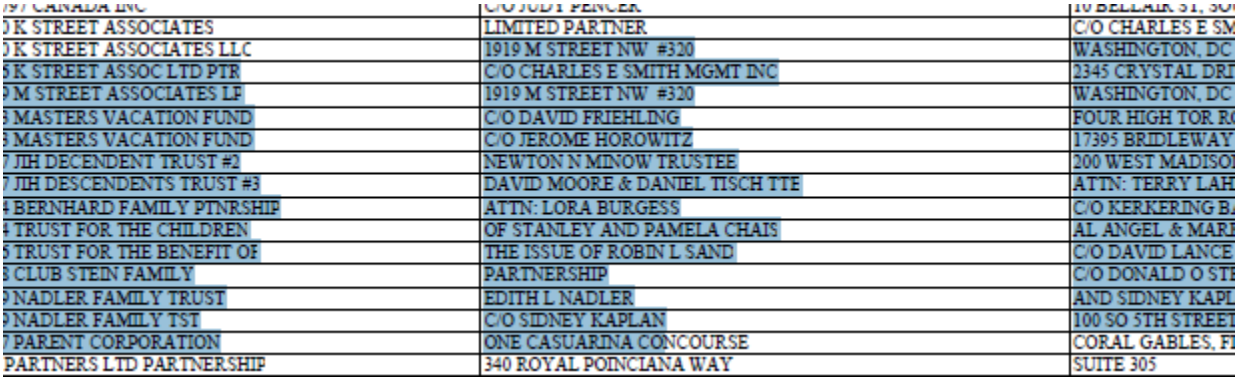
Extracting Data from Image-Based PDFs

You'll face two basic scenarios when extracting data from PDFs: documents that are text-based and documents that are image-based.

When the document is text-based, it's often fairly easy to extract reliable information. Image-based documents generally present many more problems.

Any easy way to determine whether your PDF contains text or images is to try to highlight the content using the mouse.

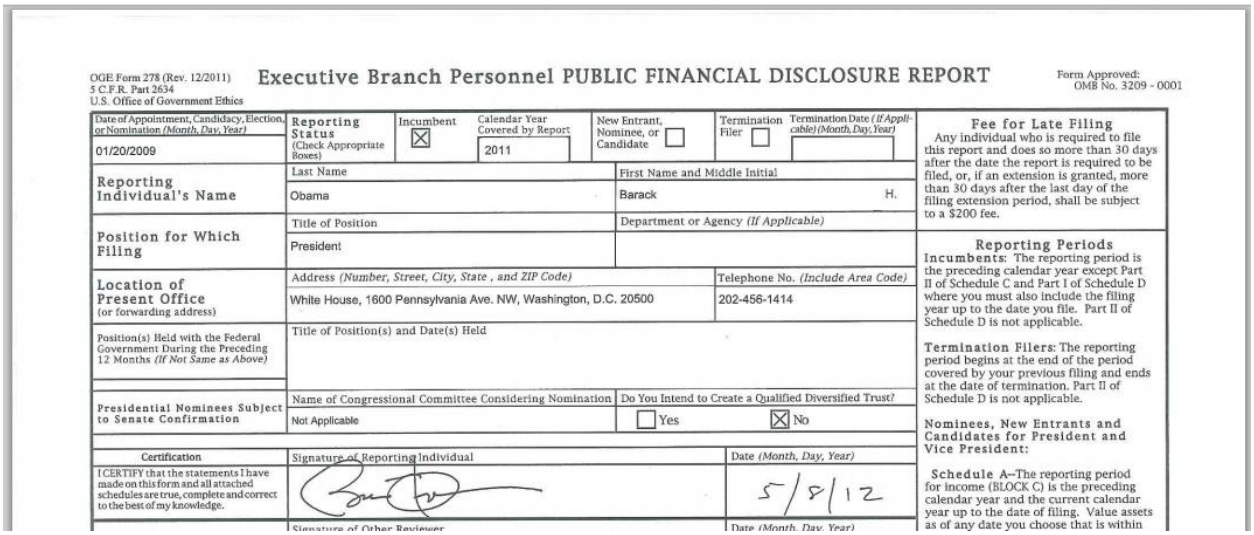


K STREET ASSOCIATES	LIMITED PARTNER	C/O CHARLES E SM
K STREET ASSOCIATES LLC	1919 M STREET NW #320	WASHINGTON, DC
K STREET ASSOC LTD PTR	C/O CHARLES E SMITH MGMT INC	2345 CRYSTAL DRI
M STREET ASSOCIATES LP	1919 M STREET NW #320	WASHINGTON, DC
MASTERS VACATION FUND	C/O DAVID FRIEHLING	FOUR HIGH TOR R
MASTERS VACATION FUND	C/O JEROME HOROWITZ	17395 BRIDLEWAY
JIH DECENDENT TRUST #2	NEWTON N MINOW TRUSTEE	200 WEST MADISO
JIH DESCENDENTS TRUST #3	DAVID MOORE & DANIEL TISCH TTE	ATTN: TERRY LAH
BERNHARD FAMILY PTNRSHIP	ATTN: LORA BURGESS	C/O KERKERING B
TRUST FOR THE CHILDREN	OF STANLEY AND PAMELA CHAIS	AL ANGEL & MAR
TRUST FOR THE BENEFIT OF	THE ISSUE OF ROBIN L SAND	C/O DAVID LANCE
CLUB STEIN FAMILY	PARTNERSHIP	C/O DONALD O STE
NADLER FAMILY TRUST	EDITH L NADLER	AND SIDNEY KAPL
NADLER FAMILY TST	C/O SIDNEY KAPLAN	100 SO 5TH STREE
PARENT CORPORATION	ONE CASUARINA CONCOURSE	CORAL GABLES, FI
PARTNERS LTD PARTNERSHIP	340 ROYAL POINCIANA WAY	SUITE 305

If Acrobat automatically selects the text (as in the image above), you can be relatively sure the PDF is text-based. This means you'll probably be able to use one of the many free PDF data extraction tools (like Tabula) to pull your records.

If, on the other hand, you can't select the text, you probably have an image-based PDF. This generally means the document has been scanned from a paper copy. Government agencies will often respond to public records requests in this format.

As an example, take a look at the first page of Barack Obama's 2011 public financial disclosure report (available here):



OGE Form 278 (Rev. 12/2011) 5 C.F.R. Part 2634 U.S. Office of Government Ethics

Executive Branch Personnel PUBLIC FINANCIAL DISCLOSURE REPORT

Form Approved: OMB No. 3209 - 0001

Date of Appointment, Candidacy, Election, or Nomination (Month, Day, Year)	Reporting Status (Check Appropriate Boxes)	Incumbent	Calendar Year Covered by Report	New Entrant, Nominee, or Candidate	Termination Filer	Termination Date (If Applicable) (Month, Day, Year)
01/20/2009		☒	2011	☐	☐	

Reporting Individual's Name	Last Name	First Name and Middle Initial
	Obama	Barack H.

Position for Which Filing	Title of Position	Department or Agency (If Applicable)
	President	

Location of Present Office (or forwarding address)	Address (Number, Street, City, State, and ZIP Code)	Telephone No. (Include Area Code)
	White House, 1600 Pennsylvania Ave. NW, Washington, D.C. 20500	202-456-1414

Position(s) Held with the Federal Government During the Preceding 12 Months (If Not Same as Above)	Title of Position(s) and Date(s) Held

Presidential Nominees Subject to Senate Confirmation	Name of Congressional Committee Considering Nomination	Do You Intend to Create a Qualified Diversified Trust?
	Not Applicable	☐ Yes ☒ No

Certification	Signature of Reporting Individual	Date (Month, Day, Year)
I CERTIFY that the statements I have made on this form and all attached schedules are true, complete and correct to the best of my knowledge.	[signature]	5/8/12
	Signature of Other Reviewer	Date (Month, Day, Year)

Fee for Late Filing

Any individual who is required to file this report and does so more than 30 days after the date the report is required to be filed, or, if an extension is granted, more than 30 days after the last day of the filing extension period, shall be subject to a $200 fee.

Reporting Periods

Incumbents: The reporting period is the preceding calendar year except Part II of Schedule C and Part I of Schedule D where you must also include the filing year up to the date you file. Part II of Schedule D is not applicable.

Termination Filers: The reporting period begins at the end of the period covered by your previous filing and ends at the date of termination. Part II of Schedule D is not applicable.

Nominees, New Entrants and Candidates for President and Vice President:

Schedule A–The reporting period for income (BLOCK C) is the preceding calendar year and the current calendar year up to the date of filing. Value assets as of any date you choose that is within

While this document was obviously created by a computer, it was printed out and then scanned back in. This process effectively destroys the ability of your computer to simply read the text.

The core challenge with image-based PDFs is to convert the document back into text and recognize relationships between different regions on the document (tables, etc.)

One of the best tools for accomplishing this task is ABBYY FineReader:

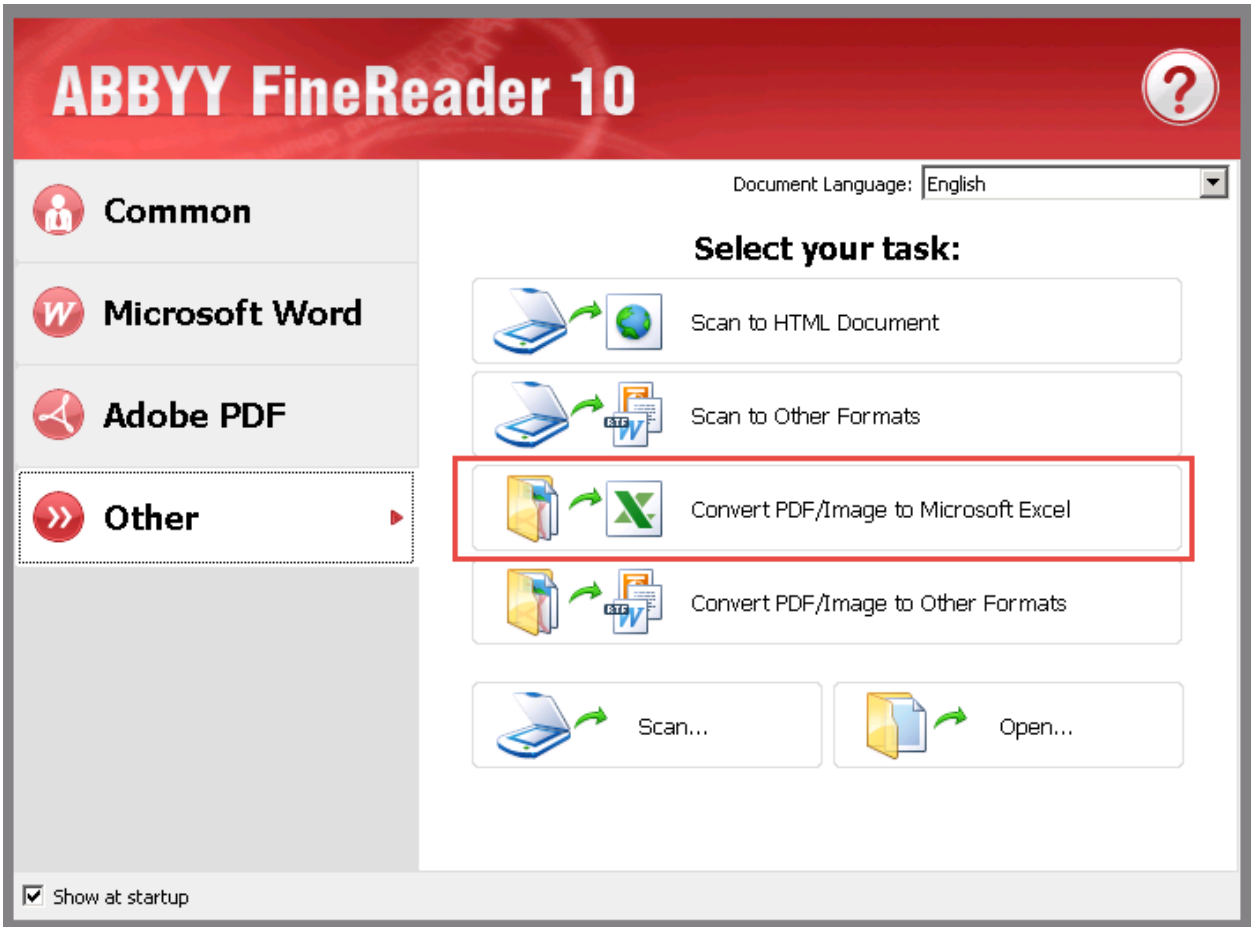
ABBYY FineReader
http://finereader.abbyy.com/
Starts at $99 (student discounts may be available)

This tool analyzes the contents of PDF files using a process called **optical character recognition (OCR)**.

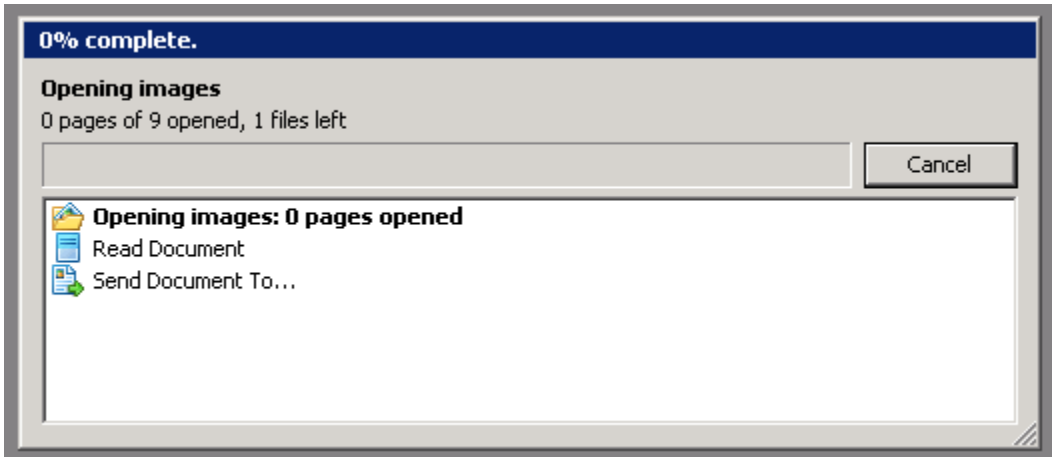
ABBYY FineReader's OCR process is very powerful and can help accomplish two key goals:

- Convert images to text that can be searched for keywords.
- Recognize and extract tables into Excel or CSV.

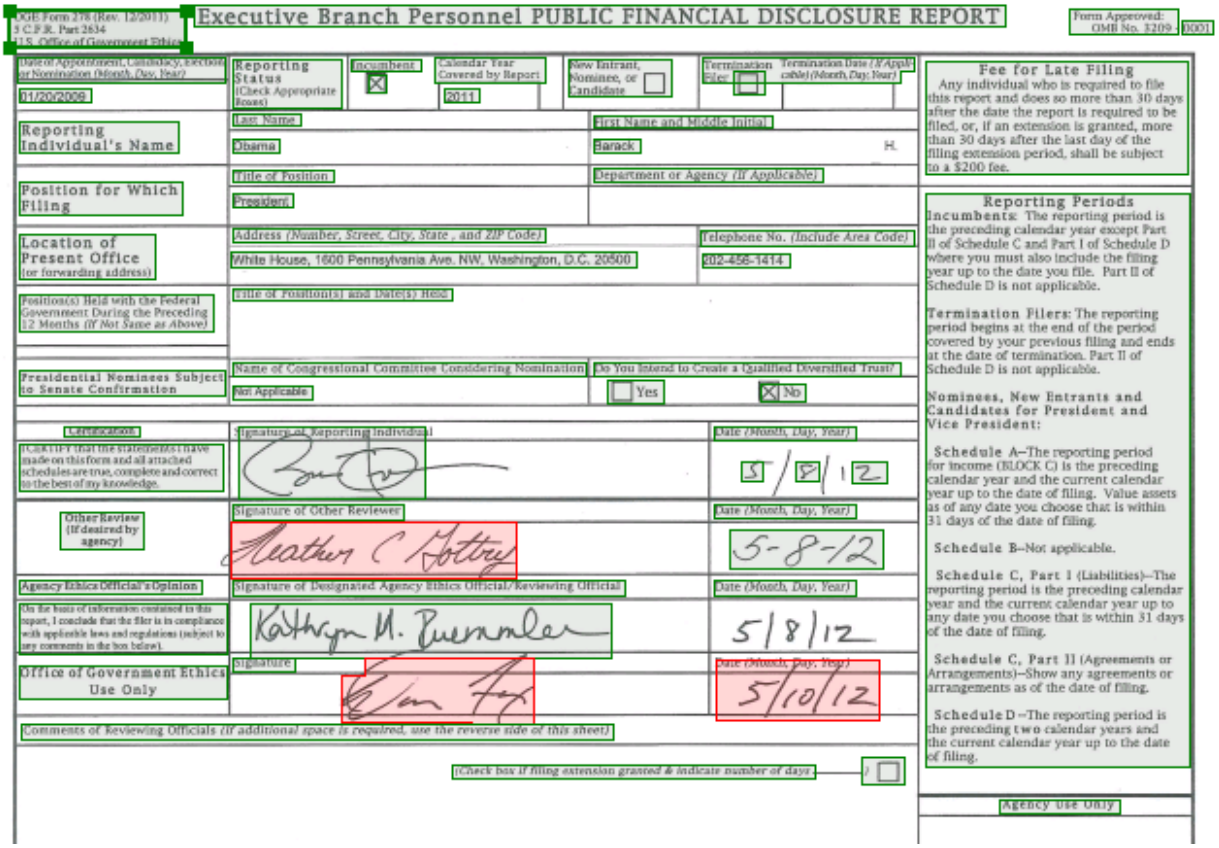
To demonstrate how this program works, I'm going to load Obama's 2011 disclosure into ABBYY and instruct it to convert the PDF into an Excel file.



Depending on the size of the document, the process can take quite some time. Obama's disclosure is only nine pages long, so for us, it's fairly quick.



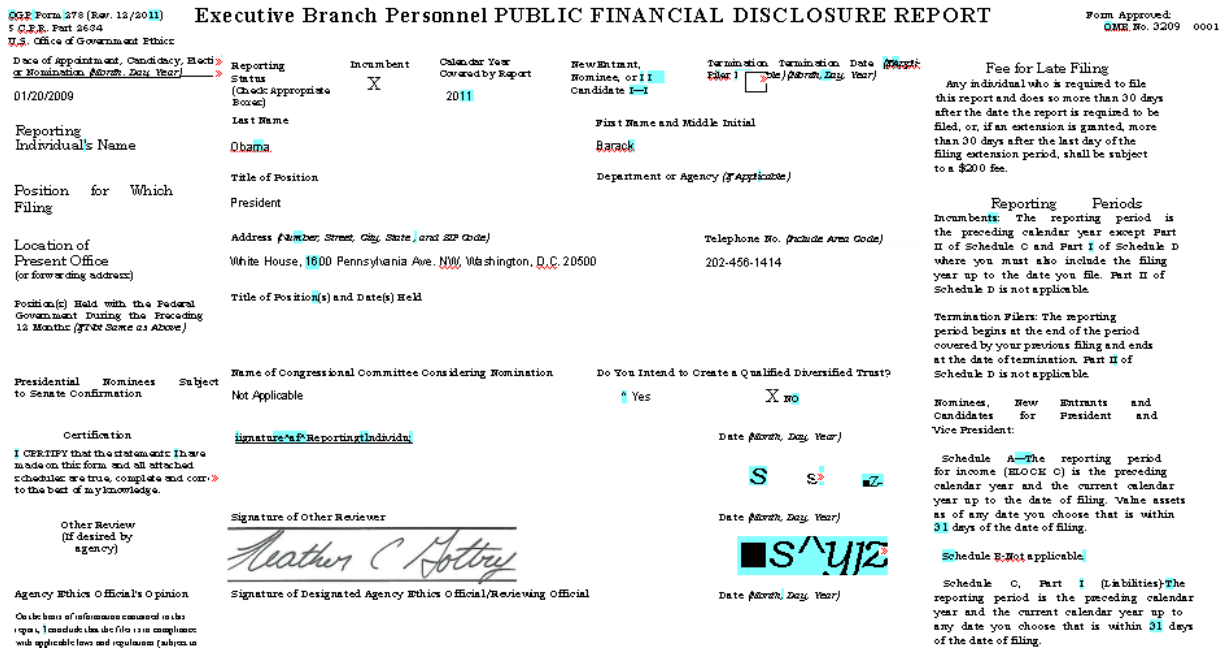
Once the process is complete, ABBYY will return its best effort at recognizing the component pieces of each page. For instance, here is ABBYY's interpretation of the first page of Obama's disclosure:



Abby automatically outlines the various sections of the document that it thinks go together. This is a very complex task and in many cases the program will need your guidance in correctly identifying tables and text.

In the case of Obama's disclosure, you can see the program automatically highlighted numerous regions of the document's front page. It gets it somewhat right and correctly determines that parts of the content are unreadable (outlined in red).

Here's what the program spits out:



OGE Form 278 (Rev. 12/2011)
5 C.F.R. Part 2634
U.S. Office of Government Ethics

Executive Branch Personnel PUBLIC FINANCIAL DISCLOSURE REPORT

Form Approved:
OMB No. 3209 0001

Date of Appointment, Candidacy, Election or Nomination (Month, Day, Year)
01/20/2009

Reporting Status (Check Appropriate Box(es)): Incumbent X
Calendar Year Covered by Report: 2011
New Entrant, Nominee, or I I Candidate I—I
Termination Filer I X I Termination Date (If Applicable) (Month, Day, Year)

Reporting Individual's Name
Last Name: Obama
First Name and Middle Initial: Barack

Position for Which Filing
Title of Position: President
Department or Agency (If Applicable)

Location of Present Office (or forwarding address)
Address (Number, Street, City, State, and ZIP Code): White House, 1600 Pennsylvania Ave. NW, Washington, D.C. 20500
Telephone No. (Include Area Code): 202-456-1414

Position(s) Held with the Federal Government During the Preceding 12 Months: (If Same as Above)
Title of Position(s) and Date(s) Held

Presidential Nominees Subject to Senate Confirmation
Name of Congressional Committee Considering Nomination: Not Applicable
Do You Intend to Create a Qualified Diversified Trust? I Yes X NO

Certification
I CERTIFY that the statements I have made on this form and all attached schedules are true, complete and correct to the best of my knowledge.
Signature of Reporting Individual
Date (Month, Day, Year)

Other Review (If desired by agency)
Signature of Other Reviewer
Date (Month, Day, Year)

Agency Ethics Official's Opinion
On the basis of information contained in this report, I conclude that the filer is in compliance with applicable laws and regulations (subject to
Signature of Designated Agency Ethics Official/Reviewing Official
Date (Month, Day, Year)

Fee for Late Filing
Any individual who is required to file this report and does so more than 30 days after the date the report is required to be filed, or, if an extension is granted, more than 30 days after the last day of the filing extension period, shall be subject to a $200 fee.

Reporting Periods
Incumbents: The reporting period is the preceding calendar year except Part II of Schedule C and Part I of Schedule D where you must also include the filing year up to the date you file. Part II of Schedule D is not applicable.

Termination Filers: The reporting period begins at the end of the period covered by your previous filing and ends at the date of termination. Part II of Schedule D is not applicable.

Nominees, New Entrants and Candidates for President and Vice President:

Schedule A—The reporting period for income (BLOCK C) is the preceding calendar year and the current calendar year up to the date of filing. Value assets as of any date you choose that is within 31 days of the date of filing.

Schedule B-Not applicable.

Schedule C, Part I (Liabilities)-The reporting period is the preceding calendar year and the current calendar year up to any date you choose that is within 31 days of the date of filing.

You can see much of the analysis was successful. However, the signature blocks gave ABBYY some trouble. While it correctly recognized one of the signatures as an unreadable image, it took a stab at other signatures and some of the dates. The results are nonsensical text.

One of the challenges of dealing with image-based PDF is weeding out and correcting these types of errors.

The rest of the document is analyzed fairly neatly and the contents are pumped into Excel.

For example, schedule A, which, in the original disclosure document, looked like this:

Reporting Individual's Name	SCHEDULE A	Page Number
Obama, Barack H.		2 of 8

Assets and Income	Valuation of Assets at close of reporting period														Income: type and amount. If "None (or less than $201)" is checked, no other entry is needed in Block C for that item.															
BLOCK A	BLOCK B														BLOCK C															
															Type				Amount											
For you, your spouse, and dependent children, report each asset held for investment or the production of income which had a fair market value exceeding $1,000 at the close of the reporting period, or which generated more than $200 in income during the reporting period, together with such income. For yourself, also report the source and actual amount of earned income exceeding $200 (other than from the U.S. Government). For your spouse, report the source but not the amount of earned income of more than $1,000 (except report the actual amount of any honoraria over $200 of your spouse). None ☐	None (or less than $1,001)	$1,001 - $15,000	$15,001 - $50,000	$50,001 - $100,000	$100,001 - $250,000	$250,001 - $500,000	$500,001 - $1,000,000	Over $1,000,000*	$1,000,001 - $5,000,000	$5,000,001 - $25,000,000	$25,000,001 - $50,000,000	Over $50,000,000	Excepted Investment Fund	Excepted Trust	Qualified Trust	Dividends	Rent and Royalties	Interest	Capital Gains	None (or less than $201)	$201 - $1,000	$1,001 - $2,500	$2,501 - $5,000	$5,001 - $15,000	$15,001 - $50,000	$50,001 - $100,000	$100,001 - $1,000,000	Over $1,000,000*	$1,000,001 - $5,000,000	Over $5,000,000	Other Income (Specify Type & Actual Amount)	Date (Mo., Day, Yr.) Only if Honoraria
Examples: Central Airlines Common				x											x						x									
Examples: Doe Jones & Smith, Hometown, State			x																										Law Partnership Income $130,000	
Examples: Kempstone Equity Fund					x								x									x								
Examples: IRA: Heartland 500 Index Fund							x						x											x						
1 JPMorgan Chase Private Client Asset Mgmt Checking Account (J)							X										X		X											
2 Northern Trust Checking Account (J)			X														X		X											
3 Vanguard 500 Index Fund (Retirement)				X									X								X									
4 State of Illinois General Assembly Defined Benefit Pension Plan				X															X											

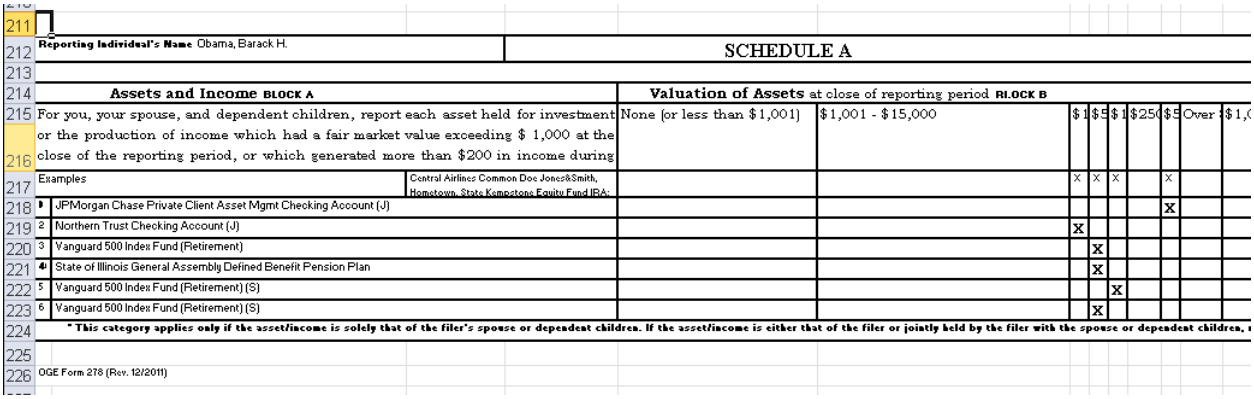
Was correctly interpreted by ABBYY as a large table:

OGE Form 278 (Rev. 12/2011)
5 C.F.R. Part 2634
U.S. Office of Government Ethics

Reporting Individual's Name	SCHEDULE A	Page Number
Obama, Barack H.		2 of 8

Assets and Income	Valuation of Assets at close of reporting period														Income: type and amount. If "None (or less than $201)" is checked, no other entry is needed in Block C for that item.															
BLOCK A	BLOCK B														BLOCK C															
															Type				Amount											
For you, your spouse, and dependent children, report each asset held for investment or the production of income which had a fair market value exceeding $1,000 at the close of the reporting period, or which generated more than $200 in income during the reporting period, together with such income. For yourself, also report the source and actual amount of earned income exceeding $200 (other than from the U.S. Government). For your spouse, report the source but not the amount of earned income of more than $1,000 (except report the actual amount of any honoraria over $200 of your spouse). None ☐	None (or less than $1,001)	$1,001 - $15,000	$15,001 - $50,000	$50,001 - $100,000	$100,001 - $250,000	$250,001 - $500,000	$500,001 - $1,000,000	Over $1,000,000*	$1,000,001 - $5,000,000	$5,000,001 - $25,000,000	$25,000,001 - $50,000,000	Over $50,000,000	Excepted Investment Fund	Excepted Trust	Qualified Trust	Dividends	Rent and Royalties	Interest	Capital Gains	None (or less than $201)	$201 - $1,000	$1,001 - $2,500	$2,501 - $5,000	$5,001 - $15,000	$15,001 - $50,000	$50,001 - $100,000	$100,001 - $1,000,000	Over $1,000,000*	$1,000,001 - $5,000,000	Over $5,000,000	Other Income (Specify Type & Actual Amount)	Date (Mo., Day, Yr.) Only if Honoraria
Examples: Central Airlines Common				x											x						x									
Examples: Doe Jones & Smith, Hometown, State			x																										Law Partnership Income $130,000	
Examples: Kempstone Equity Fund					x								x									x								
Examples: IRA: Heartland 500 Index Fund							x						x											x						
1 JPMorgan Chase Private Client Asset Mgmt Checking Account (J)							X										X		X											

That looks fairly reasonable in Excel:



Reporting Individual's Name Obama, Barack H.

SCHEDULE A

Assets and Income BLOCK A		Valuation of Assets at close of reporting period BLOCK B								
For you, your spouse, and dependent children, report each asset held for investment or the production of income which had a fair market value exceeding $ 1,000 at the close of the reporting period, or which generated more than $200 in income during		None (or less than $1,001)	$1,001 - $15,000	$ 1	$5	$ 1	$25(	$5	Over $1,(
Examples	Central Airlines Common Doc Jones&Smith, Hometown, State Kempstone Equity Fund IRA:			X	X	X		X	
1 JPMorgan Chase Private Client Asset Mgmt Checking Account (J)								X	
2 Northern Trust Checking Account (J)				X					
3 Vanguard 500 Index Fund (Retirement)					X				
4 State of Illinois General Assembly Defined Benefit Pension Plan					X				
5 Vanguard 500 Index Fund (Retirement) (S)						X			
6 Vanguard 500 Index Fund (Retirement) (S)					X				

* This category applies only if the asset/income is solely that of the filer's spouse or dependent children. If the asset/income is either that of the filer or jointly held by the filer with the spouse or dependent children,

OGE Form 278 (Rev. 12/2011)

Depending on the structure of your data, tools like Open Refine may help clean the results of an OCR'd document.

More expensive versions of ABBYY have the ability to perform **batch processes**, meaning you can OCR multiple documents without having to manually load each of them into the program.

This can be extremely useful. For example, in late 2012, the University of Colorado released about 2,700 messages from James Eagan Holmes' email account. The documents were image-based PDFs and it wasn't possible to search them for key words (like "guns"). ABBYY's batch processor was incredibly useful for the reporters working on this story, as it allowed us to convert the documents into searchable PDFs.