**Introduction**

The field of Digital Humanities offers interesting new scholarly and pedagogical possibilities for literary studies as a whole and early modern and eighteenth-century studies in particular. The Early Modern OCR Project (eMOP) at the Initiative for Digital Humanities, Media, and Culture at Texas A&M seeks to improve the usability of texts available through the EEBO and ECCO databases. As part of this project, four undergraduate students with little to no advanced technical knowledge are taught to train OCR engines to read and recognize fonts in text images from the EEBO and ECCO collection. This hands-on project has served as a means of introducing students to texts, fonts, and unique print characters of the period.

**Step One: Naming Font Sets**

Page images are named according to publisher, publication year, and the 20-line height. For example, a text image from the 1702 *Anno Regni* text would be named cbil1702_116.

**Step Two: Page Segmentation and Analysis**

Layout regions, lines, words, and individual glyphs in a text are identified and defined using Aletheia Desktop, a tool developed by PRImA Lab at the University of Salford that performs page segmentation and text recognition, among other tasks.

**Step Three: Text Correction**

Aletheia analyzes the page image and generates text for each character. Students are then responsible for fixing any “misreads” by typing in the correct text in the Text Content Box for each corresponding character—paying special attention to unique characters such as long S’s, ligatures, italics, rotunda R’s, suspension marks, and printer’s marks.

**Step Four: OCR Training**

10-15 page images are processed for each font type and comprise a training set. A Tesseract training tool uses the font set to train the OCR engine to read and recognize that particular font.