**EB Project Log**

7-20-16

* Gary: Finished creating single .rtf file of eb03 output.
  + Encountered major problem when trying to combine the odd/even pages together. AFR loses track of the text information associated with each image.
  + **For next time, keep odd/even combined.**
* Minju: Finished OCR for eb09 and created single .rtf file.
* Next: Once we have single r.rtf file for each 250-page section of the four editions, then identify the best method of segmenting them by article.
* Set up filenaming convention for article files, using *edition\_segment\_article* format.

7-14-16

* Meeting with Gary and Minju to set priorities, following completion of OCR stage
  + Integrate odd/even files and output text as single .rtf file, one for each of the four editions. Convert .rtf to a second .txt version.
    - Gary to experiment with merging odd/even within AFR
    - Still need to decide how to handle the footnotes—*I think we should just exclude them, at this point.*
  + Divide the text into articles.
    - Minju to look at editions to see what identifying traits will work, using R on plain text. It may be preferable to divide it while it’s still in .rtf.
  + Tag the articles with subject headings
    - Use Universal Decimal Classification, [www.udc.org](http://www.udc.org) .
    - Need to decide how fine-grained to go, ie, top level + 1? +2?
    - Tags need to be stored in a spreadsheet.
  + Chunk article files into ideal lengths (500 words?).
* Looked at **plain text files and metadata**. They can’t store it. All of their file properties are generated by the file system itself and stored separately from the .txt file.
* Looked at **Mallet and file formats**—it transforms all files into plain text, but .txt and .csv are the best options.
* Looked at **R and file formats**—it will scan anything, but it doesn’t recognize file formats and incorporates all of the formatting and hidden coding as text, even for .rtf, making it unusable. Or it finds null characters where they aren’t supposed to be and doesn’t recognize much of the file information.

4-29-16

* Identifont ([www.identifont.com](http://www.identifont.com)) identifies eb03 typeface as Caslon. A manual check shows that it is Caslon, and not Fell. (Fell has an angled top to the A, rather than a flat one, and the front foot of the R in Fell tapers, while the font in my files is flat with a forward serif, like Caslon.)

4-18-16

* Began comprehensive mapping of eb editions in HT, preparatory to applying for their ACS award.
* Hired Gary Scales and Minju Bae for summer work, June 1-Aug 23.
* Met with Yuhong Guo in Computer Science (on 4-13-16), who confirmed that topic modelling was the best method for identifying the migration of topics between fields.
* Trouble trying to create HT workset for eb03-gri-gri. It doesn’t show up in search mechanism, and downloading the metadata leads to volume identifiers that the “upload workset” command can’t recognize. I tried the same procedure to upload a workset for eb03-gb-um, and it worked flawlessly.

3-25-16

* Lecture on eb project at U Florida.

3-20-16

* Txt files for eb03, 7, 9 pilot slice have reversed the right and left columns, beginning with the right instead of the left. Not sure even how this was possible, but they need to be redone. Joyce has the AFR files on the 1TB HD, at least. Eb11 is okay.

3-19-16

* Began AFR tests on eb03 for higher accuracy, including limiting loaded typefaces to Fell, Caslon, Baskerville, Garamond. Had to dl Fell. Can’t find the other common font, Fleischmann. Accuracy seems better, but it needs training, and the ligatures are not included in the current language file, so need to be added.
* Eb03 pilot slice: search and replace to correct common errors. These are very dirty, though, and unreliable, so don’t do anything further on this set.
  + Eb03 pilot slice: 24867 / 281176 \* 100 = 8.84

3-18-16

* Saved AntConc output for eb03 pilot as XL file and indicated word corrections for the long “s” problem.
* Used AntConc to calculate TTR:
  + eb03 pilot slice: 25158 / 282391 \* 100 = 8.91 (uncorrected files with long “s” problem)
  + eb07 pilot slice: 20329 / 316427 \* 100 = 6.42
  + eb09 pilot slice: 25993 / 334792 \* 100 = 7.76
  + eb11 pilot slice: 26710 / 344085 \*100 = 7.76

1/21/16

* Confirmed that all files on the 1TB HD have been moved to the 3TB HD. I removed them from the 1TB HD and moved only the files needed for OCR onto it, so I can give it to Joyce to work with.
* How many pages does “R” amount to in all four editions? How many words/page?

1-19-16

* Copied all files from entries\_r/eb09\_r to EB\_PROJECT/working, for renaming.
* files for eb09 are consistently numbered even/odd for left/right pages.
* Page numbers differ from file numbers. The volume includes full-page plates, with blank pages on the reverse, and these are unnumbered in the pagination while numbered in the filename, of course, so the fn and pn cannot be reconciled easily. Not necessary, either, so leave that alone in the renaming.
* File renaming: eb+*edition*+*volume*\_*repo+digitizer+original*-*derivative*\_*nnnn*
  + eb0920\_iacdluc-dv\_0104
  + for 9th ed v.20\_internet archive California digital archive-derivative\_file number

1-13-16

* Added eb07 originals to EB\_PROJECT/archive
* Added letter-R image files for eb07 to EB\_PROJECT/entries\_r
* Need to finalize naming convention for files.
  + Original files: preserve original filename as downloaded.
  + Derivative volume files: edition\_volume\_source, ie:
    - eb07\_05\_ht
    - eb03\_11\_ia
  + Derivative page files: TO DO

12-19-15

* Organize original image files into storage archive on 3TB HD, named EB\_PROJECT. Missing eb07 files—they’re on the 1TB LOGAN HD in office.
* Created new folder in EB\_PROJECT HD to hold copies of all of the letter-R image files, “entries\_r.” All are there except eb07.

11-17-15

* Next: calculate words-per-page averages for all three editions, and then convert total pages to total words.
* Finished adding eb11 to map\_by\_letter.xlsx
  + Eb11 integrates plates with articles, unlike eb03 and eb07, so that increases article page lengths, separate from word counts.

11-13-15

* Finished adding eb07 to map\_by\_letter.xlsx

11-12-15

* Started adding eb07 to map\_by\_letter.xlsx
* Discovered that eb07\_3, v1, is a series of 4 “preliminary dissertations” on the history of philosophy (by Dugald Stewart) and math/science. How do I integrate these into calculations??

11-11-15

* Created XL chart with all alpha-segment lengths for eb03, map\_by\_letter.xlsx.
* Mapped alphabet-segment lengths in eb03. Average is 600 pp. The closest to average are
  + **g** 540
  + **h** 591
  + **r** 607
  + **e** 654
  + **l** 662

11-10-15

* Kaelin assigned to eb07 downloads.
* Checked entries for eb03, ”E”. Compare with those for “G” to decide which to use. Both are 2% of all words in English, but *a better comparison would be page numbers*.
* For eb07, comparing v11, ht version is cleaner scan than gb, so ask Kaelin to download all of the hathi files (v1-9, missing v10, v11 done, v12-21).
* Also need to download all of the eb11 images from ia-BeLL Reference.
* For eb03, gb contains OCR, but the ia-getty files do not.

11-6-15

* Next—install ABBY Fine Reader. It’s $110 for educator, $70 as upgrade, and I think my little freebie program qualifies for that.
* Tried “FREE PDF TO WORD DOC.” Rudimentary but worked.
* These PDFs do not have text—they haven’t been OCR’d.
* Tried CNET, but could not install DEBENU PDF TOOLS—system thinks it is malware.
* Tried importing into Word. It does it, but only imports the images of pages, not the OCR text.
* Tried exporting text using Adobe Reader. Requires an Adobe subscription to use.
* Tried extracting text using Adobe Reader. Took 5 minutes and produced a 1k file. Useless.
* Looked up BeLL Reference. Basic e-Learning Library, in Cambridge, generating free online resources for basic learning in education. They won’t be doing anymore eb’s, IMO—just the 11th, which I have.
* Eb07 notes
  + Downloaded vol 3 from ht to check file size, compared to gb

11-4-15

* Eb07 notes
  + Created workset at HT, with all vols except 10 and 22
  + downloaded vol 10, 16, 17 from gb

10-28-15

* Need to create a timeline—monthly project goals, in order to have enough data and analysis to use in an NEH Start-Up Grant in September.
* Focus on eb07 for 3rd edition …

10-22-15

* Eb-sup 4, 5, 6 (6 vols)– excellent scans from getty research dl’d from ia. But these supplements are treatises, out of alphabetical order. That’s a problem for working with a slice defined by alphabet (like all ‘G). Better solution is to look for the best scans of 7th edition. So far, I’ve only been able to locate GB scans.
* Eb03 is all done and high quality scans are dl by Kaelin
* Eb11 – dl good scans from Bell Reference; still need Kaelin to dl the originals.

9-9-15

* Since eb02 is not fully digitized I’ll have to try eb03. That’s 18v+2 supp, so it’s the first edition to approximate the size of the c19 editions.
* eb03: 1788-97, 18v + 2v supp 1801, rev 1803.
* **3 editions for initial trials: eb03 + supp (1801), 7 (1842), and 11 (1911)**. Those are all major editions, so good choice to work with.

9-8-15

* EB02 problem: only vols 5, 8, 9, 10 are digitized in Google books.
* Nothing in Internet Archive (checked both spellings of “ency”)
* HathiTrust has only the same 5 vols of EB02 (“Encyclopaedia” search needed to find them)

9-2-15

* DL eb02-5, eb02-08 from GB
* Downloaded eb02-9 and eb02-10, one from HT, the other from GB. There’s no difference in image quality—the HT version is from GB but takes 10 minutes to assemble and download. Direct from GB is much quicker.
* EDITIONS: instead of eb01 (1771), use eb02 (1784). Eb01 is 3 vols, compared with 10v for eb02, and that will give me a much better base for tracking changes. That still gives me a 127-year span to eb11, so:
  + **3 editions for initial trials: eb02 (1784), eb07 (1842), and eb11 (1911)**.
* Reviewed everything. First project milestone: get full texts for the 3 editions I want to run tests on

6-10-15

* NEXT: see about running tests within Hathi
* Downloaded Eb03, v7 abd 8 (G entries) from HT.
* Created collection for Eb03 in HT.
* Browsed IA more and found other editions, noted in spreadsheet.
* Did wordcounts for “G” in eb01, 7, 11.
* Downloaded 1st-letter word frequency chart.
* Eb07 (1842) “History” entry uses Genesis as the only true history, and dates the world’s history at 6000 years. Evidence of Egyptian and Chinese antiquity is dismissed as “absurd”.
* Eb11 (1911) “History” is entirely dedicated to historiography itself—the “history of history.”

6-2-15

* Browsed IA—lots of editions, some digitized by UC instead of Google—need to look more for my key editions.
* Initiated Torrent download of eb11 from Internet Archive in .djvu format.
* Download eb07, v10-1, v10-2 (gb); v11, v22 (ht)
* Create collection of complete eb07 in Hathi
* Revise prototype plan: use 3 eds for initial trials, 1 (1771), 7 (1842), 11 (1911). Those 3 are all 70 years apart.
* Updated EB\_editions.xlsx with more information on first 3 eds from EB’s history article. Need to continue with eb04-12.
* Experiments with extracting text from pdf; quality is very poor for eb01; see notes.

5-28-15

* eb01—full text downloaded from Google Books for all 3 vols in pdf.
* eb01.v2, letter “G” only—extracted from pdf in txt and dox formats.

5-27-15

* Eb11, letter “G” only—full text downloaded from PG in txt and epub formats

2-23-15

* Download eb03 v1-3 in pdf from Hathi.