The eMOP Project: Partnering with Proprietors

Laura Mandell
@mandelllc
4 March 2015
http://www.ar-c.org

Welcome to ARC
The Advanced Research Consortium

The Advanced Research Consortium (ARC)
A consortium of scholars committed to advancing research in the humanities through period-specific online communities built around peer-review, aggregation, and searching digital data.
The Advanced Research Consortium (ARC)
Building Bridges between Academic Networks

Elizabeth Grumbach [@EMGrumbach], ARC Project Manager
Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University

As a hub of humanities research nodes, ARC contains scholarly resources spanning the bulk of existing Western-written documents, scanned page images, text transcriptions, scholarly research, and teaching and research tools to help humanities researchers work with digital text and images. ARC coordinates the resources of these research environments.
The Advanced Research Consortium (ARC)
Building Bridges between Academic Networks

Elizabeth Grumbach [@EMGrumbach], ARC Project Manager
Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University

Data Aggregation
- ARC
- RDF
- Solr
- Index

Peer Review for digital projects
- Archives submit for peer review
- 18th NINES
- P&T Letter
- Metadata Ingestion

Metadata Standards
- Dublin Core
- Custom <collex> elements
- Consultation & Standards Meetings

Software Sustainability
- COLLEX
  a COLlections & EXhibits builder
  bit.ly/COLLEX
- TypeWright
  crowd-sourced OCR correction
  bit.ly/typewright
- Juxta
  textual collation on the web
  bit.ly/jcommons
- Apache Solr
  customized for ARC
  bit.ly/arc-solr

http://emop.tamu.edu
Warnings
<table>
<thead>
<tr>
<th>Dec</th>
<th>Hx</th>
<th>Oct</th>
<th>Char</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>000</td>
<td>NUL</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>001</td>
<td>SOH</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>002</td>
<td>STX</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>003</td>
<td>ETX</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>004</td>
<td>EOT</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>005</td>
<td>ENQ</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>006</td>
<td>ACK</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>007</td>
<td>BEL</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>010</td>
<td>BS</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>011</td>
<td>TAB</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>012</td>
<td>LF</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>013</td>
<td>VT</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>014</td>
<td>FF</td>
</tr>
<tr>
<td>13</td>
<td>D</td>
<td>015</td>
<td>CR</td>
</tr>
<tr>
<td>14</td>
<td>E</td>
<td>016</td>
<td>SO</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>017</td>
<td>SI</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>020</td>
<td>DLE</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>021</td>
<td>DC1</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>022</td>
<td>DC2</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>023</td>
<td>DC3</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>024</td>
<td>DC4</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
<td>025</td>
<td>NAK</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>026</td>
<td>SYN</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
<td>027</td>
<td>ETB</td>
</tr>
<tr>
<td>24</td>
<td>18</td>
<td>030</td>
<td>CAN</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
<td>031</td>
<td>EM</td>
</tr>
<tr>
<td>26</td>
<td>1A</td>
<td>032</td>
<td>SUB</td>
</tr>
<tr>
<td>27</td>
<td>1B</td>
<td>033</td>
<td>ESC</td>
</tr>
<tr>
<td>28</td>
<td>1C</td>
<td>034</td>
<td>FS</td>
</tr>
<tr>
<td>29</td>
<td>1D</td>
<td>035</td>
<td>GS</td>
</tr>
<tr>
<td>30</td>
<td>1E</td>
<td>036</td>
<td>RS</td>
</tr>
<tr>
<td>31</td>
<td>1F</td>
<td>037</td>
<td>US</td>
</tr>
</tbody>
</table>

Source: www.LookupTables.com
ASCII Decimal

\&#67;\&#65;\&#75;\&#69
Binary Data
01000011 01000001 01001011 01000101
“As computational data structures, images differ radically and fundamentally from electronic text... [A]ll raster-based image formats—GIF, TIFF, JPEG, and BMP, to name some of the most common—are stored... as a matrix or grid in which each pixel is assigned a color value...,” as “a table of these pixel values.”

It is for this reason that “images remain largely opaque to the algorithmic eyes of the machine.”

THE

First volume of the Chronicles of England, Scotland, and Irelande.

Conteyning,
The description and Chronicles of England from the first inhabiting unto the conquest.
The description and Chronicles of Scotland, from the first original of the Scottes nation, until the yeare of our Lorde 1571.
The description and Chronicles of Irelande, likewise from the first original of that nation, until the yeare 1547.

Faithfully gathered and set forth, by Raphael Holinshed.

At London, Imprinted for John Harrison.
“Dirty OCR”
A Detection
and Querimonie of the daily
eveniments and abuses
committt in physick,
Concerning the
two parts ther-
of: that is,
The Physlions part, The part of
the Surgeons, and the arte of
Poctlaries.Dedicated untv
the two most famous
universties Oxford
and Cambridge.

by John Securis
physitian.

Non serit hac medica præstantes arte Securis,
Nec medici officio qui bene functus erit,
Non serit insignes chirurgos nec myropolas,
Ars quibus & pieter sunt bene imitata simul.

1566.
EEBO – 12 million page images

Eighteenth Century Collections Online (ECCO)

"ECCO is an amazingly rich resource: it puts a magnificent library of eighteenth-century printed material on the desktops of scholars and students. It vastly improves access; it allows the scholar to discover new seams of material; it gives students unprecedented access to masses of primary source material. And of course it supports new kinds of searching. It is very well presented and in general easy to use."

Joanna Innes, Somerville College, Oxford University

Read more testimonials here>>

Latest News

A new national license agreement for French institutions>>

Eighteenth Century Collections Online (ECCO) brings learning to life and new dimension to history at Leeds University >>
ECCO – 33 million page images
Scanned $\leftrightarrow$ 1970s

Microfilmed $\leftrightarrow$ 1980s

Digitized $\leftrightarrow$ 1990s
The text on the image is not legible due to the quality of the scan. It appears to be a page with text, but the content cannot be accurately transcribed.
Early Modern Printing

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

Days.

at Waltham, count. Waltham.

at Earls, count. Earls.

at Downton, count. Downton.

at St. Albans, count. St. Albans.

at St. Anne's, count. St. Anne's.

at Shrewsbury, count. Shrewsbury.


at Exeter, count. Exeter.

at Plymouth, count. Plymouth.

at Dartmouth, count. Dartmouth.

at Truro, count. Truro.

at Newbury, count. Newbury.

at Weymouth, count. Weymouth.

at Salisbury, count. Salisbury.

at Winchester, count. Winchester.

at Southampton, count. Southampton.

at Chichester, count. Chichester.

at Portsmouth, count. Portsmouth.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Bristol, count. Bristol.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestoft, count. Lowestoft.


at Harwich, count. Harwich.

at Lowestoft, count. Lowestoft.

at Yarmouth, count. Yarmouth.

at Lowestof...
Image Quality

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

Aletheia

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

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

http://emop.tamu.edu
Brazos High Performance Computing Cluster

CPU Hours: Total: by Group

1. IDHMC
2. HEPX
3. MATH
4. IAMCS
5. PETE
6. General
7. College of Agriculture
8. Cluster Admins
9. CMS

CPU Hour

[Bar chart showing CPU hours for each group]
!! results !!

27 pages per second
15 million pages done

On ECCO: 86% Correct
This will be the first time that EEBO documents are available as searchable text, and that they can be corrected by scholars for use and/or publication.

TypeWright

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

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

lity in so easy and agreeable a manner, as
shall render them equally delightful and

9 I F to inculcate Religion and Mor-
10 lity info easy and agreeable a manner, as
11 shall render them equally delightful and

In this document the original spelling should be retained.

Instructions:
Use arrow buttons above-right or click on the thumbnail to jump around the page.

- If a word or portion of a word is illegible, type "8" in its place; please do not make any guesses about what a word might be.
- Copy original spelling and punctuation, typing what you see on the page, except in the case of the long 's' use 's' and not 'f' when 's' is called for.
- Include end-of-line hyphens, preserving the syllables as they occur on each line.
Gale and 18thConnect Partner to Improve Access to Eighteenth Century Documents

By Dana Wheeles on November 5, 2010

[18thConnect's work with Gale/Cengage has been publicized by a press release, copied in this message. Users of 18thConnect already have access to citations from Gale's ECCO catalog here.]

Farmington Hills, Mich., Nov. 4, 2010 — Gale, part of Cengage Learning, and 18thConnect, a scholarly organization dedicated to forging links between eighteenth-century archives and today's digital research environment, today announced a partnership to share scholarly content and improve the searchability of documents within Gale's Eighteenth Century Collections Online (ECCO) archive.

"Gale's partnership with 18thConnect gives us a unique opportunity to collaborate with a leading scholarly organization in order to improve upon the user experience within ECCO, the leading database for research and teaching of the eighteenth century," said Jim Draper, Vice President and Publisher, Gale.

Gale's ECCO archive, one of the largest academic research collections of its kind, contains more than 180,000 key English and foreign language titles published primarily in the United Kingdom. Despite Gale's use of the best in Optical Character Recognition (OCR) technology, eighteenth-century typefaces can still be challenging to capture with perfect accuracy, which may impact results when searching or data-mining.

Recently, 18thConnect was awarded National Endowment for the Humanities (NEH) sponsored supercomputer time to re-run page images from the ECCO archive through an open-source OCR program that will generate cleaner texts. This improved OCR-created text will be incorporated into ECCO, resulting in improved searching within the resource. In addition, registered 18thConnect users will then have the opportunity to review the improved texts and correct them using a tool housed on the 18thConnect website. The correction tool will be built thanks to a grant awarded to Miami University of Ohio from the Mellon Foundation. Using this crowd-sourced correction tool, users can further correct issues not caught by the OCR process, and in exchange they will have the option to download the amended books in dedicated editions. 18thConnect will provide a limited number of copies of these new editions for distribution to libraries, making it easier for scholars and students to access the original texts.
TEXT AND META DATA SHARING AGREEMENT

THIS TEXT AND METADATA SHARING AGREEMENT (this “Agreement”) is entered into by and between PROQUEST LLC, a Delaware Limited Liability Company (“PROQUEST”), with offices at 789 e. Eisenhower Parkway, Ann Arbor, MI 48106-1346 and Networked Interface for Nineteenth-century Electronic Scholarship, a not-for-profit endeavor from Texas A&M University (“ARC” representing “REKN/18THCONNECT”) as of the Contract Date specified below (the “Effective Date”).

WHEREAS, PROQUEST’S mission is to create indispensable research solutions that connect researchers with information and empower them to use, create and share content, including but not limited to its EEBO Database which includes page images of works originating in 14th to 17th century; and

WHEREAS, REKN/18THCONNECT seeks to enable REKN and 18thConnect users to explore 14th to 17th-century works currently available in digital form by providing scholars with the ability to conduct research within federated digital environments, as well as to produce personal annotated collections and exhibits using an integrated tool.

NOW, THEREFORE, in consideration of the premises and the mutual promises contained herein and other good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged, the parties agree as follows:

1. DEFINITIONS
Online Humanities Scholarship, The Shape of Things to Come is a three day conference to explore how to develop and sustain online humanities research and publication. Nine scholarly papers and eighteen responses will leverage discussion by a broad group of persons invited to the conference to contribute their expertise. This group includes scholars working on other projects and persons from funding agencies, publishers, museums, libraries, and professional organizations. The conference is closed to this group in order to provide maximum focus to the discussions.

The 'content focus' is a set of editorial projects that implement digital technology in different scholarly disciplines: specifically literary and cultural studies, history, classics, archaeology, and music. While the technical features of these projects will inevitably come into discussion, the social and institutional issues facing all scholarly ventures of these kinds are the imperative concern: how to sustain the life of such projects; how to address their institutional obstacles and financial demands; how to involve the greater community of students and scholars in online research and publication; how to integrate these resources with our inherited material and paper-based depositories; how to promote institutional collaborations to support these scholarly efforts; how to link online resources, which are now largely dispersed and isolated, into a connected network. The sustainability and institutional problems should be considered the two overriding topics.
Digital Editions