OCR Technique: Input

NEW YORK.

Chase National Bank, New York.

H. W. CANNON, *President.* No. 2370. J. T. MILLS, JR., *Cashier.*

Resources.		Liabilities.	
Loans and discounts	$14,954,408.80	Capital stock paid in	$500,000.00
Overdrafts	4,683.41		
U. S. bonds to secure circulation	50,000.00	Surplus fund	1,000,000.00
U. S. bonds to secure deposits	200,000.00	Undivided profits, less current expenses and taxes paid	284,769.70
U. S. bonds on hand	167,350.00		
Premiums on U. S. bonds	26,782.06	National-bank notes outstanding	45,000.00
Stocks, securities, etc	2,882,266.11	State-bank notes outstanding	.....
Bank'g house, furniture, and fixtures	.....		
Other real estate and mortg's owned	.....	Due to other national banks	9,309,113.60
Due from other national banks	759,750.57	Due to State banks and bankers	9,312,523.30
Due from State banks and bankers	239,149.84		
Due from approved reserve agents	.....	Dividends unpaid	.....
Checks and other cash items	5,855.93	Individual deposits	4,641,779.71
Exchanges for clearing house	375,878.15	Certified checks	153,674.33
Bills of other national banks	45,250.00	United States deposits	.....
Fractional currency, nickels, cents	262.13	Deposits of U.S. disbursing officers	110,450.36
Specie	739,586.00		
Legal-tender notes	883,838.00	Notes and bills rediscounted	.....
U. S. certificates of deposit	4,020,000.00	Bills payable	.....
Redemption fund with Treas. U. S.	2,250.00		
Due from Treasurer U. S.	.....	Liabilities other than those above stated	.....
Total	25,357,311.00	Total	25,357,311.00

back

OCR Technique: Naïve Output

```
REPORT OF DL A I 'THE COMPTROLLER OF THE CURRENCY.NE.
Bank, Thomaston.Georges National Epwarp O'BRIEN
U. S. certificates of deposit ..... mense e eae nce
Bills PAY@DIGÂ® . 000020 see ce ek ces
Due from U. S. Treasurer .........4,500 00
_-TOb@L 220000.245, 606 05000.00}
245, 606 05
```

- OCR on its own useless!
- Three ways of improving OCR:
  1. Multiple sources and engines
  2. Topology of source
  3. Autocorrection

OCR Technique: Naïve Output

```
REPORT OF DL A I 'THE COMPTROLLER OF THE CURRENCY.NE.
Bank, Thomaston.Georges National Epwarp O'BRIEN
U. S. certificates of deposit ..... mense e eae nce
Bills PAY@DIGÂ® . 000020 see ce ek ces
Due from U. S. Treasurer .........4,500 00
_-TOb@L 220000.245, 606 05000.00}
245, 606 05
```

→ OCR on its own useless!

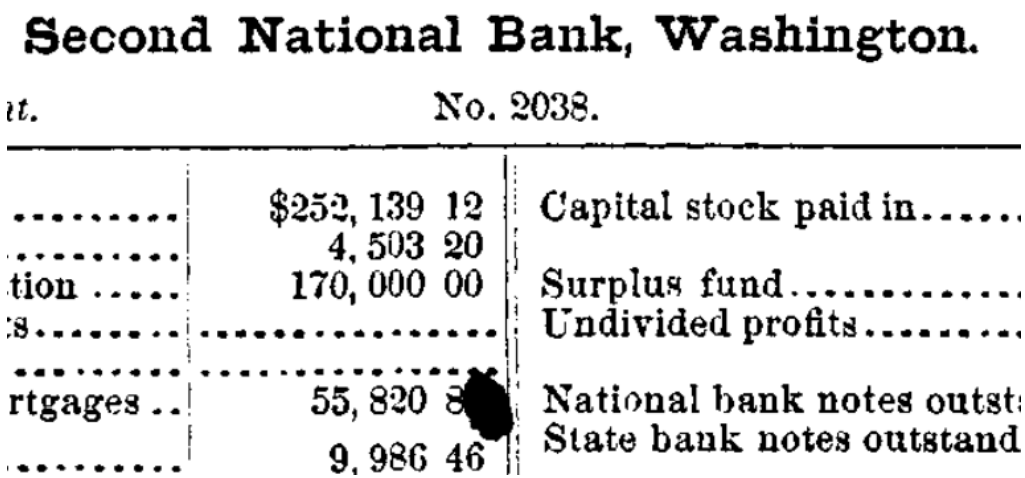
Three ways of improving OCR:

1. Multiple sources and engines
2. Topology of source
3. Autocorrection

back

OCR Improvement 1: Ensemble of Scans and Engines

- Old documents often have stains, ink blots and broken pages
- Different OCR engines make different errors: one confuses *BOB* with *808*; another confuses *bonds* with *bomls*



Second National Bank, Washington.

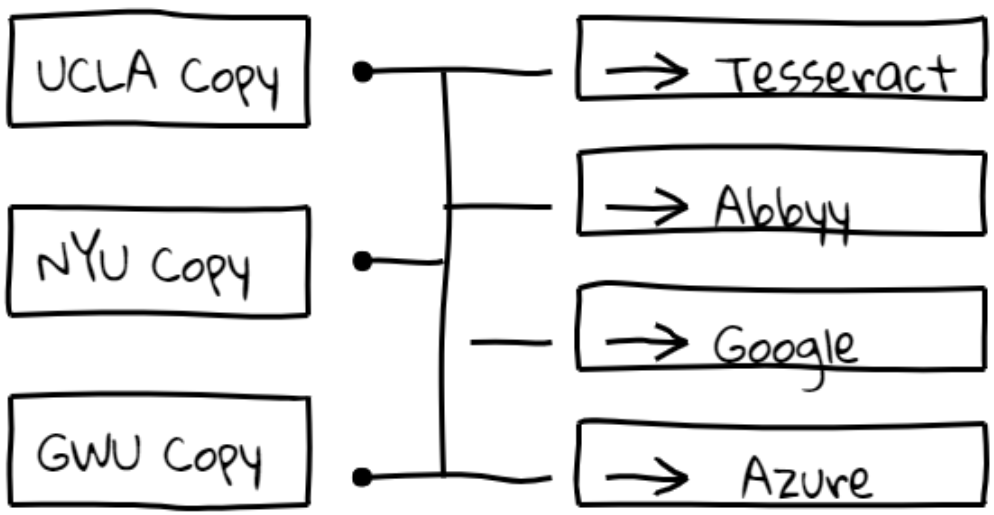
it. No. 2038.

.....	$252, 139 12	Capital stock paid in.....
.....	4, 503 20	
tion	170, 000 00	Surplus fund.....
s.....	.....	Undivided profits.....
.....	.....	
rtgages ..	55, 820 8[redacted]	National bank notes outst
.....	9, 986 46	State bank notes outstand

back

OCR Improvement 1: Ensemble of Scans and Engines

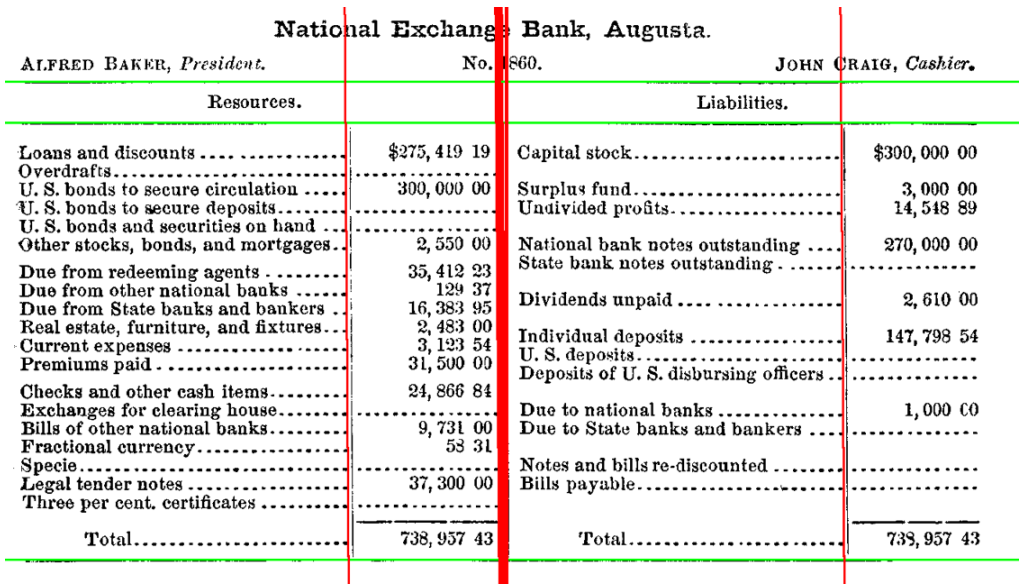
- We gather as many different documents as possible and apply several modern OCR engines
- For example, in 1895 there are 12 inputs from three copies and four OCR engines:



back

OCR Improvement 2: Use Computer Vision to Detect Columns and Rows

- To separate each column, we use OpenCV to detect all lines and separate columns (Konrad 2017)



National Exchange Bank, Augusta.

ALFRED BAKER, *President.* No. 1860. JOHN CRAIG, *Cashier.*

Resources.		Liabilities.	
Loans and discounts	$275,419 19	Capital stock	$300,000 00
Overdrafts			
U. S. bonds to secure circulation	300,000 00	Surplus fund	3,000 00
U. S. bonds to secure deposits		Undivided profits	14,548 89
U. S. bonds and securities on hand			
Other stocks, bonds, and mortgages	2,550 00	National bank notes outstanding	270,000 00
		State bank notes outstanding	
Due from redeeming agents	35,412 23		
Due from other national banks	129 37	Dividends unpaid	2,610 00
Due from State banks and bankers	16,383 95		
Real estate, furniture, and fixtures	2,483 00	Individual deposits	147,798 54
Current expenses	3,123 54	U. S. deposits	
Premiums paid	31,500 00	Deposits of U. S. disbursing officers	
Checks and other cash items	24,866 84		
Exchanges for clearing house		Due to national banks	1,000 00
Bills of other national banks	9,731 00	Due to State banks and bankers	
Fractional currency	58 31		
Specie		Notes and bills re-discounted	
Legal tender notes	37,300 00	Bills payable	
Three per cent. certificates			
Total	738,957 43	Total	738,957 43

back

OCR Improvement 3: Autocorrect Errors

- Construct list of possible balance sheet items, city names, etc.
- Feed to a spelling corrector (Norvig 2016)
- Example: `Washlngion` → `Washington`
- For numbers, enforce heuristics (no leading 0, three digits between commas, etc.)
- Allow OCR engines to vote, digit by digit:

True value:	123456	
Abbyy OCR:	23456	Skipped first digit
Tesseract OCR:	128450	Confused 3 and 8
Azure OCR:	123156	Confused 1 and 4
Select common digit:	123456	

OCR Improvement 3: Autocorrect Errors

- Construct list of possible balance sheet items, city names, etc.
- Feed to a spelling corrector (Norvig 2016)
- Example: `Washlngion` → `Washington`
- For numbers, enforce heuristics (no leading 0, three digits between commas, etc.)
- Allow OCR engines to vote, digit by digit:

True value:	123456	
Abbyy OCR:	_23456	Skipped first digit
Tesseract OCR:	128450	Confused 3 and 8
Azure OCR:	123156	Confused 1 and 4
Select common digit:	123456	

back

Data

- Accurately identified all banks, including charter numbers
  - Geolocated cities (Schmidt 2017)
  - Major events: entry, exit (receivership, liquidation, mergers)
- Assess accuracy through balance sheet identities:
  - Total Assets $=$ Total Liabilities (97% of cases)
  - Total Assets $= \Sigma$ Assets$_i$ (65%)
  - Total Liabilities $= \Sigma$ Liabilities$_i$ (80%)
  - **All three: 52% of cases.**
- Residual gets fixed by hand.

back