_16" lang="SC8b-R8-D2b"
                title="bbox 1440 628 1679 701; x_wconf 74; pred 1; noiseConf 0.0030"
                    ><em>forthe</em></span>
            <span class="ocrx_word" dir="ltr" id="word_17" lang="SC8b-R8-D2b"
                title="bbox 1720 636 2003 716; x_wconf 80; pred 1; noiseConf 0.0028"
                    ><em>absolute</em></span>
        </span>
    </p>
</div>
<div class="ocr_carea" id="block_4_4" title="bbox 1200 724 1939 1019">
    <p class="ocr_par" dir="ltr" id="par_4" title="bbox 1200 724 1939 1019">
        <span class="ocr_line" id="line_6"
            title="bbox 1222 724 1918 807; baseline 0.018678161 -22"><span
                class="ocrx_word" dir="ltr" id="word_18" lang="SC8b-R8-D2b"
                title="bbox 1222 724 1469 807; x_wconf 55; pred 1; noiseConf 0.1122"
                    ><em>flaïc(:efficy</em></span>
            <span class="ocrx_word" dir="ltr" id="word_19" lang="SC8b-R8-D2b"
                title="bbox 1518 728 1610 786; x_wconf 25; pred 1; noiseConf 0.0346"
                    ><em>dof</em></span>
            <span class="ocrx_word" dir="ltr" id="word_20" lang="SC8b-R8-D2b"
                title="bbox 1661 736 1918 797; x_wconf 82; pred 1; noiseConf 0.0011"
                    ><em>Inhcrcnt</em></span>
        </span>
        <span class="ocr_line" id="line_7"
            title="bbox 1277 807 1874 871; baseline 0.0050251256 -16"><span
                class="ocrx_word" dir="ltr" id="word_21" lang="SC8b-R8-D2b"
                title="bbox 1277 808 1643 871; x_wconf 40; pred 1; noiseConf 0.0420"
                    ><em>-Rjghteousness</em></span>
            <span class="ocrx_word" dir="ltr" id="word_22" lang="SC8b-R8-D2b"
                title="bbox 1676 807 1722 859; x_wconf 67; pred 1; noiseConf 0.0112"
                    ><em>ilf</em></span>
            <span class="ocrx_word" dir="ltr" id="word_23" lang="SC8b-R8-D2b"
                title="bbox 1748 817 1874 865; x_wconf 69; pred 1; noiseConf 0.0036"
                >those</span>
        </span>
        <span class="ocr_line" id="line_8"
            title="bbox 1335 872 1823 926; baseline 0.020491803 -11"><span
```

```
                        class="ocrx_word" dir="ltr" id="word_24" lang="SC8b-R8-D2b"
                        title="bbox 1335 872 1426 917; x_wconf 87; pred 1; noiseConf 0.0021"
                        >that</span>
                    <span class="ocrx_word" dir="ltr" id="word_25" lang="SC8b-R8-D2b"
                        title="bbox 1445 873 1554 926; x_wconf 77; pred 1; noiseConf 0.0022"
                        >hope</span>
                    <span class="ocrx_word" dir="ltr" id="word_26" lang="SC8b-R8-D2b"
                        title="bbox 1570 887 1613 918; x_wconf 87; pred 1; noiseConf
0.0028"
                        >to</span>
                    <span class="ocrx_word" dir="ltr" id="word_27" lang="SC8b-R8-D2b"
                        title="bbox 1631 873 1678 920; x_wconf 77; pred 1; noiseConf
0.0039"
                            ><em>be</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_28" lang="SC8b-R8-D2b"
                        title="bbox 1694 877 1823 924; x_wconf 78; pred 1; noiseConf
0.0020"
                            ><em>saved.</em></span>
                </span>
                <span class="ocr_line" id="line_9"
                    title="bbox 1406 928 1454 958; baseline 0.083333333 -4"><span
                        class="ocrx_word" dir="ltr" id="word_29" lang="SC8b-R8-D2b"
                        title="bbox 1406 928 1454 958; x_wconf 61; pred 1; noiseConf
0.0121"
                            ><em>Z-</em></span>
                </span>
                <span class="ocr_line" id="line_10"
                    title="bbox 1200 928 1939 1019; baseline 0.01082544 -19"><span
                        class="ocrx_word" dir="ltr" id="word_30" lang="SC8b-R8-D2b"
                        title="bbox 1200 958 1262 1014; x_wconf 89; pred 1; noiseConf
0.0076"
                            ><em>By</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_31" lang="SC8b-R8-D2b"
                        title="bbox 1287 928 1662 1019; x_wconf 48; pred 1; noiseConf
0.4293"
                            ><em>YT'o.JlWQdsworth-j</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_32" lang="SC8b-R8-D2b"
                        title="bbox 1683 963 1878 1009; x_wconf 74; pred 1; noiseConf
0.0018"
                            ><em>preacher</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_33" lang="SC8b-R8-D2b"
                        title="bbox 1893 979 1939 1011; x_wconf 89; pred 1; noiseConf
0.0044"
                            >to</span>
                </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_5_5" title="bbox 1291 1013 1878 1128">
            <p class="ocr_par" dir="ltr" id="par_5" title="bbox 1291 1013 1878 1128">
                <span class="ocr_line" id="line_11"
                    title="bbox 1291 1013 1878 1074; baseline 0.0085178876 -19"><span
                        class="ocrx_word" dir="ltr" id="word_34" lang="SC8b-R8-D2b"
                        title="bbox 1291 1014 1359 1059; x_wconf 86; pred 1; noiseConf
0.0028"
                            ><em>che</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_35" lang="SC8b-R8-D2b"
                        title="bbox 1385 1013 1550 1057; x_wconf 38; pred 1; noiseConf
0.0395"
                            ><em>Glyn-reb</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_36" lang="SC8b-R8-D2b"
```

```
                            title="bbox 1565 1029 1604 1058; x_wconf 83; pred 1; noiseConf
0.0030"
                            >at</span>
                        <span class="ocrx_word" dir="ltr" id="word_37" lang="SC8b-R8-D2b"
                            title="bbox 1622 1018 1878 1074; x_wconf 67; pred 1; noiseConf
0.0030"
                                ><em>Newington-</em></span>
                    </span>
                    <span class="ocr_line" id="line_12"
                        title="bbox 1412 1071 1748 1128; baseline 0.020833333 -20"><span
                            class="ocrx_word" dir="ltr" id="word_38" lang="SC8b-R8-D2b"
                            title="bbox 1412 1073 1529 1110; x_wconf 81; pred 1; noiseConf
0.0017"
                                ><em>Butts</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_39" lang="SC8b-R8-D2b"
                            title="bbox 1542 1071 1582 1113; x_wconf 80; pred 1; noiseConf
0.0032"
                                ><em>in</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_40" lang="SC8b-R8-D2b"
                            title="bbox 1597 1076 1748 1128; x_wconf 73; pred 1; noiseConf 0.0022"
                                ><em>Surrey.</em></span>
                    </span>
                </p>
            </div>
            <div class="ocr_carea" id="block_6_6" title="bbox 1165 1153 1911 1170">
                <p class="ocr_par" dir="ltr" id="par_6" title="bbox 1165 1153 1911 1170">
                    <span class="ocr_line" id="line_13"
                        title="bbox 1165 1153 1911 1170; baseline 0 672"><span class="ocrx_word"
                            dir="ltr" id="word_41" lang="SC8b-R8-D2b"
                            title="bbox 1165 1153 1911 1170; x_wconf 95; pred 0; noiseConf 0.9500"
                                ><em> </em></span>
                    </span>
                </p>
            </div>
            <div class="ocr_carea" id="block_7_7" title="bbox 1917 1166 2003 1176">
                <p class="ocr_par" dir="ltr" id="par_7" title="bbox 1917 1166 2003 1176">
                    <span class="ocr_line" id="line_14"
                        title="bbox 1917 1166 2003 1176; baseline 0.023255814 -2"><span
                            class="ocrx_word" dir="ltr" id="word_42" lang="SC8b-R8-D2b"
                            title="bbox 1917 1166 2003 1176; x_wconf 73; pred 0; noiseConf 0.9910"
                                ><em>-.--</em></span>
                    </span>
                </p>
            </div>
            <div class="ocr_carea" id="block_8_8" title="bbox 1155 1200 2003 1410">
                <p class="ocr_par" dir="ltr" id="par_8" title="bbox 1155 1200 2003 1410">
                    <span class="ocr_line" id="line_15"
                        title="bbox 1432 1200 1705 1260; baseline -0.0036630037 -13"><span
                            class="ocrx_word" dir="ltr" id="word_43" lang="SC8b-R8-D2b"
                            title="bbox 1432 1200 1543 1247; x_wconf 81; pred 1; noiseConf 0.0015"
                                ><em>Heb.</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_44" lang="SC8b-R8-D2b"
                            title="bbox 1562 1221 1577 1247; x_wconf 76; pred 1; noiseConf 0.0129"
                            >r</span>
                        <span class="ocrx_word" dir="ltr" id="word_45" lang="SC8b-R8-D2b"
                            title="bbox 1590 1218 1626 1245; x_wconf 88; pred 1; noiseConf 0.0028"
                                ><em>2.</em></span>
                        <span class="ocrx_word" dir="ltr" id="word_46" lang="SC8b-R8-D2b"
                            title="bbox 1640 1221 1654 1246; x_wconf 82; pred 1; noiseConf 0.0115"
                            >1</span>
                        <span class="ocrx_word" dir="ltr" id="word_47" lang="SC8b-R8-D2b"
                            title="bbox 1665 1224 1705 1260; x_wconf 65; pred 1; noiseConf 0.0091"
                                ><em>4.</em></span>
```

```
                </span>
                <span class="ocr_line" id="line_16"
                    title="bbox 1155 1259 2003 1322; baseline 0.014150943 -27"><span
                        class="ocrx_word" dir="ltr" id="word_48" lang="SC8b-R8-D2b"
                        title="bbox 1155 1259 1285 1298; x_wconf 74; pred 1; noiseConf 0.0052"
                            ><em>Follow</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_49" lang="SC8b-R8-D2b"
                        title="bbox 1293 1270 1400 1311; x_wconf 78; pred 1; noiseConf 0.0021"
                            ><em>peace</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_50" lang="SC8b-R8-D2b"
                        title="bbox 1413 1260 1510 1300; x_wconf 79; pred 1; noiseConf 0.0018"
                            ><em>with</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_51" lang="SC8b-R8-D2b"
                        title="bbox 1524 1278 1546 1299; x_wconf 77; pred 1; noiseConf 0.0067"
                            ><em>a</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_52" lang="SC8b-R8-D2b"
                        title="bbox 1562 1265 1576 1300; x_wconf 84; pred 1; noiseConf 0.0186"
                            ><em>l</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_53" lang="SC8b-R8-D2b"
                        title="bbox 1591 1277 1672 1301; x_wconf 77; pred 1; noiseConf 0.0038"
                            ><em>men</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_54" lang="SC8b-R8-D2b"
                        title="bbox 1685 1294 1698 1312; x_wconf 83; pred 1; noiseConf 0.0218"
                        >'</span>
                    <span class="ocrx_word" dir="ltr" id="word_55" lang="SC8b-R8-D2b"
                        title="bbox 1729 1264 1808 1305; x_wconf 75; pred 1; noiseConf 0.0031"
                            ><em>and</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_56" lang="SC8b-R8-D2b"
                        title="bbox 1833 1264 2003 1322; x_wconf 61; pred 1; noiseConf 0.0126"
                            ><em>holincss,</em></span>
                </span>
                <span class="ocr_line" id="line_17"
                    title="bbox 1213 1314 2002 1376; baseline 0.020278834 -26"><span
                        class="ocrx_word" dir="ltr" id="word_57" lang="SC8b-R8-D2b"
                        title="bbox 1213 1314 1376 1352; x_wconf 65; pred 1; noiseConf 0.0054"
                            ><em>vvithout</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_58" lang="SC8b-R8-D2b"
                        title="bbox 1405 1315 1533 1356; x_wconf 69; pred 1; noiseConf 0.0031"
                            ><em>which</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_59" lang="SC8b-R8-D2b"
                        title="bbox 1557 1333 1600 1359; x_wconf 78; pred 1; noiseConf 0.0034"
                            ><em>no</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_60" lang="SC8b-R8-D2b"
                        title="bbox 1621 1320 1823 1375; x_wconf 68; pred 1; noiseConf 0.0034"
                            ><em>m.-mstqall</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_61" lang="SC8b-R8-D2b"
                        title="bbox 1844 1324 1908 1376; x_wconf 70; pred 1; noiseConf 0.0073"
                            ><em>stc</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_62" lang="SC8b-R8-D2b"
                        title="bbox 1937 1325 2002 1366; x_wconf 72; pred 1; noiseConf 0.0070"
                            ><em>tht</em></span>
                </span>
                <span class="ocr_line" id="line_18"
                    title="bbox 1212 1370 1328 1410; baseline 0.0086206897 -2"><span
                        class="ocrx_word" dir="ltr" id="word_63" lang="SC8b-R8-D2b"
                        title="bbox 1212 1370 1328 1410; x_wconf 82; pred 1; noiseConf 0.0036"
                            ><em>Lord.</em></span>
                </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_9_9" title="bbox 1153 1460 2000 1466">
            <p class="ocr_par" dir="ltr" id="par_9" title="bbox 1153 1460 2000 1466">
                <span class="ocr_line" id="line_19"
                    title="bbox 1153 1460 2000 1466; baseline 0 376"><span class="ocrx_word"
                        dir="ltr" id="word_64" lang="SC8b-R8-D2b"
```

```
                    title="bbox 1153 1460 2000 1466; x_wconf 95; pred 0; noiseConf 0.9500"
                        ><em> </em></span>
                </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_10_10" title="bbox 1149 1486 2179 1716">
            <p class="ocr_par" dir="ltr" id="par_10" title="bbox 1149 1486 2179 1675">
                <span class="ocr_line" id="line_20"
                    title="bbox 1422 1486 2179 1566; baseline 0.0052840159 -24"><span
                        class="ocrx_word" dir="ltr" id="word_65" lang="SC8b-R8-D2b"
                        title="bbox 1422 1504 1460 1542; x_wconf 87; pred 1; noiseConf 0.0023"
                            ><em>L</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_66" lang="SC8b-R8-D2b"
                        title="bbox 1475 1509 1506 1546; x_wconf 89; pred 1; noiseConf 0.0027"
                            ><em>o</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_67" lang="SC8b-R8-D2b"
                        title="bbox 1518 1506 1568 1545; x_wconf 89; pred 1; noiseConf 0.0017"
                            ><em>N</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_68" lang="SC8b-R8-D2b"
                        title="bbox 1582 1507 1617 1544; x_wconf 89; pred 1; noiseConf 0.0023"
                        >D</span>
                    <span class="ocrx_word" dir="ltr" id="word_69" lang="SC8b-R8-D2b"
                        title="bbox 1633 1509 1665 1544; x_wconf 83; pred 1; noiseConf 0.0029"
                            ><em>o</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_70" lang="SC8b-R8-D2b"
                        title="bbox 1682 1510 1743 1556; x_wconf 87; pred 1; noiseConf 0.0018"
                            ><em>N'</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_71" lang="SC8b-R8-D2b"
                        title="bbox 1872 1557 1875 1563; x_wconf 43; pred 0; noiseConf 0.9630"
                        >.</span>
                    <span class="ocrx_word" dir="ltr" id="word_72" lang="SC8b-R8-D2b"
                        title="bbox 2128 1511 2135 1520; x_wconf 84; pred 0; noiseConf 0.9521"
                        >-</span>
                    <span class="ocrx_word" id="word_73" lang="SC8b-R8-D2b"
                        title="bbox 2158 1486 2179 1566; x_wconf 0; pred 0; noiseConf 0.9996"
                        ></span>
                </span>
                <span class="ocr_line" id="line_21"
                    title="bbox 1149 1554 2179 1615; baseline 0.0077669903 -15"><span
                        class="ocrx_word" dir="ltr" id="word_74" lang="SC8b-R8-D2b"
                        title="bbox 1149 1554 1315 1600; x_wconf 77; pred 1; noiseConf 0.0023"
                            ><em>Psïnted</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_75" lang="SC8b-R8-D2b"
                        title="bbox 1330 1555 1384 1612; x_wconf 86; pred 1; noiseConf 0.0034"
                            ><em>by</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_76" lang="SC8b-R8-D2b"
                        title="bbox 1402 1564 1451 1601; x_wconf 81; pred 1; noiseConf 0.0021"
                            ><em>R-.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_77" lang="SC8b-R8-D2b"
                        title="bbox 1465 1563 1502 1601; x_wconf 79; pred 1; noiseConf 0.0030"
                            ><em>I.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_78" lang="SC8b-R8-D2b"
                        title="bbox 1525 1558 1590 1604; x_wconf 83; pred 1; noiseConf 0.0016"
                            ><em>for</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_79" lang="SC8b-R8-D2b"
                        title="bbox 1608 1564 1765 1603; x_wconf 73; pred 1; noiseConf 0.0020"
                            ><em>Andrew</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_80" lang="SC8b-R8-D2b"
                        title="bbox 1784 1563 1949 1614; x_wconf 75; pred 1; noiseConf 0.0028"
                            ><em>Kembc'</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_81" lang="SC8b-R8-D2b"
                        title="bbox 1962 1576 2003 1608; x_wconf 83; pred 1; noiseConf 0.0082"
                            ><em>at</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_82" lang="SC8b-R8-D2b"
                        title="bbox 2163 1569 2179 1615; x_wconf 54; pred 0; noiseConf 0.9278"
```

```
                    >j</span>
                </span>
                <span class="ocr_line" id="line_22"
                    title="bbox 1163 1614 2179 1675; baseline 0.018700787 -25"><span
                        class="ocrx_word" dir="ltr" id="word_83" lang="SC8b-R8-D2b"
                        title="bbox 1163 1614 1452 1672; x_wconf 74; pred 1; noiseConf 0.0016"
                            ><em>sr.Ma,-gare:s</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_84" lang="SC8b-R8-D2b"
                        title="bbox 1468 1615 1561 1660; x_wconf 79; pred 1; noiseConf 0.0016"
                        >Hill</span>
                    <span class="ocrx_word" dir="ltr" id="word_85" lang="SC8b-R8-D2b"
                        title="bbox 1577 1618 1617 1660; x_wconf 82; pred 1; noiseConf 0.0025"
                            ><em>iu</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_86" lang="SC8b-R8-D2b"
                        title="bbox 1635 1619 1987 1675; x_wconf 73; pred 1; noiseConf 0.0015"
                            ><em>Scm-hwark;And</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_87" lang="SC8b-R8-D2b"
                        title="bbox 2175 1622 2179 1659; x_wconf 65; pred 0; noiseConf 1.0000"
                        >z</span>
                </span>
            </p>
        </div>
        <div class="ocr_carea" id="block_11_11" title="bbox 1195 1668 1953 1758">
            <p class="ocr_par" dir="ltr" id="par_11" title="bbox 1195 1668 1953 1758">
                <span class="ocr_line" id="line_23"
                    title="bbox 1195 1668 1953 1719; baseline 0.0092348285 -15"><span
                        class="ocrx_word" dir="ltr" id="word_88" lang="SC8b-R8-D2b"
                        title="bbox 1195 1682 1246 1705; x_wconf 79; pred 1; noiseConf 0.0030"
                            ><em>are</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_89" lang="SC8b-R8-D2b"
                        title="bbox 1258 1682 1293 1704; x_wconf 85; pred 1; noiseConf 0.0027"
                        >to</span>
                    <span class="ocrx_word" dir="ltr" id="word_90" lang="SC8b-R8-D2b"
                        title="bbox 1303 1668 1360 1704; x_wconf 83; pred 1; noiseConf 0.0018"
                            ><em>bee</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_91" lang="SC8b-R8-D2b"
                        title="bbox 1375 1668 1442 1705; x_wconf 81; pred 1; noiseConf 0.0018"
                            ><em>fold</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_92" lang="SC8b-R8-D2b"
                        title="bbox 1455 1673 1550 1707; x_wconf 87; pred 1; noiseConf 0.0013"
                            ><em>under</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_93" lang="SC8b-R8-D2b"
                        title="bbox 1565 1675 1610 1708; x_wconf 83; pred 1; noiseConf 0.0020"
                            ><em>St.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_94" lang="SC8b-R8-D2b"
                        title="bbox 1621 1675 1748 1719; x_wconf 64; pred 1; noiseConf 0.0058"
                            ><em>,M.:rga.</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_95" lang="SC8b-R8-D2b"
                        title="bbox 1758 1683 1803 1709; x_wconf 71; pred 1; noiseConf 0.0051"
                            ><em>ers</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_96" lang="SC8b-R8-D2b"
                        title="bbox 1820 1676 1953 1714; x_wconf 70; pred 1; noiseConf 0.0034"
                            ><em>Church</em></span>
                </span>
                <span class="ocr_line" id="line_24"
                    title="bbox 1287 1718 1848 1758; baseline 0.016042781 -9"><span
                        class="ocrx_word" dir="ltr" id="word_97" lang="SC8b-R8-D2b"
                        title="bbox 1287 1727 1331 1750; x_wconf 72; pred 1; noiseConf 0.0049"
                            ><em>on</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_98" lang="SC8b-R8-D2b"
                        title="bbox 1346 1718 1616 1755; x_wconf 76; pred 1; noiseConf 0.0010"
                            ><em>New-Filhstreet</em></span>
                    <span class="ocrx_word" dir="ltr" id="word_99" lang="SC8b-R8-D2b"
                        title="bbox 1628 1721 1711 1755; x_wconf 81; pred 1; noiseConf 0.0016"
                        >Hill.</span>
```

```
                        <span class="ocrx_word" dir="ltr" id="word_100" lang="SC8b-R8-D2b"
                            title="bbox 1744 1729 1848 1758; x_wconf 76; pred 1; noiseConf 0.0026"
                                ><em>166.-).</em></span>
                    </span>
                </p>
            </div>
            <div class="ocr_carea" id="block_12_12" title="bbox 0 0 2179 1842">
                <p class="ocr_par" dir="ltr" id="par_12" title="bbox 0 0 2179 1842">
                    <span class="ocr_line" id="line_25" title="bbox 0 0 2179 1842; baseline 0 0"
                            ><span class="ocrx_word" dir="ltr" id="word_101" lang="SC8b-R8-D2b"
                            title="bbox 0 0 2179 1842; x_wconf 95; pred 0; noiseConf 0.9500"
                                ><em> </em></span>
                    </span>
                </p>
            </div>
        </div>
    </body>
</html>
```

A.3 De-noised, Corrected ALTO Output (eMOP work_id 32, page1)

```
<alto xmlns="http://schema.ccs-gmbh.com/ALTO" xmlns:emop="http://emop.tamu.edu">
  <Description>
    <MeasurementUnit>pixel</MeasurementUnit>
    <sourceImageInformation>
      <filename>/dh/data/eebo/e0031/40133/00001.000.001.tif</filename>
    </sourceImageInformation>
    <OCRProcessing>
      <preProcessingStep/>
      <ocrProcessingStep>
        <processingSoftware>
          <softwareCreator>Google</softwareCreator>
          <softwareName>Tesseract</softwareName>
          <softwareVersion>tesseract 3.03</softwareVersion>
        </processingSoftware>
      </ocrProcessingStep>
      <postProcessingStep>
        <processingSoftware>
          <softwareCreator> Illinois Informatics Institute, University of Illinois at
            Urbana-Champaign http://www.informatics.illinois.edu </softwareCreator>
          <softwareName>PageCorrector</softwareName>
          <softwareVersion>1.10.0-SNAPSHOT</softwareVersion>
        </processingSoftware>
      </postProcessingStep>
    </OCRProcessing>
  </Description>
  <Layout>
    <Page ID="page_1" WIDTH="2179" HEIGHT="1842">
      <PrintSpace>
        <TextBlock ID="par_1" WIDTH="968" HEIGHT="288" HPOS="1211" VPOS="22">
          <TextLine ID="line_1" WIDTH="619" HEIGHT="105" HPOS="22" VPOS="1560">
            <String ID="word_1" WIDTH="65" HEIGHT="70" HPOS="24" VPOS="1560" CONTENT="A" WC="79"
              emop:DNC="0.0096"> </String>
          </TextLine>
```

```
      <TextLine ID="line_2" WIDTH="736" HEIGHT="83" HPOS="132" VPOS="1211">
        <String ID="word_3" WIDTH="260" HEIGHT="77" HPOS="132" VPOS="1211" CONTENT="Serious"
          WC="73" emop:DNC="0.0048"> </String>
        <SP WIDTH="10"/>
        <String ID="word_4" WIDTH="437" HEIGHT="83" HPOS="132" VPOS="1510" CONTENT="Exhortation"
          WC="72" emop:DNC="0.0048">
          <ALTERNATIVE>Exho.rtation</ALTERNATIVE>
        </String>
      </TextLine>
      <TextLine ID="line_3" WIDTH="671" HEIGHT="57" HPOS="253" VPOS="1438">
        <String ID="word_5" WIDTH="114" HEIGHT="55" HPOS="253" VPOS="1438" CONTENT="To" WC="84"
          emop:DNC="0.0024">
          <ALTERNATIVE>T()</ALTERNATIVE>
        </String>
        <SP WIDTH="10"/>
        <String ID="word_6" WIDTH="46" HEIGHT="49" HPOS="260" VPOS="1599" CONTENT="A" WC="89"
          emop:DNC="0.0048"> </String>
        <SP WIDTH="10"/>
        <String ID="word_7" WIDTH="54" HEIGHT="52" HPOS="258" VPOS="1666" CONTENT="N" WC="92"
          emop:DNC="0.0083"> </String>
      </TextLine>
    </TextBlock>
    <TextBlock ID="par_2" WIDTH="788" HEIGHT="214" HPOS="1225" VPOS="334">
      <TextLine ID="line_4" WIDTH="788" HEIGHT="214" HPOS="334" VPOS="1225">
        <String ID="word_11" WIDTH="338" HEIGHT="199" HPOS="334" VPOS="1225" CONTENT="floly"
          WC="79" emop:DNC="0.0242"> </String>
        <SP WIDTH="10"/>
        <String ID="word_12" WIDTH="295" HEIGHT="149" HPOS="343" VPOS="1635" CONTENT="Life."
          WC="69" emop:DNC="0.0965"> </String>
      </TextLine>
    </TextBlock>
    <TextBlock ID="par_3" WIDTH="861" HEIGHT="91" HPOS="1142" VPOS="625">
      <TextLine ID="line_5" WIDTH="861" HEIGHT="91" HPOS="625" VPOS="1142">
        <String ID="word_14" WIDTH="80" HEIGHT="69" HPOS="629" VPOS="1142" CONTENT=".A" WC="21"
          emop:DNC="0.0883"> </String>
        <SP WIDTH="10"/>
        <String ID="word_15" WIDTH="153" HEIGHT="73" HPOS="625" VPOS="1241" CONTENT="Plea"
          WC="83" emop:DNC="0.0041"> </String>
        <SP WIDTH="10"/>
        <String ID="word_16" WIDTH="239" HEIGHT="73" HPOS="628" VPOS="1440" CONTENT="forthe"
          WC="74" emop:DNC="0.003"> </String>
        <SP WIDTH="10"/>
        <String ID="word_17" WIDTH="283" HEIGHT="80" HPOS="636" VPOS="1720" CONTENT="absolute"
          WC="80" emop:DNC="0.0028"> </String>
      </TextLine>
    </TextBlock>
    <TextBlock ID="par_4" WIDTH="739" HEIGHT="295" HPOS="1200" VPOS="724">
      <TextLine ID="line_6" WIDTH="696" HEIGHT="83" HPOS="724" VPOS="1222">
        <String ID="word_18" WIDTH="247" HEIGHT="83" HPOS="724" VPOS="1222"
          CONTENT="flaïc(:efficy" WC="55" emop:DNC="0.1122"> </String>
        <SP WIDTH="10"/>
        <String ID="word_19" WIDTH="92" HEIGHT="58" HPOS="728" VPOS="1518" CONTENT="dof" WC="25"
          emop:DNC="0.0346"> </String>
        <SP WIDTH="10"/>
        <String ID="word_20" WIDTH="257" HEIGHT="61" HPOS="736" VPOS="1661" CONTENT="Inhcrcnt"
          WC="82" emop:DNC="0.0011"> </String>
      </TextLine>
      <TextLine ID="line_7" WIDTH="597" HEIGHT="64" HPOS="807" VPOS="1277">
        <String ID="word_21" WIDTH="366" HEIGHT="63" HPOS="808" VPOS="1277"
          CONTENT="Righteousness" WC="40" emop:DNC="0.042">
          <ALTERNATIVE>-Rjghteousness</ALTERNATIVE>
        </String>
        <SP WIDTH="10"/>
        <String ID="word_22" WIDTH="46" HEIGHT="52" HPOS="807" VPOS="1676" CONTENT="ill" WC="67"
          emop:DNC="0.0112">
```

```
            <ALTERNATIVE>ils</ALTERNATIVE>
            <ALTERNATIVE>iis</ALTERNATIVE>
            <ALTERNATIVE>ilf</ALTERNATIVE>
          </String>
          <SP WIDTH="10"/>
          <String ID="word_23" WIDTH="126" HEIGHT="48" HPOS="817" VPOS="1748" CONTENT="those"
            WC="69" emop:DNC="0.0036"> </String>
        </TextLine>
        <TextLine ID="line_8" WIDTH="488" HEIGHT="54" HPOS="872" VPOS="1335">
          <String ID="word_24" WIDTH="91" HEIGHT="45" HPOS="872" VPOS="1335" CONTENT="that"
            WC="87" emop:DNC="0.0021"> </String>
          <SP WIDTH="10"/>
          <String ID="word_25" WIDTH="109" HEIGHT="53" HPOS="873" VPOS="1445" CONTENT="hope"
            WC="77" emop:DNC="0.0022"> </String>
          <SP WIDTH="10"/>
          <String ID="word_26" WIDTH="43" HEIGHT="31" HPOS="887" VPOS="1570" CONTENT="to" WC="87"
            emop:DNC="0.0028"> </String>
          <SP WIDTH="10"/>
          <String ID="word_27" WIDTH="47" HEIGHT="47" HPOS="873" VPOS="1631" CONTENT="be" WC="77"
            emop:DNC="0.0039"> </String>
          <SP WIDTH="10"/>
          <String ID="word_28" WIDTH="129" HEIGHT="47" HPOS="877" VPOS="1694" CONTENT="saved."
            WC="78" emop:DNC="0.002"> </String>
        </TextLine>
        <TextLine ID="line_9" WIDTH="48" HEIGHT="30" HPOS="928" VPOS="1406">
          <String SUBS_TYPE="HypPart1" SUBS_CONTENT="ZBy" ID="word_29" WIDTH="48" HEIGHT="30"
            HPOS="928" VPOS="1406" CONTENT="Z" WC="61" emop:DNC="0.0121"> </String>
          <HYP WIDTH="10" CONTENT="-"/>
        </TextLine>
        <TextLine ID="line_10" WIDTH="739" HEIGHT="91" HPOS="928" VPOS="1200">
          <String SUBS_TYPE="HypPart2" SUBS_CONTENT="ZBy" ID="word_30" WIDTH="62" HEIGHT="56"
            HPOS="958" VPOS="1200" CONTENT="By" WC="89" emop:DNC="0.0121"> </String>
          <SP WIDTH="10"/>
          <String ID="word_31" WIDTH="375" HEIGHT="91" HPOS="928" VPOS="1287"
            CONTENT="YT'o.JlWQdsworth-j" WC="48" emop:DNC="0.4293"> </String>
          <SP WIDTH="10"/>
          <String ID="word_32" WIDTH="195" HEIGHT="46" HPOS="963" VPOS="1683" CONTENT="preacher"
            WC="74" emop:DNC="0.0018"> </String>
          <SP WIDTH="10"/>
          <String ID="word_33" WIDTH="46" HEIGHT="32" HPOS="979" VPOS="1893" CONTENT="to" WC="89"
            emop:DNC="0.0044"> </String>
        </TextLine>
      </TextBlock>
      <TextBlock ID="par_5" WIDTH="587" HEIGHT="115" HPOS="1291" VPOS="1013">
        <TextLine ID="line_11" WIDTH="587" HEIGHT="61" HPOS="1013" VPOS="1291">
          <String ID="word_34" WIDTH="68" HEIGHT="45" HPOS="1014" VPOS="1291" CONTENT="one"
            WC="86" emop:DNC="0.0028">
            <ALTERNATIVE>oho</ALTERNATIVE>
            <ALTERNATIVE>olio</ALTERNATIVE>
            <ALTERNATIVE>che</ALTERNATIVE>
          </String>
          <SP WIDTH="10"/>
          <String ID="word_35" WIDTH="165" HEIGHT="44" HPOS="1013" VPOS="1385" CONTENT="Glyn-reb"
            WC="38" emop:DNC="0.0395"> </String>
          <SP WIDTH="10"/>
          <String ID="word_36" WIDTH="39" HEIGHT="29" HPOS="1029" VPOS="1565" CONTENT="at" WC="83"
            emop:DNC="0.003"> </String>
          <SP WIDTH="10"/>
          <String SUBS_TYPE="HypPart1" SUBS_CONTENT="NewingtonButts" ID="word_37" WIDTH="256"
            HEIGHT="56" HPOS="1018" VPOS="1622" CONTENT="Newington" WC="67" emop:DNC="0.003">
</String>
          <HYP WIDTH="10" CONTENT="-"/>
        </TextLine>
        <TextLine ID="line_12" WIDTH="336" HEIGHT="57" HPOS="1071" VPOS="1412">
          <String SUBS_TYPE="HypPart2" SUBS_CONTENT="NewingtonButts" ID="word_38" WIDTH="117"
```

```
              HEIGHT="37" HPOS="1073" VPOS="1412" CONTENT="Butts" WC="81" emop:DNC="0.003">
</String>
            <SP WIDTH="10"/>
            <String ID="word_39" WIDTH="40" HEIGHT="42" HPOS="1071" VPOS="1542" CONTENT="in" WC="80"
              emop:DNC="0.0032"> </String>
            <SP WIDTH="10"/>
            <String ID="word_40" WIDTH="151" HEIGHT="52" HPOS="1076" VPOS="1597" CONTENT="Surrey."
              WC="73" emop:DNC="0.0022"> </String>
          </TextLine>
        </TextBlock>
        <TextBlock ID="par_6" WIDTH="746" HEIGHT="17" HPOS="1165" VPOS="1153"> </TextBlock>
        <TextBlock ID="par_7" WIDTH="86" HEIGHT="10" HPOS="1917" VPOS="1166"> </TextBlock>
        <TextBlock ID="par_8" WIDTH="848" HEIGHT="210" HPOS="1155" VPOS="1200">
          <TextLine ID="line_15" WIDTH="273" HEIGHT="60" HPOS="1200" VPOS="1432">
            <String ID="word_43" WIDTH="111" HEIGHT="47" HPOS="1200" VPOS="1432" CONTENT="Heb."
              WC="81" emop:DNC="0.0015"> </String>
            <SP WIDTH="10"/>
            <String ID="word_44" WIDTH="15" HEIGHT="26" HPOS="1221" VPOS="1562" CONTENT="r" WC="76"
              emop:DNC="0.0129"> </String>
            <SP WIDTH="10"/>
            <String ID="word_45" WIDTH="36" HEIGHT="27" HPOS="1218" VPOS="1590" CONTENT="2." WC="88"
              emop:DNC="0.0028"> </String>
            <SP WIDTH="10"/>
            <String ID="word_46" WIDTH="14" HEIGHT="25" HPOS="1221" VPOS="1640" CONTENT="1" WC="82"
              emop:DNC="0.0115"> </String>
            <SP WIDTH="10"/>
            <String ID="word_47" WIDTH="40" HEIGHT="36" HPOS="1224" VPOS="1665" CONTENT="4." WC="65"
              emop:DNC="0.0091"> </String>
          </TextLine>
          <TextLine ID="line_16" WIDTH="848" HEIGHT="63" HPOS="1259" VPOS="1155">
            <String ID="word_48" WIDTH="130" HEIGHT="39" HPOS="1259" VPOS="1155" CONTENT="Follow"
              WC="74" emop:DNC="0.0052"> </String>
            <SP WIDTH="10"/>
            <String ID="word_49" WIDTH="107" HEIGHT="41" HPOS="1270" VPOS="1293" CONTENT="peace"
              WC="78" emop:DNC="0.0021"> </String>
            <SP WIDTH="10"/>
            <String ID="word_50" WIDTH="97" HEIGHT="40" HPOS="1260" VPOS="1413" CONTENT="with"
              WC="79" emop:DNC="0.0018"> </String>
            <SP WIDTH="10"/>
            <String ID="word_51" WIDTH="22" HEIGHT="21" HPOS="1278" VPOS="1524" CONTENT="a" WC="77"
              emop:DNC="0.0067"> </String>
            <SP WIDTH="10"/>
            <String ID="word_52" WIDTH="14" HEIGHT="35" HPOS="1265" VPOS="1562" CONTENT="l" WC="84"
              emop:DNC="0.0186"> </String>
            <SP WIDTH="10"/>
            <String ID="word_53" WIDTH="81" HEIGHT="24" HPOS="1277" VPOS="1591" CONTENT="men"
              WC="77" emop:DNC="0.0038"> </String>
            <SP WIDTH="10"/>
            <String ID="word_54" WIDTH="13" HEIGHT="18" HPOS="1294" VPOS="1685" CONTENT="'" WC="83"
              emop:DNC="0.0218"> </String>
            <SP WIDTH="10"/>
            <String ID="word_55" WIDTH="79" HEIGHT="41" HPOS="1264" VPOS="1729" CONTENT="and"
              WC="75" emop:DNC="0.0031"> </String>
            <SP WIDTH="10"/>
            <String ID="word_56" WIDTH="170" HEIGHT="58" HPOS="1264" VPOS="1833" CONTENT="holiness,"
              WC="61" emop:DNC="0.0126">
              <ALTERNATIVE>nosiness,</ALTERNATIVE>
              <ALTERNATIVE>holincss,</ALTERNATIVE>
            </String>
          </TextLine>
          <TextLine ID="line_17" WIDTH="789" HEIGHT="62" HPOS="1314" VPOS="1213">
            <String ID="word_57" WIDTH="163" HEIGHT="38" HPOS="1314" VPOS="1213" CONTENT="without"
              WC="65" emop:DNC="0.0054">
              <ALTERNATIVE>vvithout</ALTERNATIVE>
            </String>
```

```
        <SP WIDTH="10"/>
        <String ID="word_58" WIDTH="128" HEIGHT="41" HPOS="1315" VPOS="1405" CONTENT="which"
          WC="69" emop:DNC="0.0031"> </String>
        <SP WIDTH="10"/>
        <String ID="word_59" WIDTH="43" HEIGHT="26" HPOS="1333" VPOS="1557" CONTENT="no" WC="78"
          emop:DNC="0.0034"> </String>
        <SP WIDTH="10"/>
        <String ID="word_60" WIDTH="202" HEIGHT="55" HPOS="1320" VPOS="1621"
          CONTENT="m.-mstqall" WC="68" emop:DNC="0.0034"> </String>
        <SP WIDTH="10"/>
        <String ID="word_61" WIDTH="64" HEIGHT="52" HPOS="1324" VPOS="1844" CONTENT="stc"
          WC="70" emop:DNC="0.0073"> </String>
        <SP WIDTH="10"/>
        <String ID="word_62" WIDTH="65" HEIGHT="41" HPOS="1325" VPOS="1937" CONTENT="tht"
          WC="72" emop:DNC="0.007"> </String>
      </TextLine>
      <TextLine ID="line_18" WIDTH="116" HEIGHT="40" HPOS="1370" VPOS="1212">
        <String ID="word_63" WIDTH="116" HEIGHT="40" HPOS="1370" VPOS="1212" CONTENT="Lord."
          WC="82" emop:DNC="0.0036"> </String>
      </TextLine>
    </TextBlock>
    <TextBlock ID="par_9" WIDTH="847" HEIGHT="6" HPOS="1153" VPOS="1460"> </TextBlock>
    <TextBlock ID="par_10" WIDTH="1030" HEIGHT="189" HPOS="1149" VPOS="1486">
      <TextLine ID="line_20" WIDTH="757" HEIGHT="80" HPOS="1486" VPOS="1422">
        <String ID="word_65" WIDTH="38" HEIGHT="38" HPOS="1504" VPOS="1422" CONTENT="L" WC="87"
          emop:DNC="0.0023"> </String>
        <SP WIDTH="10"/>
        <String ID="word_66" WIDTH="31" HEIGHT="37" HPOS="1509" VPOS="1475" CONTENT="o" WC="89"
          emop:DNC="0.0027"> </String>
        <SP WIDTH="10"/>
        <String ID="word_67" WIDTH="50" HEIGHT="39" HPOS="1506" VPOS="1518" CONTENT="N" WC="89"
          emop:DNC="0.0017"> </String>
        <SP WIDTH="10"/>
        <String ID="word_68" WIDTH="35" HEIGHT="37" HPOS="1507" VPOS="1582" CONTENT="D" WC="89"
          emop:DNC="0.0023"> </String>
        <SP WIDTH="10"/>
        <String ID="word_69" WIDTH="32" HEIGHT="35" HPOS="1509" VPOS="1633" CONTENT="o" WC="83"
          emop:DNC="0.0029"> </String>
        <SP WIDTH="10"/>
        <String ID="word_70" WIDTH="61" HEIGHT="46" HPOS="1510" VPOS="1682" CONTENT="N'" WC="87"
          emop:DNC="0.0018"> </String>
      </TextLine>
      <TextLine ID="line_21" WIDTH="1030" HEIGHT="61" HPOS="1554" VPOS="1149">
        <String ID="word_74" WIDTH="166" HEIGHT="46" HPOS="1554" VPOS="1149" CONTENT="Psïnted"
          WC="77" emop:DNC="0.0023"> </String>
        <SP WIDTH="10"/>
        <String ID="word_75" WIDTH="54" HEIGHT="57" HPOS="1555" VPOS="1330" CONTENT="by" WC="86"
          emop:DNC="0.0034"> </String>
        <SP WIDTH="10"/>
        <String ID="word_76" WIDTH="49" HEIGHT="37" HPOS="1564" VPOS="1402" CONTENT="R-."
          WC="81" emop:DNC="0.0021"> </String>
        <SP WIDTH="10"/>
        <String ID="word_77" WIDTH="37" HEIGHT="38" HPOS="1563" VPOS="1465" CONTENT="I." WC="79"
          emop:DNC="0.003"> </String>
        <SP WIDTH="10"/>
        <String ID="word_78" WIDTH="65" HEIGHT="46" HPOS="1558" VPOS="1525" CONTENT="for"
          WC="83" emop:DNC="0.0016"> </String>
        <SP WIDTH="10"/>
        <String ID="word_79" WIDTH="157" HEIGHT="39" HPOS="1564" VPOS="1608" CONTENT="Andrew"
          WC="73" emop:DNC="0.002"> </String>
        <SP WIDTH="10"/>
        <String ID="word_80" WIDTH="165" HEIGHT="51" HPOS="1563" VPOS="1784" CONTENT="Kembc'"
          WC="75" emop:DNC="0.0028"> </String>
        <SP WIDTH="10"/>
        <String ID="word_81" WIDTH="41" HEIGHT="32" HPOS="1576" VPOS="1962" CONTENT="at" WC="83"
```

```
            emop:DNC="0.0082"> </String>
        </TextLine>
        <TextLine ID="line_22" WIDTH="1016" HEIGHT="61" HPOS="1614" VPOS="1163">
          <String ID="word_83" WIDTH="289" HEIGHT="58" HPOS="1614" VPOS="1163"
            CONTENT="sr.Ma,-gare:s" WC="74" emop:DNC="0.0016"> </String>
          <SP WIDTH="10"/>
          <String ID="word_84" WIDTH="93" HEIGHT="45" HPOS="1615" VPOS="1468" CONTENT="Hill"
            WC="79" emop:DNC="0.0016"> </String>
          <SP WIDTH="10"/>
          <String ID="word_85" WIDTH="40" HEIGHT="42" HPOS="1618" VPOS="1577" CONTENT="iu" WC="82"
            emop:DNC="0.0025"> </String>
          <SP WIDTH="10"/>
          <String ID="word_86" WIDTH="352" HEIGHT="56" HPOS="1619" VPOS="1635"
            CONTENT="Scm-hwark;And" WC="73" emop:DNC="0.0015"> </String>
        </TextLine>
      </TextBlock>
      <TextBlock ID="par_11" WIDTH="758" HEIGHT="90" HPOS="1195" VPOS="1668">
        <TextLine ID="line_23" WIDTH="758" HEIGHT="51" HPOS="1668" VPOS="1195">
          <String ID="word_88" WIDTH="51" HEIGHT="23" HPOS="1682" VPOS="1195" CONTENT="are"
            WC="79" emop:DNC="0.003"> </String>
          <SP WIDTH="10"/>
          <String ID="word_89" WIDTH="35" HEIGHT="22" HPOS="1682" VPOS="1258" CONTENT="to" WC="85"
            emop:DNC="0.0027"> </String>
          <SP WIDTH="10"/>
          <String ID="word_90" WIDTH="57" HEIGHT="36" HPOS="1668" VPOS="1303" CONTENT="bee"
            WC="83" emop:DNC="0.0018"> </String>
          <SP WIDTH="10"/>
          <String ID="word_91" WIDTH="67" HEIGHT="37" HPOS="1668" VPOS="1375" CONTENT="fold"
            WC="81" emop:DNC="0.0018"> </String>
          <SP WIDTH="10"/>
          <String ID="word_92" WIDTH="95" HEIGHT="34" HPOS="1673" VPOS="1455" CONTENT="under"
            WC="87" emop:DNC="0.0013"> </String>
          <SP WIDTH="10"/>
          <String ID="word_93" WIDTH="45" HEIGHT="33" HPOS="1675" VPOS="1565" CONTENT="St."
            WC="83" emop:DNC="0.002"> </String>
          <SP WIDTH="10"/>
          <String ID="word_94" WIDTH="127" HEIGHT="44" HPOS="1675" VPOS="1621" CONTENT=",M.:rga."
            WC="64" emop:DNC="0.0058"> </String>
          <SP WIDTH="10"/>
          <String ID="word_95" WIDTH="45" HEIGHT="26" HPOS="1683" VPOS="1758" CONTENT="ers"
            WC="71" emop:DNC="0.0051"> </String>
          <SP WIDTH="10"/>
          <String ID="word_96" WIDTH="133" HEIGHT="38" HPOS="1676" VPOS="1820" CONTENT="Church"
            WC="70" emop:DNC="0.0034"> </String>
        </TextLine>
        <TextLine ID="line_24" WIDTH="561" HEIGHT="40" HPOS="1718" VPOS="1287">
          <String ID="word_97" WIDTH="44" HEIGHT="23" HPOS="1727" VPOS="1287" CONTENT="on" WC="72"
            emop:DNC="0.0049"> </String>
          <SP WIDTH="10"/>
          <String ID="word_98" WIDTH="270" HEIGHT="37" HPOS="1718" VPOS="1346"
            CONTENT="New-Filhstreet" WC="76" emop:DNC="0.001"> </String>
          <SP WIDTH="10"/>
          <String ID="word_99" WIDTH="83" HEIGHT="34" HPOS="1721" VPOS="1628" CONTENT="Hill."
            WC="81" emop:DNC="0.0016"> </String>
          <SP WIDTH="10"/>
          <String ID="word_100" WIDTH="104" HEIGHT="29" HPOS="1729" VPOS="1744" CONTENT="166.-)."
            WC="76" emop:DNC="0.0026"> </String>
        </TextLine>
      </TextBlock>
      <TextBlock ID="par_12" WIDTH="2179" HEIGHT="1842" HPOS="0" VPOS="0"> </TextBlock>
    </PrintSpace>
  </Page>
 </Layout>
</alto>
```

Appendix B

B.1 Example JSON Input (request from DB)


```
[
    {
        "status": {
            "id": 2,
            "name": "Processing"
        },
        "proc_id": "20150307090823396",
        "work": {
            "wks_primary_print_font": null,
            "wks_pub_date": "1728",
            "wks_eebo_citation_id": null,
            "wks_ecco_directory": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300",
            "wks_eebo_image_id": null,
            "wks_estc_number": "W037851",
            "id": 444001,
            "wks_eebo_directory": null,
            "wks_author": "Vincent, Nathanael",
            "wks_last_trawled": "2013-07-04",
            "wks_ecco_gale_ocr_xml_path": "/data/ecco/ECCOII/RelAndPhil/XML/1474101300.xml",
            "wks_organizational_unit": 444,
            "wks_ecco_corrected_text_path": null,
            "wks_bib_name": null,
            "wks_marc_record": null,
            "wks_book_id": null,
            "wks_tcp_number": null,
            "wks_ecco_uncorrected_gale_ocr_path": null,
            "wks_tcp_bibno": null,
            "wks_ecco_corrected_xml_path": null,
            "wks_ecco_number": "1474101300",
            "wks_publisher": "Boston : Re-printed for Alford Butler, and sold at his shop, the lower
end of King-Street, near the Crown Coffee-House, 1728.",
            "wks_word_count": null,
            "wks_title": "The day of grace, in which the chief of sinners may be turn'd and healed.
By Nathaniel Vincent. [Two lines from II Corinthians]",
            "wks_eebo_url": null
        },
        "results": null,
        "id": 30855678,
        "tries": 0,
        "postproc_result": null,
        "page_result": null,
        "batch_job": {
            "ocr_engine": {
                "id": 2,
                "name": "Tesseract"
            },
            "name": "ECCO w/o GT (SC8b-R7-D2b)",
            "parameters": "",
            "notes": "levenshtein; m=20; Xms/x=2048; timeout=300; ",
            "job_type": {
                "id": 2,
                "name": "OCR"
            },
            "font": {
```

```
                "font_bold": null,
                "font_line_height": null,
                "font_fixed": null,
                "font_fraktur": null,
                "font_italic": null,
                "font_library_path": null,
                "font_serif": null,
                "id": 207,
                "font_name": "SC8b-R7-D2b"
            },
            "id": 1
        },
        "page": {
            "pg_ref_number": 43,
            "work": {
                "wks_primary_print_font": null,
                "wks_pub_date": "1728",
                "wks_eebo_citation_id": null,
                "wks_ecco_directory": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300",
                "wks_eebo_image_id": null,
                "wks_estc_number": "W037851",
                "id": 444001,
                "wks_eebo_directory": null,
                "wks_author": "Vincent, Nathanael",
                "wks_last_trawled": "2013-07-04",
                "wks_ecco_gale_ocr_xml_path": "/data/ecco/ECCOII/RelAndPhil/XML/1474101300.xml",
                "wks_organizational_unit": 444,
                "wks_ecco_corrected_text_path": null,
                "wks_bib_name": null,
                "wks_marc_record": null,
                "wks_book_id": null,
                "wks_tcp_number": null,
                "wks_ecco_uncorrected_gale_ocr_path": null,
                "wks_tcp_bibno": null,
                "wks_ecco_corrected_xml_path": null,
                "wks_ecco_number": "1474101300",
                "wks_publisher": "Boston : Re-printed for Alford Butler, and sold at his shop, the
lower end of King-Street, near the Crown Coffee-House, 1728.",
                "wks_word_count": null,
                "wks_title": "The day of grace, in which the chief of sinners may be turn'd and
healed. By Nathaniel Vincent. [Two lines from II Corinthians]",
                "wks_eebo_url": null
            },
            "pg_image_path": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300/147410130000430.TIF",
            "pg_ground_truth_file": null,
            "pg_gale_ocr_file": null,
            "id": 40691559
        }
    },
    {
        "status": {
            "id": 2,
            "name": "Processing"
        },
        "proc_id": "20150307090823396",
        "work": {
            "wks_primary_print_font": null,
            "wks_pub_date": "1728",
            "wks_eebo_citation_id": null,
            "wks_ecco_directory": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300",
            "wks_eebo_image_id": null,
            "wks_estc_number": "W037851",
            "id": 444001,
            "wks_eebo_directory": null,
            "wks_author": "Vincent, Nathanael",
```

```
            "wks_last_trawled": "2013-07-04",
            "wks_ecco_gale_ocr_xml_path": "/data/ecco/ECCOII/RelAndPhil/XML/1474101300.xml",
            "wks_organizational_unit": 444,
            "wks_ecco_corrected_text_path": null,
            "wks_bib_name": null,
            "wks_marc_record": null,
            "wks_book_id": null,
            "wks_tcp_number": null,
            "wks_ecco_uncorrected_gale_ocr_path": null,
            "wks_tcp_bibno": null,
            "wks_ecco_corrected_xml_path": null,
            "wks_ecco_number": "1474101300",
            "wks_publisher": "Boston : Re-printed for Alford Butler, and sold at his shop, the lower
end of King-Street, near the Crown Coffee-House, 1728.",
            "wks_word_count": null,
            "wks_title": "The day of grace, in which the chief of sinners may be turn'd and healed.
By Nathaniel Vincent. [Two lines from II Corinthians]",
            "wks_eebo_url": null
        },
        "results": null,
        "id": 30855679,
        "tries": 0,
        "postproc_result": null,
        "page_result": null,
        "batch_job": {
            "ocr_engine": {
                "id": 2,
                "name": "Tesseract"
            },
            "name": "ECCO w/o GT (SC8b-R7-D2b)",
            "parameters": "",
            "notes": "levenshtein; m=20; Xms/x=2048; timeout=300; ",
            "job_type": {
                "id": 2,
                "name": "OCR"
            },
            "font": {
                "font_bold": null,
                "font_line_height": null,
                "font_fixed": null,
                "font_fraktur": null,
                "font_italic": null,
                "font_library_path": null,
                "font_serif": null,
                "id": 207,
                "font_name": "SC8b-R7-D2b"
            },
            "id": 1
        },
        "page": {
            "pg_ref_number": 44,
            "work": {
                "wks_primary_print_font": null,
                "wks_pub_date": "1728",
                "wks_eebo_citation_id": null,
                "wks_ecco_directory": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300",
                "wks_eebo_image_id": null,
                "wks_estc_number": "W037851",
                "id": 444001,
                "wks_eebo_directory": null,
                "wks_author": "Vincent, Nathanael",
                "wks_last_trawled": "2013-07-04",
                "wks_ecco_gale_ocr_xml_path": "/data/ecco/ECCOII/RelAndPhil/XML/1474101300.xml",
                "wks_organizational_unit": 444,
                "wks_ecco_corrected_text_path": null,
```

```
                "wks_bib_name": null,
                "wks_marc_record": null,
                "wks_book_id": null,
                "wks_tcp_number": null,
                "wks_ecco_uncorrected_gale_ocr_path": null,
                "wks_tcp_bibno": null,
                "wks_ecco_corrected_xml_path": null,
                "wks_ecco_number": "1474101300",
                "wks_publisher": "Boston : Re-printed for Alford Butler, and sold at his shop, the
lower end of King-Street, near the Crown Coffee-House, 1728.",
                "wks_word_count": null,
                "wks_title": "The day of grace, in which the chief of sinners may be turn'd and
healed. By Nathaniel Vincent. [Two lines from II Corinthians]",
                "wks_eebo_url": null
            },
            "pg_image_path": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300/147410130000440.TIF",
            "pg_ground_truth_file": null,
            "pg_gale_ocr_file": null,
            "id": 40691560
        }
    },
    {
        "status": {
            "id": 2,
            "name": "Processing"
        },
        "proc_id": "20150307090823396",
        "work": {
            "wks_primary_print_font": null,
            "wks_pub_date": "1728",
            "wks_eebo_citation_id": null,
            "wks_ecco_directory": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300",
            "wks_eebo_image_id": null,
            "wks_estc_number": "W037851",
            "id": 444001,
            "wks_eebo_directory": null,
            "wks_author": "Vincent, Nathanael",
            "wks_last_trawled": "2013-07-04",
            "wks_ecco_gale_ocr_xml_path": "/data/ecco/ECCOII/RelAndPhil/XML/1474101300.xml",
            "wks_organizational_unit": 444,
            "wks_ecco_corrected_text_path": null,
            "wks_bib_name": null,
            "wks_marc_record": null,
            "wks_book_id": null,
            "wks_tcp_number": null,
            "wks_ecco_uncorrected_gale_ocr_path": null,
            "wks_tcp_bibno": null,
            "wks_ecco_corrected_xml_path": null,
            "wks_ecco_number": "1474101300",
            "wks_publisher": "Boston : Re-printed for Alford Butler, and sold at his shop, the lower
end of King-Street, near the Crown Coffee-House, 1728.",
            "wks_word_count": null,
            "wks_title": "The day of grace, in which the chief of sinners may be turn'd and healed.
By Nathaniel Vincent. [Two lines from II Corinthians]",
            "wks_eebo_url": null
        },
        "results": null,
        "id": 30855680,
        "tries": 0,
        "postproc_result": null,
        "page_result": null,
        "batch_job": {
            "ocr_engine": {
                "id": 2,
                "name": "Tesseract"
```

```
            },
            "name": "ECCO w/o GT (SC8b-R7-D2b)",
            "parameters": "",
            "notes": "levenshtein; m=20; Xms/x=2048; timeout=300; ",
            "job_type": {
                "id": 2,
                "name": "OCR"
            },
            "font": {
                "font_bold": null,
                "font_line_height": null,
                "font_fixed": null,
                "font_fraktur": null,
                "font_italic": null,
                "font_library_path": null,
                "font_serif": null,
                "id": 207,
                "font_name": "SC8b-R7-D2b"
            },
            "id": 1
        },
        "page": {
            "pg_ref_number": 45,
            "work": {
                "wks_primary_print_font": null,
                "wks_pub_date": "1728",
                "wks_eebo_citation_id": null,
                "wks_ecco_directory": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300",
                "wks_eebo_image_id": null,
                "wks_estc_number": "W037851",
                "id": 444001,
                "wks_eebo_directory": null,
                "wks_author": "Vincent, Nathanael",
                "wks_last_trawled": "2013-07-04",
                "wks_ecco_gale_ocr_xml_path": "/data/ecco/ECCOII/RelAndPhil/XML/1474101300.xml",
                "wks_organizational_unit": 444,
                "wks_ecco_corrected_text_path": null,
                "wks_bib_name": null,
                "wks_marc_record": null,
                "wks_book_id": null,
                "wks_tcp_number": null,
                "wks_ecco_uncorrected_gale_ocr_path": null,
                "wks_tcp_bibno": null,
                "wks_ecco_corrected_xml_path": null,
                "wks_ecco_number": "1474101300",
                "wks_publisher": "Boston : Re-printed for Alford Butler, and sold at his shop, the
lower end of King-Street, near the Crown Coffee-House, 1728.",
                "wks_word_count": null,
                "wks_title": "The day of grace, in which the chief of sinners may be turn'd and
healed. By Nathaniel Vincent. [Two lines from II Corinthians]",
                "wks_eebo_url": null
            },
            "pg_image_path": "/data/ecco/ECCOII/RelAndPhil/Images/1474101300/147410130000450.TIF",
            "pg_ground_truth_file": null,
            "pg_gale_ocr_file": null,
            "id": 40691561
        }
    },
...
}
```

B.2 Example JSON Output (write to DB)


```
{
    "postproc_results": [
        {
            "skew_idx": "0.200000,",
            "multicol": "-1376.00,-1376.00,-1376.00",
            "pp_health":
"{\"total\":226,\"ignored\":13,\"correct\":134,\"corrected\":38,\"unchanged\":41}",
            "batch_job_id": 1,
            "pp_ecorr": "0.945652",
            "pp_noisemsr": "0.0934",
            "pp_pg_quality": "0.707317",
            "page_id": 40691559
        },
        {
            "skew_idx": "0.000000,",
            "multicol": "-1344.00,-1344.00,-1344.00",
            "pp_health":
"{\"total\":214,\"ignored\":11,\"correct\":129,\"corrected\":29,\"unchanged\":45}",
            "batch_job_id": 1,
            "pp_ecorr": "0.971910",
            "pp_noisemsr": "0.1070",
            "pp_pg_quality": "0.752174",
            "page_id": 40691560
        },
        {
            "skew_idx": "0.200000,",
            "multicol": "-1376.00,-1376.00,-1376.00",
            "pp_health":
"{\"total\":230,\"ignored\":7,\"correct\":166,\"corrected\":30,\"unchanged\":27}",
            "batch_job_id": 1,
            "pp_ecorr": "0.983051",
            "pp_noisemsr": "0.0667",
            "pp_pg_quality": "0.710204",
            "page_id": 40691561
        },
...
}
```