CODERS' CORNER

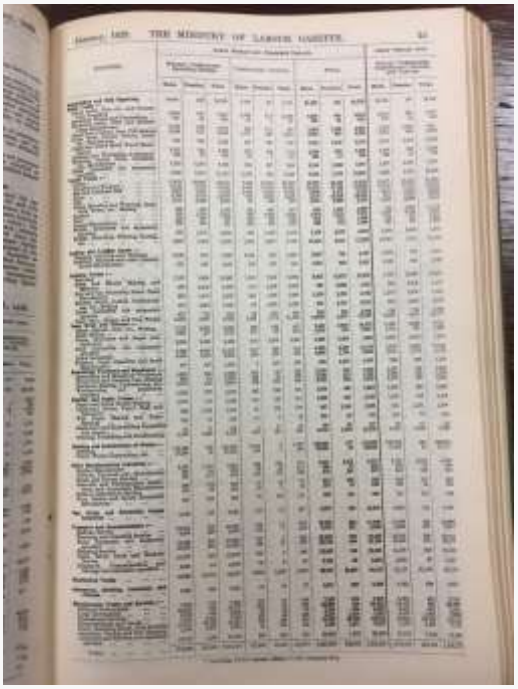
csae
CENTRE FOR THE STUDY OF AFRICAN ECONOMIES

HOW TO DIGITIZE A DATASET USING YOUR PHONE

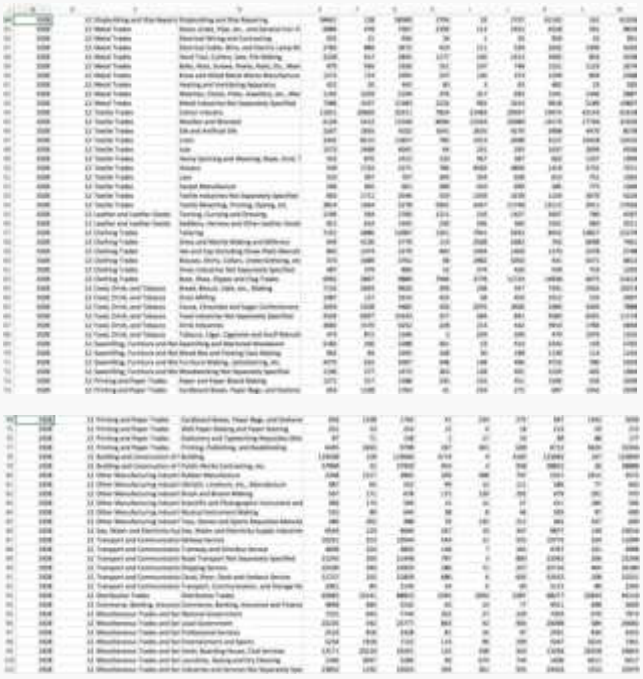
This summer and fall, I decided that I wanted to digitize 600+ tables from the 1923-1936 *Ministry of Labour Gazette,* a British government publication with detailed monthly data on unemployment in different industries during the interwar period.

While many, many academic papers have used aggregated or partial data based on this series, to digitize the entire dataset by hand would be an overwhelming (perhaps impossible) task. But there are some dimensions of interwar unemployment that we can only really understand with a complete series, including the different experiences of men and women under the unemployment insurance system, the scale and nature of short-time work, and the distinct patterns of unemployment across specific industries.

Motivated by these questions and my optimism about the power of new technology, I set out to find a method of digitizing this data that was time efficient, cheap, and accurate. Here's what I came up with!



Before (data in book)



After (data on computer)

WHAT YOU'LL NEED

1. DATA

This method will work best with data that:

- Is typewritten rather than handwritten. The technology for digitizing handwriting exists and is improving, but in my judgement is not reliable or accurate enough yet to use for research purposes.
- Has clear formatting like headlines and column separators. This allows the software to recognize the table structure better.
- Is in a large table. The real time savings come when you're working with tables with 50+ lines.
- Is repeated in a similar format. I had a two-page table for every month in the same format, meaning I could use the same system for checking the accuracy of the digitizing over and over.
- Added plus: This method works great with giant books that don't fit well on a traditional scanner!

2. OCR SOFTWARE - ABBYY FINESCANNER IPHONE APP

OCR stands for "optical character recognition," where text can be extracted from an image that is scanned or uploaded somehow. For example, if you take a picture of a book and send it to your computer, you will need to use OCR software to make the text searchable or editable.

In the early 2000s, the primary way people used OCR was scanning images into a desktop computer program. Now, there are a plethora of OCR programs, apps, websites, and even pens that use a variety of different methods for different purposes. For example, both Google Drive and Dropbox have proprietary OCR models that can automatically recognize text in stored images or PDFs. A number of open-source models also exist which are easy to adjust to meet the needs of a specific project.

For this project I used the ABBYY FineScanner iPhone App (available here). ABBYY is the industry standard for commercial OCR for good reason — the results are accurate and table recognition is built-in. Though I was prepared to train my own OCR to recognize these tables, there is no need to reinvent the wheel when the technology is this advanced.

The Premium subscription for the FineScanner App, including all features, is only $9.99 for the first year. This is a huge savings on the ABBYY desktop program, which as of writing costs a whopping $119.99. More on how the app works below!

STEP 1: SCANNING YOUR DATA

OPEN THE FINESCANNER APP AND CREATE A NEW DOCUMENT

FineScanner organizes your images into "Documents" which makes it easy to scan things that are multiple pages. In my case, I made one document for each year of data I was scanning, which contained the tables for each month in that year.

To create a new document, click the red plus on the bottom of the screen.

PICK YOUR CAMERA SETTINGS

Across the top of the camera there are a number of options for your scanning. Here are the available options and why I chose NOT to use any of them:

- Autocapture: Automatically takes a photo once the edges of a document are recognized, without you having to click the photo button. This is finicky in my opinion, and I often needed more time to line up and flatten out what I was scanning.
- Flash: Self-explanatory, but you should be doing your scanning in a place with good light so that your photos are clearer.
- Book Scan: Scans the left and right page of an open book simultaneously, but doesn't work well with oversized books like what I was using
- Best of 3: This takes 3 photos and selects the best one automatically for you. I used this for about half of my scanning but it made the process go very slowly. I got equally good (if not better) results from just a single shot, which was also more efficient.

LINE UP YOUR DATA FOR THE SHOT

In this very DIY setup, I didn't have any fancy scanning equipment or book holders, so every page I scanned I needed to get flat on the table. By only scanning one page at a time (i.e. not over the binding of the book), I could put that half of the book flat on the table with the other side of the book held up at a 90 degree angle.

Flat on the table

TAKE THE PICTURE

Hold your phone flat above the data, lining it up as straight as possible with your data. Click the big button in the center to take the picture.

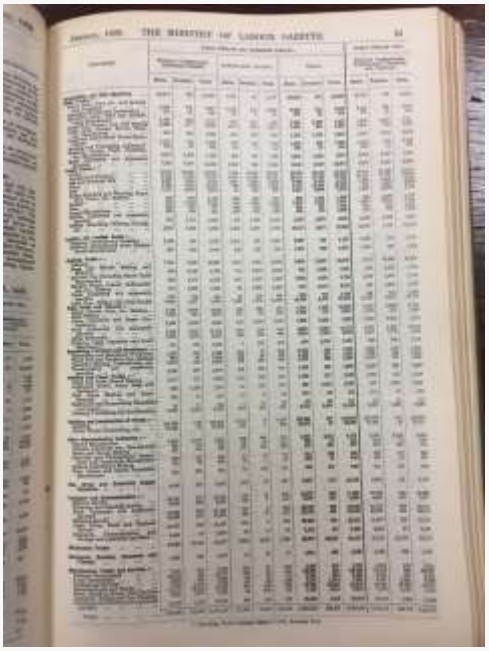
You may see a prompt asking if you want to save the image to your Photos. This saves a copy of the raw image into your iPhone's camera roll, so choose whatever is best for you. It's good to have this for backup if it will be difficult to reference the original data again. But if you're scanning hundreds of things, this does take up some significant room on your phone and might be overkill.

After your first picture, it will take you back to the camera. This is how you can take multiple photos at once. I wouldn't recommend taking more than 5 or so in one batch just because that makes it harder to root out and correct photography errors (see next step). But up to you!

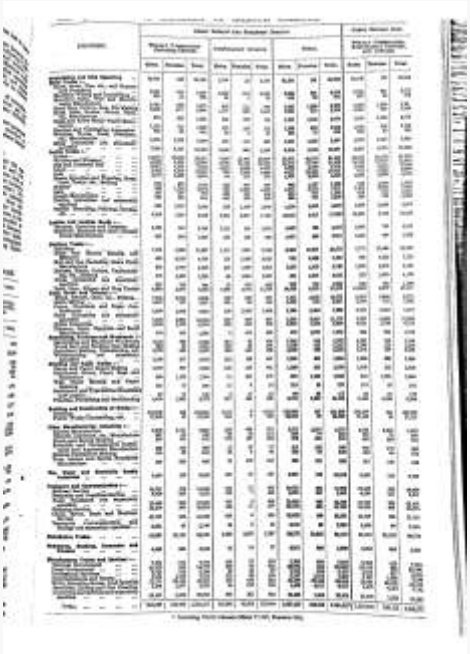
CHECK PHOTOS FOR QUALITY

On the bottom right you should see a thumbnail of your image with a little red number indicating how many photos you've taken. Click on this to check how your scanning went.

Each photo will have been edited by the app for optimal OCR. Typically the contrast is increased and borders may have been cropped out.



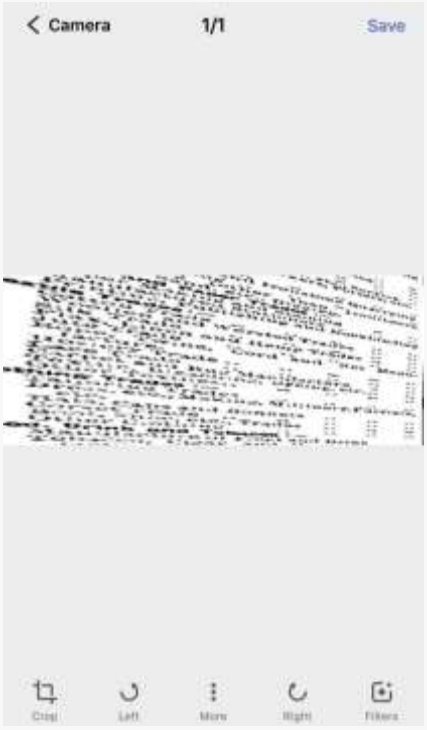
Original data in the book



A good scan of it

Occasionally the app will fail comically. In these cases, you'll want to retake the image. If you click the three vertical dots on the bottom of the screen, you'll be able to select "Retake" and see the camera again.

Oops



Oops again

SAVE

Click "Save" in the upper right corner of the screen. This will bring up a screen called Document Properties where you can enter a title for the document and optional tags.
Once you click "Done" (again in the upper right corner), the document will show up on the main app screen under My Documents. From here, you can click on it to add more images if you need to, or to edit again the document's properties.

REPEAT HUNDREDS OF TIMES

[Make sure to move onto the next steps with a small test case of your data to make sure this method works before spending hours scanning hundreds more pages!]

Listen to a podcast or some local indie music. This is the most time consuming part. Remind yourself that it could be worse!

STEP 2: RECOGNIZE/OCR THE SCANS

Now for the easy part, where ABBYY does all the work.

"RECOGNIZE" THE DOCUMENT

Simply press the purple "Recognize" button on a document from the My Documents screen. This will prompt you to pick the language of the document and the file type you want the text put into. ABBYY supports OCR for over 100 languages which gives you a lot of flexibility!

I OCR'd all of my PDFs into DOCX (Word) format because they had a mix of words and numbers, with tables that did not have many row lines. If you have a very clean or straightforward table you're scanning, feel free to experiment with scanning straight into XLSX (Excel) format.

When you click "Done" the app will upload your PDF to the ABBYY FineReader servers where their proprietary OCR model (usually costing hundreds of dollars) will be used on your PDF.

After recognizing your document, it will show up under My Documents as a DOCX.

SEND YOUR DOCX AND PDF TO YOUR COMPUTER

From here, it's easy to get the DOCX off your phone. Click the share button and you can upload it directly to the Cloud (e.g. Dropbox) or send via email. I would recommend sending both the DOCX and the Large PDF to your computer so that you can reference the original data when you get to data cleaning.

STEP 3: CLEAN AND COMPILE YOUR DATA

OPEN THE DOCX FILE ON YOUR COMPUTER AND EVALUATE

In the best case scenario, which happened for me around 75% of the time, the data was appropriately placed into a table in Word. In this picture, you can see that the numbers are neatly aligned into the table. When your output is like this, it's easy to highlight the relevant cells of the table and paste it into Excel. You'll also notice that the first column with the text has some issues. This doesn't matter in my case

because I could use the same industry list from month to month, just copying and pasting over into Excel the updated numbers.

In the medium case scenario, which happened for me around 20% of the time, the data was in long skinny text boxes rather than in a table. This was annoying but you could copy and paste the numbers over to Excel anyways to save on a little effort. Frequently, only a couple of columns would be like this and the rest would be in an efficient table.

Sometimes, for whatever reason, the OCR completely failed to read the numbers and I had to enter the data by hand into Excel. This was obviously the worst case scenario, and it happened around 5% of the time even though my PDF scans seemed neat and clear. Though this is not fun, had you been digitizing by hand this is what you'd be doing all the time!

DOCX Output

BRING DATA OVER TO EXCEL AND CLEAN

Before compiling your data into a final Excel sheet, you might want to clean the data to ensure it is accurate enough for your purposes. In my case, I put in many extra hours to make sure the data was absolutely perfect, using a special Excel sheet to clean and verify each table.

The structure of my data included a variety of "total" columns, making it relatively easier to clean and verify. Rather than check every number, I could replace the "total" columns with an Excel formula. You can see in this screen shot that the third column of numbers is a formula that has been filled down, summing the first two columns of numbers. If the formula threw #VALUE, then I checked by hand the two cells that went into it to make sure there were no accidental characters inserted by the OCR.

Then, I summed up each column and compared my results to the column totals given in the original data. If one of these numbers did not match exactly the original data totals, then I went through by hand and checked that the numbers in the column matched the relevant data. In most cases, developing this reusable spreadsheet reduced the amount of work I had to check by hand by around 80%.

Printing and Paper Trades	Cardboard Boxes, Paper Bags, and Stationery	1954	2,358	4,292	193	629	822	2,127	2987	5,114
Printing and Paper Trades	Wall Paper Making and Paper Staining	362	137	499	32	25	57	394	162	556
Printing and Paper Trades	Stationery and Typewriting Requisites (Not Paper)	327	254	581	14	87	101	341	341	682
Printing and Paper Trades	Printing, Publishing, and Bookbinding	18601	6,253	24,854	951	1113	2064	19,552	7366	26,918
Building and Construction of Works	Building	169556	353	169,909	5576	16	5592	175,132	389	175,501
Building and Construction of Works	Public Works Contracting, etc.	117309	37	117,346	2175	3	2178	119,484	40	119,524
Other Manufacturing Industries	Rubber Manufacture	5867	2005	7,872	733	460	1193	6,600	2465	9,065
Other Manufacturing Industries	Oilcloth, Linoleum, etc., Manufacture	1468	165	1,633	215	28	243	1,683	193	1,876
Other Manufacturing Industries	Brush and Broom Making	1,057	336	1,393	359	314	673	1,416	650	2,066
Other Manufacturing Industries	Scientific and Photographic Instrument and Appar	1673	397	2,070	228	111	339	1,901	508	2,409
Other Manufacturing Industries	Musical Instrument Making	4100	642	4,742	794	243	1037	4,894	885	5,779
Other Manufacturing Industries	Toys, Games and Sports Requisites Manufacture	819	553	1,372	254	290	544	1,073	843	1,916
Gas, Water and Electricity Supply Indu	Gas, Water and Electricity Supply Industries	17760	181	17,941	694	14	708	18,454	195	18,649
Transport and Communication	Railway Service	18572	259	18,831	795	6	801	19,367	265	19,632
Transport and Communication	Tramway and Omnibus Service	9089	457	9,546	514	12	526	9,603	469	10,072
Transport and Communication	Road Transport Not Separately Specified	40815	422	41,237	2219	14	2233	43,034	436	43,470
Transport and Communication	Shipping Service	50807	531	51,338	415	44	459	51,222	575	51,797
Transport and Communication	Canal, River, Dock and Harbour Service	52801	148	52,949	1728	38	1766	54,529	186	54,715
Transport and Communication	Transport, Communication, and Storage Not Sepa	3,228	148	3,376	88	4	92	3,316	152	3,468
Distributive Trades	Distributive Trades	164,712	49346	214,058	7519	3474	10993	172,231	52820	225,051
Commerce, Banking, Insurance and Fi	Commerce, Banking, Insurance and Finance	9,785	1587	11,372	140	18	158	9,925	1605	11,530
Miscellaneous Trades and Services	National Government	13,216	951	14,167	410	38	448	13,626	989	14,615
Miscellaneous Trades and Services	Local Government	59189	638	59,827	1549	44	1593	60,738	682	61,420
Miscellaneous Trades and Services	Professional Services	5,677	1837	7,514	244	73	317	5,921	1910	7,831
Miscellaneous Trades and Services	Entertainment and Sports	15587	4,667	20,254	457	235	692	16,044	4902	20,946
Miscellaneous Trades and Services	Hotel, Boarding House, Club Services	26770	28065	54,835	351	761	1112	27,121	28826	55,947
Miscellaneous Trades and Services	Laundries, Dyeing and Dry Cleaning	3,763	6741	10,504	166	1570	1736	3,929	8311	12,240
Miscellaneous Trades and Services	Industries and Services Not Separately Specified	47737	2363	50,100	1407	731	2138	49,144	3094	52,238
		1,756,106	273,079	2,029,185	361,199	107,669	468,868	2,117,305	380,748	2,498,053

For your own project, you'll have to devise a system that meets your needs, making it as reusable as possible. In my case, I used the same sheet over and over so that it would be already set up for each batch of new numbers to be cleaned. Feel free to experiment with your own improvements!

PUT DATA INTO FINAL EXCEL SHEET

The last step is the simplest of them all. Just copy the data from your cleaning worksheet and paste (as values) into your master data file!

ENDING THOUGHTS

This method definitely saved me time over digitizing all of this data by hand. However, it still took months to scan 600+ tables into the app and then clean it meticulously.

Why use this method over easier methods like hiring an RA or mailing the books to another country to have digitized? At least for this project, I really value the deep familiarity I developed with the data in the process of digitizing it. When you write a PhD, you're becoming an expert in something, and a key part of that is knowing the ins and outs of what you're working with. I also appreciated having control over every stage of the process so I could ensure it meets the highest standards of accuracy, which is not always needed but is essential to my current project.

Some other strengths of this method are that it was cheap ($9.99 + effort), I didn't need a fancy scanner, and I didn't have to take the original data sources out of the library.

Meredith Paker, DPhil Candidate in Economic History, Nuffield College, Oxford
Originally posted at meridithcpaker.com, re-posted with permission
06 May 2019