WEEK 4b: DIGITAL ARCHIVES, GALE PRIMARY SOURCES AND THE DIGITAL SCHOLAR LAB

> The first step of the OCR software is to analyse the structure of the newspaper page. It divides the page into elements such as blocks of texts (columns), tables, images, etc. The lines are divided into words and then into characters. Once the characters have been singled out, the program compares them with a set of pattern images stored in its database. It analyzes the stroke edge, the line of discontinuity between the text characters, and the background. Allowing for irregularities of printed ink on paper, each algorithm averages the light and dark along the side of a stroke, and advances numerous hypotheses about what this character is. Finally, the software makes a best guess decision on the character. This character is given a confidence rating…A secondary level analysis may then take place at word level (since now a word is formed). The built-in English dictionaries and possibly dictionaries of other languages are checked to see if the word matches.
>
> Rose Holley, "How Good Can It Get?" (2009)

Creating Digital Archives

This OCR section will give you an understanding of HOW the text you are working with came to be created, where it came from and what factors influence its quality. As digital humanists, it is crucial to develop this awareness in order to make informed decisions about your data and its limitations.

Curious to try working with OCR text for yourself? ABBYY offers a 30-day free trial if you want to test the process.

Transkribus is an HTR (Handwritten Text Recognition) platform which is also worth exploring.

READING

Begin by reading the following articles:

How does OCR document scanning work? https://www.explainthatstuff.com/how-ocr-works.html

Rose Holley, 'How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs', *DLib Magazine* 15, n. 3/4, March/April 2009 http://www.dlib.org/dlib/march09/holley/03holley.html

What about OCR accuracy? https://www.hsassocs.com/what-is-ocr-accuracy/

Strange, Carolyn, et al. "Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers ." Digital Humanities Quarterly, vol. 008, no. 1, Apr. 2014.

Abstract*: "Digital humanities research that requires the digitization of medium-scale, project-specific texts confronts a significant methodological and practical question: is labour-intensive cleaning of the Optical Character Recognition (OCR) output necessary to produce robust results through text mining analysis? This paper traces the steps taken in a collaborative research project that aimed to analyze newspaper coverage of a high-profile murder trial, which occurred in New York City in 1873. A corpus of approximately one-half million words was produced by converting original print sources and image files into digital texts, which produced a substantial rate of OCR-generated errors. We then corrected the scans and added document-level genre metadata. This allowed us to evaluate the impact of our quality upgrade procedures when we tested for possible differences in word usage across two key phases in the trial's coverage using log likelihood ratio [Dunning 1993]. The same tests were run on each dataset – the original OCR scans, a subset of OCR scans selected through the addition of genre metadata, and the metadata-enhanced scans corrected to 98% accuracy. Our results revealed that error correction is desirable but not essential. However, metadata to distinguish between different genres of trial coverage, obtained during the correction process, had a substantial impact. This was true both when investigating all words and when testing for a subset of "judgment words" we created to explore the murder's emotive elements and its moral implications. Deeper analysis of this case, and others like it, will require more sophisticated text mining techniques to disambiguate word sense and context, which may be more sensitive to OCR-induced errors."*

LECTURE

https://youtu.be/TOCazHs5wEE

This lecture was recorded for the Fall 2018 'Introduction to DH' class. Ray Bankoski and Michelle Fappiano from Gale joined the class via Zoom to discuss 'Making an Archive'. In it, you'll learn about the practical and business considerations that inform the selection and creation of the archives you're using in the Digital Scholar Lab. There's also an overview of what OCR and HTR are, and the limitations of both.

IN CLASS

We'll begin working with the Digital Scholar Lab in class, starting with a demo, then you'll have an opportunity for hands-on work as you begin to gather material into content sets to investigate for your research project. You'll have an opportunity to look at some of the OCR output in the DSL in detail.