Digitization

Workset 3: Digitization

Goals

1. Improve your understanding of gaps in the digitized record by finding some;
2. Gain firsthand experience with digitization practices by bringing a text through from physical to digital form.
3. Spend some time staring at in-the-wild OCR.

Part 1: Inventorying non-digitized sources

1. Reflecting on what we've read, and what you know from past experience, but *without doing any Internet searches*, write down four *printed texts* that you think may have escaped the dragnet of digitization.[1] At least two should be a book; I'll be happy if all four are, but you can also choose magazine articles, broadsides, newspaper editions, and so forth.

2. Check (figuratively and literally) whether your three texts exist in digitized, analogue, or physical forms in the following places. Put, let's say, "A" for analogue, "D" for digital, and "N" for neither.

Library	Source 1	Source 2	Source 3	Source 4
Northeastern				
Boston Public				
Any library in eastern MA				
Any library in Worldcat				
Hathi Trust				
Google Books				
Internet Archive				
A domain-appropriate engine				

3. Post a citation of some sort (don't worry about formatting) for the texts that you are confident don't exist online, but which *do* exist in a Boston-area library or of which you have a physical copy. In a sentence, describe in the *general type* of the thing you chose. E.g., if you found that the Boston Globe magazine doesn't exist, say "I figured the Boston Globe Magazine from the 80s wouldn't be digitized because typically microfilm only stores the newsprint part of the paper; this is probably true for most newspaper magazines. There is a copy at the BPL."

[1] By printed texts, I mean, affixed to paper by a printing press of some sort. By "escaped the dragnet," I mean: physical copies exist in the real world, but there aren't copies in "the cyber," whether downloadable or not. If they are online but not *accessible*, they *have* been digitized. I think I'm also OK with typewritten things that are mimeographed, but don't go nuts with these.

One thing to keep in mind here is that the Internet is absolutely the worst way to search for books that haven't been digitized.

Instead, you may do well to find a room of books somewhere that seems likely not to have been scanned. If you're totally stumped, ask around about what other people are finding. There are a lot of different strategies for this.

In my house, for example, I've turned up the following books that don't seem to have been digitized (I haven't done a totally exhaustive search yet):

1. A 1904 "Fullstandig Varlds-Atlas" published by the Svenska Amerikanska Post in Minneapolis; probably something given away to customers by an ethnic newspaper and so not collected by libraries.
2. The Telberg Book company's 1956 English translation to "Atlas Mira"; a mimeographed set of pages bound weirdly together for a small print run that would live in the atlas case at just a few libraries that owned the Soviet atlas and wanted a translation.
3. When I Wished I was Here, 2007, by Derek McGee – a short book/pamphlet published by a kind of hipster press in Brooklyn.
4. Is it Time for the Party? By Julia Summers. Hallmark Children's Editions, c. 1970. Out-of-print children's book with a plastic clock built into it.

–

Part 2: Digitization

3. Take responsibility for one of the texts identified in step 3. You have first dibs on your own. But: identify something **not in copyright**. No academic books from the last 30 years, say.

4. Scan it to an image format. If it's a whole book, you can do just a chapter or two, or split up with someone else. This step will be much easier if you work with others. (Only two people should need to go to the BPL, or Tufts, or wherever.) I'm happy for any three people to share a single book.

5. Perform OCR on the image. You can use Adobe Acrobat, Bill Turkel's instructions, or some instructions for OS X from a previous time this class was offered.. The easiest way for short sources like this is to use Google Docs, which allows you to upload a PDF and get OCR out.

6. Export the OCR'ed text as a `.txt` format; post it to your blog with a brief bibliographic preamble. (This doesn't count towards the blog post quota, it's just a place to store it).

7. Clean it up until you've fixed 50 mistakes in the main text or the whole document is spotless, whichever comes first. If you want to think about *regular expressions* as Bill Turkel does as a way to programatically fix problems, you could instead compile a list of some general changes.