Digitizing Historical Balance Sheet Data: A Practitioner's Guide

Sergio Correia[a] Stephan Luck[b]

December 3, 2021

Methodological Advances in the Extraction and Analysis of Historical Data

[a] Federal Reserve Board, sergio.a.correia@frb.gov
[b] Federal Reserve Bank of New York, stephan.luck@ny.frb.org

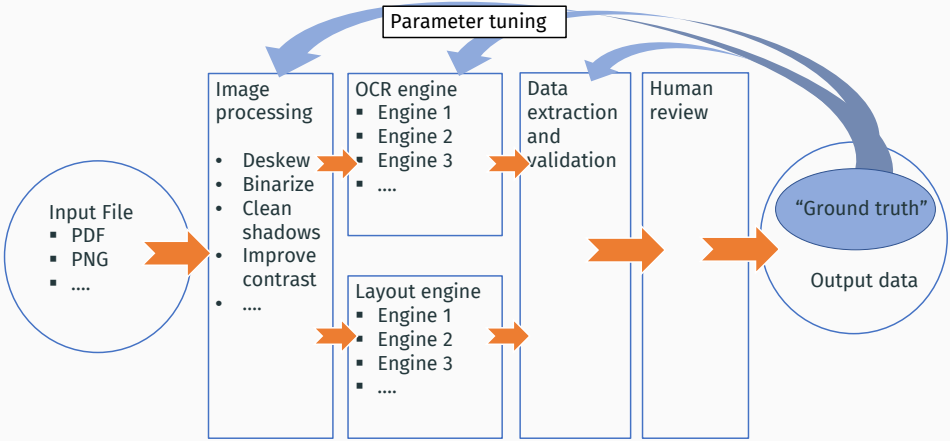
Goals of this talk (and paper)

- Share what we learned from transcribing structured historical data,
  - At *scale*
  - With limited resources (no funding, no RAs or interns)
- Show,
  - Our mistakes so you can avoid them
  - What we think works well (beyond just our projects)
- Make digitization projects more accessible:
  - Key principle: modular tools that can be mix-and-matched as needed
  - You don't have to reinvent the wheel, you *don't* have to be an expert coding
  - Open sourcing our tools: **https://github.com/sergiocorreia/quipucamayoc**

“A Practitioner’s Guide”

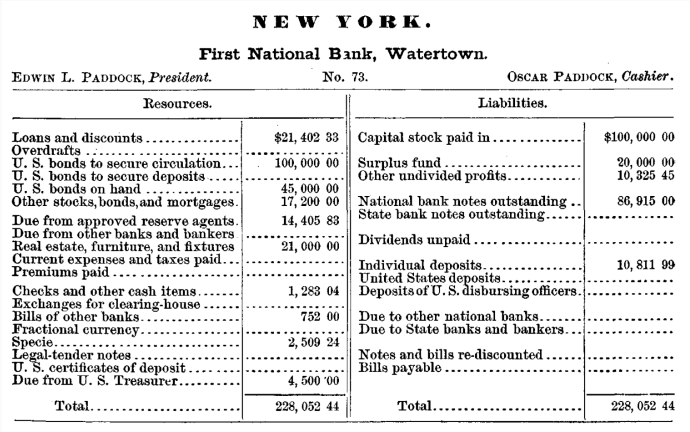
- Off-the-shelf OCR tools,
  - Have improved dramatically in recent years
  - Potentially very powerful
  - But still have high error rates, particularly for historical data
  - This makes large-scale efforts unfeasible
- Insight: we need to augment OCR with pre- and post-processing methods
- *The researcher must become the practitioner*
  - Understand their data
  - How to work around its limitations
  - How to take advantage of its characteristics

The data extraction pipeline



Example #1: OCC Annual Reports to Congress

- More than 100,000 balance sheets in *tabular form* (1867-1904)
- About 12,000 hours of work if typed by hand
- Used in Carlson et al. JPE, Forthcoming



NEW YORK.

First National Bank, Watertown.

EDWIN L. PADDOCK, *President.* No. 73. OSCAR PADDOCK, *Cashier.*

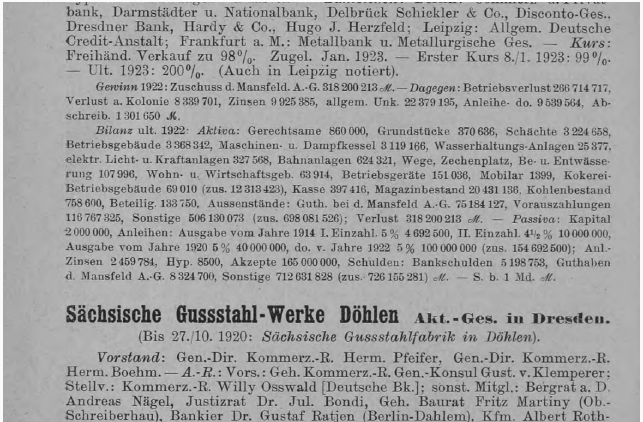
Resources.		Liabilities.	
Loans and discounts	$21, 402 33	Capital stock paid in	$100, 000 00
Overdrafts			
U. S. bonds to secure circulation	100, 000 00	Surplus fund	20, 000 00
U. S. bonds to secure deposits		Other undivided profits	10, 325 45
U. S. bonds on hand	45, 000 00		
Other stocks, bonds, and mortgages	17, 200 00	National bank notes outstanding	86, 915 00
Due from approved reserve agents	14, 405 83	State bank notes outstanding	
Due from other banks and bankers			
Real estate, furniture, and fixtures	21, 000 00	Dividends unpaid	
Current expenses and taxes paid			
Premiums paid		Individual deposits	10, 811 99
		United States deposits	
Checks and other cash items	1, 283 04	Deposits of U. S. disbursing officers	
Exchanges for clearing-house			
Bills of other banks	752 00	Due to other national banks	
Fractional currency		Due to State banks and bankers	
Specie	2, 509 24		
Legal-tender notes		Notes and bills re-discounted	
U. S. certificates of deposit		Bills payable	
Due from U. S. Treasurer	4, 500 00		
Total	228, 052 44	Total	228, 052 44

Example #2: *Saling's Börsen-Papiere*

- More than 30,000 balance sheets and income statements of German financial and non-financial firms (1915-1933)
- Used in ongoing research with Brunnermeier and Zimmermann

Example #2: *Saling's Börsen-Papiere*

- Data is in FREE FORM TEXT !!!
  - No predefined set of labels
  - Archaic abbreviations (in German)
  - Values twelve-digit long in hyperinflation years



bank, Darmstädter u. Nationalbank, Delbrück Schickler & Co., Disconto-Ges., Dresdner Bank, Hardy & Co., Hugo J. Herzfeld; Leipzig: Allgem. Deutsche Credit-Anstalt; Frankfurt a. M.: Metallbank u. Metallurgische Ges. — *Kurs:* Freihänd. Verkauf zu 98%. Zugel. Jan. 1923. — Erster Kurs 8./1. 1923: 99%. — Ult. 1923: 200%. (Auch in Leipzig notiert).

Gewinn 1922: Zuschuss d. Mansfeld. A.-G. 318 200 213 ℳ. — *Dagegen:* Betriebsverlust 236 714 717, Verlust a. Kolonie 8 339 701, Zinsen 9 925 385, allgem. Unk. 22 379 135, Anleihe- do. 9 539 564, Abschreib. 1 301 650 ℳ.

Bilanz ult. 1922: *Aktiva:* Gerechtsame 860 000, Grundstücke 370 636, Schächte 3 224 658, Betriebsgebäude 3 368 342, Maschinen- u. Dampfkessel 3 119 166, Wasserhaltungs-Anlagen 25 377, elektr. Licht- u. Kraftanlagen 327 568, Bahnanlagen 624 321, Wege, Zechenplatz, Be- u. Entwässerung 107 996, Wohn- u. Wirtschaftsgeb. 63 914, Betriebsgeräte 151 036, Mobilar 1399, Kokerei-Betriebsgebäude 69 010 (zus. 12 313 423), Kasse 397 416, Magazinbestand 20 431 136, Kohlenbestand 758 600, Beteilig. 133 750, Aussenstände: Guth. bei d. Mansfeld A.-G. 75 184 127, Vorauszahlungen 116 767 325, Sonstige 506 130 073 (zus. 698 081 526); Verlust 318 200 213 ℳ. — *Passiva:* Kapital 2 000 000, Anleihen: Ausgabe vom Jahre 1914 I. Einzahl. 5% 4 692 500, II. Einzahl. 4½% 10 000 000, Ausgabe vom Jahre 1920 5% 40 000 000, do. v. Jahre 1922 5% 100 000 000 (zus. 154 692 500); Anl.-Zinsen 2 459 784, Hyp. 8500, Akzepte 165 000 000, Schulden: Bankschulden 5 198 753, Guthaben d. Mansfeld A.-G. 8 324 700, Sonstige 712 631 828 (zus. 726 155 281) ℳ. — S. b. 1 Md. ℳ.

Sächsische Gussstahl-Werke Döhlen Akt.-Ges. in Dresden.

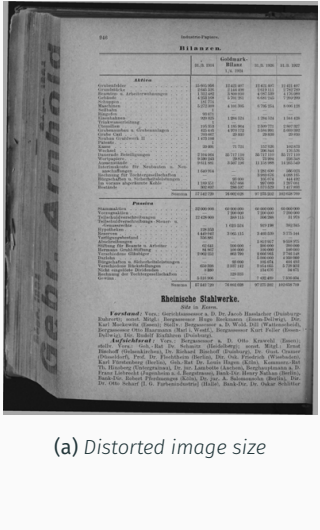
(Bis 27./10. 1920: *Sächsische Gussstahlfabrik in Döhlen*).

Vorstand: Gen.-Dir. Kommerz.-R. Herm. Pfeifer, Gen.-Dir. Kommerz.-R. Herm. Boehm. — *A.-R.:* Vors.: Geh. Kommerz.-R. Gen.-Konsul Gust. v. Klemperer; Stellv.: Kommerz.-R. Willy Osswald [Deutsche Bk.]; sonst. Mitgl.: Bergrat a. D. Andreas Nägel, Justizrat Dr. Jul. Bondi, Geh. Baurat Fritz Martiny (Ob.-Schreiberhau), Bankier Dr. Gustaf Ratjen (Berlin-Dahlem), Kfm. Albert Roth-

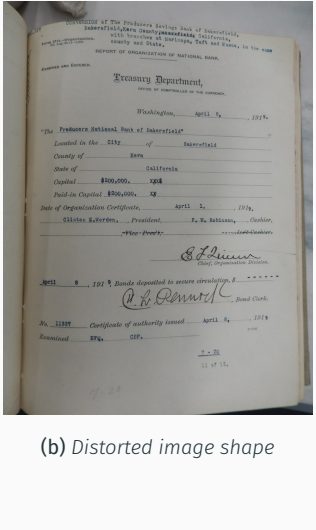
Tackling the datasets

- Each dataset required a different combination of tools
- OCC: good quality scan, already black-and-white
  - No need for preprocessing
  - But required lots of layout recognition for the tables
- *Saling's*: scan not as good, in color, often with artifacts
  - Lots of preprocessing
  - Less need for layout recognition
- Also, what are the *advantages* of these datasets?
  - Balance sheets allow us to validate data: sum of assets = total assets; assets=liabilities

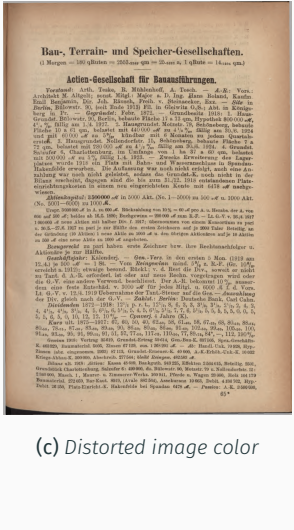
Showcase - Types of Image Distortions



(a) *Distorted image size*

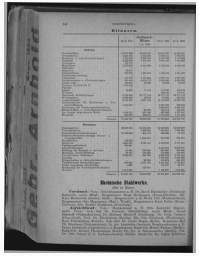


(b) *Distorted image shape*

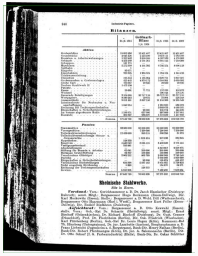


(c) *Distorted image color*

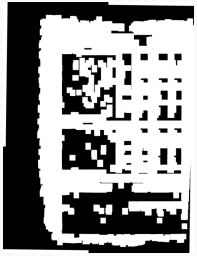
Showcase - Fixing Size Distortions



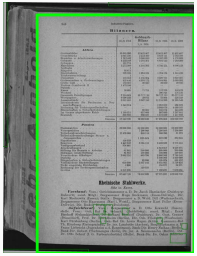
(a) *Input*



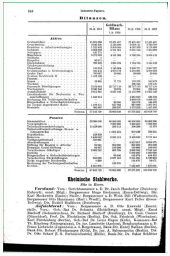
(b) *Remove noise*



(c) *Dilate white area*

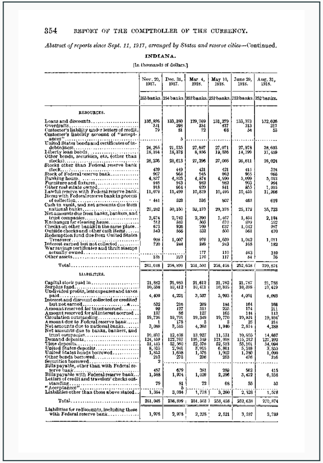


(d) *Identify white box*



(e) *Trim*

Showcase - Detecting Table Delimiters



354 REPORT OF THE COMPTROLLER OF THE CURRENCY.

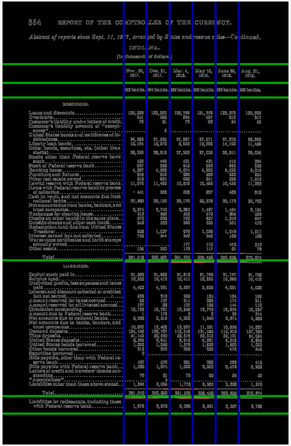
Abstract of reports since Sept. 11, 1917, arranged by States and reserve cities—Continued.

INDIANA.

[In thousands of dollars.]

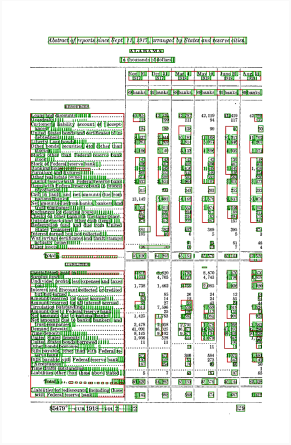
	Nov. 20, 1917.	Dec. 31, 1917.	Mar. 4, 1918.	May 10, 1918.	June 29, 1918.	Aug. 31, 1918.
	253 banks.	254 banks.	252 banks.	252 banks.	252 banks.	252 banks.
RESOURCES.						
Loans and discounts	136,896	133,380	129,769	131,279	133,373	132,626
Overdrafts	341	298	334	437	313	317
Customer's liability under letters of credit	79	81	72	68	54	53
Customer's liability account of "acceptances"		5				
United States bonds and certificates of indebtedness	24,265	21,535	27,697	27,071	27,978	38,693
Liberty loan bonds	18,164	18,378	8,956	19,036	14,199	11,459
Other bonds, securities, etc. (other than stocks)	28,226	28,613	27,296	27,005	26,611	26,024
Stocks other than Federal reserve bank stock	439	449	431	421	415	324
Stock of Federal reserve bank	967	968	945	962	965	966
Banking house	4,827	4,823	4,874	4,990	5,099	5,013
Furniture and fixtures	946	943	983	982	903	994
Other real estate owned	918	964	920	841	855	1,015
Lawful reserve with Federal reserve bank	11,079	11,499	10,819	10,495	10,455	11,266
Items with Federal reserve bank in process of collection	441	523	536	807	463	619
Cash in vault, and net amounts due from national banks	27,282	30,150	32,170	29,376	25,172	35,723
Net amounts due from banks, bankers, and trust companies	2,674	2,742	2,390	1,467	1,454	2,184
Exchanges for clearing house	512	582	505	570	599	552
Checks on other banks in the same place	673	926	709	637	1,043	297
Outside checks and other cash items	543	566	523	505	561	470
Redemption fund due from United States Treasurer	998	1,007	979	1,039	1,042	1,011
Interest earned but not collected	720	244	249	243	168	182
War savings certificates and thrift stamps actually owned			177	110	443	310
Other assets	138	223	170	117	84	76
Total	261,018	258,899	251,503	258,458	252,638	270,874
LIABILITIES.						
Capital stock paid in	21,882	21,983	21,612	21,762	21,787	21,788
Surplus fund	10,358	10,412	10,411	10,335	10,398	10,419
Undivided profits, less expenses and taxes paid	4,409	4,221	3,537	3,993	4,091	4,093
Interest and discount collected or credited but not earned	532	216	209	184	164	166
Amount reserved for taxes accrued	99	127	311	225	174	211
Amount reserved for all interest accrued	137	86	127	165	144	113
Circulation outstanding	19,726	19,791	19,349	19,770	19,924	19,934
Amount due to Federal reserve bank	1	176	5	3	25	214
Net amounts due to national banks	3,088	3,515	4,203	1,940	2,674	4,388
Net amounts due to banks, bankers, and trust companies	10,833	12,408	13,927	11,531	10,655	14,667
Demand deposits	124,459	122,767	116,319	121,098	115,212	127,292
Time deposits	51,443	52,560	52,576	52,513	53,101	54,094
United States deposits	8,525	2,941	3,915	6,861	5,316	3,855
United States bonds borrowed	1,853	1,658	1,576	1,902	1,580	1,090
Other bonds borrowed	213	275	236	263	476	216
Securities borrowed	2					
Bills payable, other than with Federal reserve bank	457	679	381	289	562	415
Bills payable with Federal reserve bank	1,588	1,974	1,029	2,296	3,472	6,258
Letters of credit and travelers' checks outstanding	79	81	72	68	55	53
"Acceptances"		5				
Liabilities other than those above stated	1,364	3,024	1,778	3,260	2,328	1,578
Total	261,048	258,899	251,503	258,458	252,638	270,874
Liabilities for rediscounts, including those with Federal reserve bank	1,978	2,978	2,228	2,521	3,587	3,789

(a) *Preprocessed image*

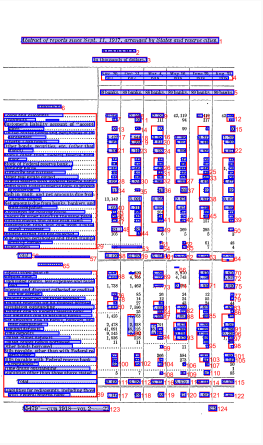


(b) *Image edges and detected lines*

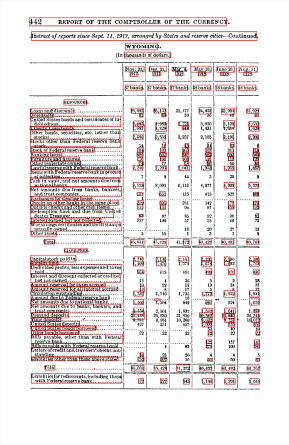
Showcase - Document Hierarchy



(a) *"Words" from GCV*

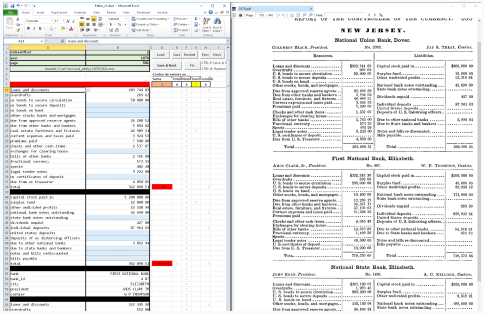


(b) *Lines from* `quipucamayoc`

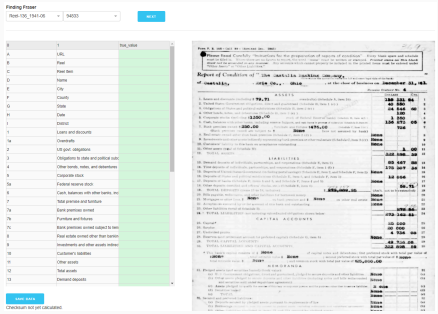


(c) *Fixed labels*

Showcase - Human Review Interfaces



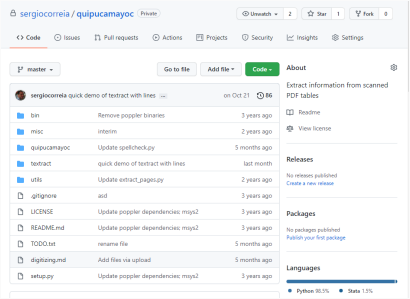
(a) *Excel+VBA*



(b) *Website*

Our tool - quipucamayoc

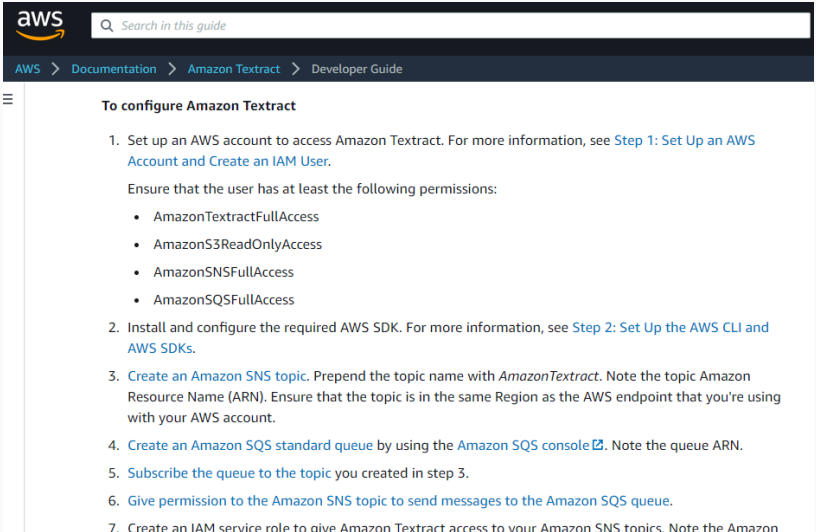
- Built in Python
- Modular tools implementing the methods we've used for these projects
- Should be easily combined with other packages (e.g. other layout parsers)




```
import quipucamayoc as q

doc=q.read('rothschild-secret-ledger.pdf')
page = doc.pages[0]

page.remove_black_background()
page.deskew()
page.binarize()

page.run_ocr(engine='amazon')
page.run_ocr(engine='google')

s = q.spellchecker('frequencies.csv')
process(page, spellchecker=s)
```

Our tool - `quipucamayoc`

`q.setup_textract()` versus:



Lessons Learned

1. Digitization at scale is a Leontief production function:

- Only as good as its weakest step
- Won't be successful if any of the steps (OCR, human review, data extraction, etc) is done poorly
- Other researchers often mention poor results with OCR digitization; we suspect this is why

2. Use a cloud OCR provider:

- Not worth it to use a cheaper (free) OCR engine and then waste lots of time (=money) improving its results.
- Cloud OCR providers, today, are performant, robust, and still quite cheap ($0.001-$0.0015 per page)
- Cloud providers work in parallel without the need of us maintaining multiple severs.

3. **Use the right tools when pre- and post-processing:**

- Python worked better than {R, Stata}
- Well supported tools such as OpenCV worked better than state-of-the-art tools not yet battle-tested

4. **Make human validation efficient:**

- Human validation can be vastly sped up by using programs that help the reviewer:
  - Keyboard shortcuts
  - Side-to-side image-data comparisons
  - Auto zoom-in
  - Auto flag potential errors
- Not just about saving time, but about preserving *focus*

Thanks!