

Scraping Data from PDFs and Images

In this lab, you will review different data formats that are available online and offline. Next, you will be introduced to tools that help convert data contained in PDFs and images into structured machine-readable formats.

Getting Started

Before we move forward, here is a refresher on data formats:

- **Machine readable, structured:** These are generated by a computer, and are organized in rows and columns. For example - CSV (comma-separated values), TSV (tab-separated values), Excel (.xls)
- **Unstructured:** These are sometimes generated by a computer, but are not organized as data tables by the computer. For example – some PDF, Word, and bitmap images (GIF, JPEG, PNG, BMP)

As part of your day-to-day work, you must have come across data in these file formats:

- **Portable Document Format (PDF):** These files may include tables that contain data, but the data is saved in a unified document with text.
- **Excel file (XLS):** These files save data as tables, which are readable by Microsoft Excel
- **Comma separated values (CSV):** These are plain text files with each data point separated by a comma

In order to analyze data that you find in a PDF, you will need to convert it in a format which is machine-readable and structured (for instance to XLS format).

Example: Argentina's Senate Expenses



In 2014, Argentina's Senate published information about their expenses from 2004 – 2013 as raw PDFs and images. This vast amount of information was unstructured and was not machine readable – making it challenging to analyze. A team from the LA NACION newspaper managed to scrape, transform, normalize, and structure these datasets and analyzed the data to produce front-page stories about the expenses. The stories elicited replies from current and former Senate presidents, and provoked a judicial investigation regarding these expenses. This series of stories also lead to different approaches to keep Senate accountable. For more information, take a look at:

<http://blogs.lanacion.com.ar/projects/data/argentina%C2%B4s-senate-expenses-2004-2013/>

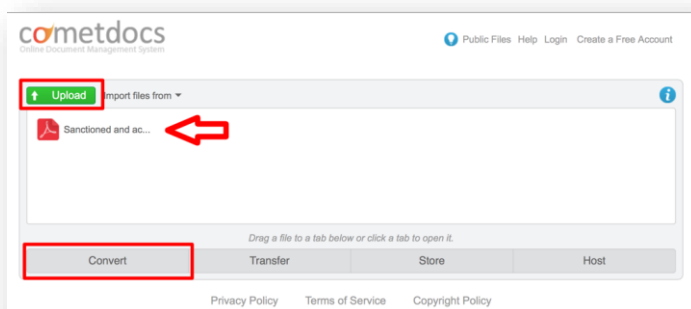
Task 1: Converting PDF to Excel Using Online Tools

Several tools help you convert a data table within a PDF file to the Excel format. Each uses slightly different technology and it is worth it to try more than one and see which results in a cleaner data file rather than wasting a lot of time cleaning a messy file. One such tool that is available online is Cometdocs. It works especially well if your table has background shading in multiple colors instead of just being a black and white table. You can access it here: www.cometdocs.com

The document work for these exercise come from the [Cabinet Secretariat Establishment Division Government of Pakistan](#) the full document is a compilation of [statistics](#) for sanctioned strength in different offices of the government. That is, the data tells us how many employees the government body has been approved for and how many each actually has. Table-I refers to Sanctioned, actual strength of employees of autonomous and Semi-autonomous bodies in corporations from 2013.

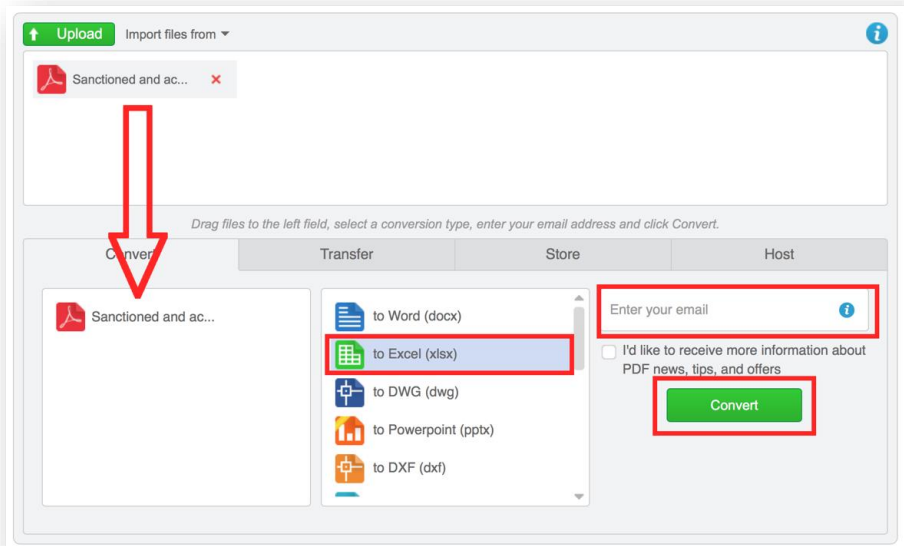
Let's try working with Cometdocs to convert a PDF to Excel. Here are the steps:

1. Open www.cometdocs.com in your web browser
2. Click the **Go to the Web App** button. The screen updates.
3. Click the **Upload** button to upload *Sanctioned and actual strength of employees.pdf*¹ to Cometdocs. Once uploaded the file displays in the window

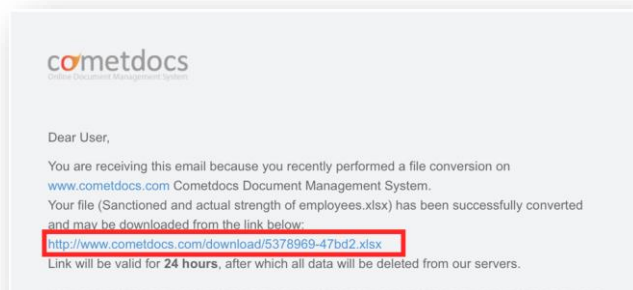


4. Click **Convert**. The screen updates with an empty box under **Convert**.
5. Now, click the PDF file icon in the box above, then drag-and-drop this file to the empty box under **Convert**.
6. The screen refreshes. Now select the conversion type **to Excel (xlsx)**
7. Next, enter your e-mail address in the 'Enter your email' field. Then, click **Convert**. Cometdocs will now send you a hyperlink to access the converted Excel file.

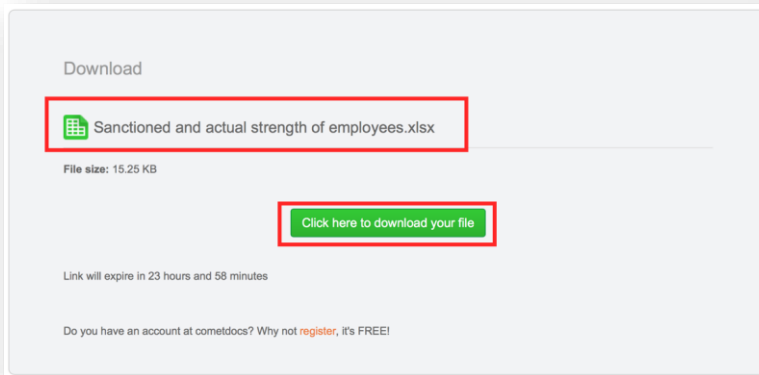
¹ Document from <http://202.83.164.29/estab/userfiles1/file/Establishment/publication/may/Statistics.pdf>



8. Open your e-mail and click the link provided in the e-mail for Cometdocs

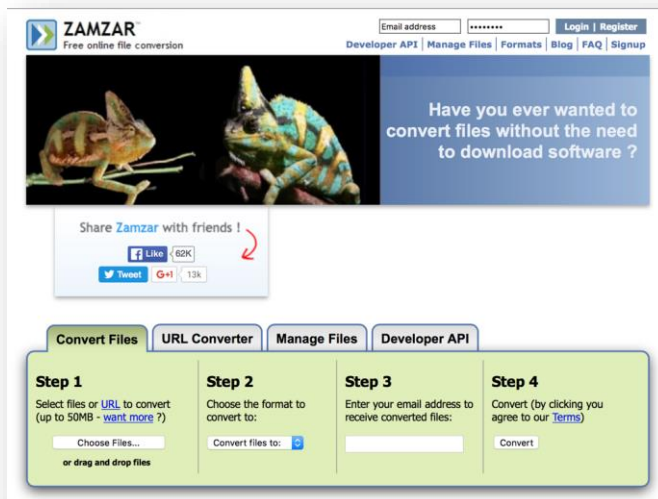


9. The link will open in your default browser; click the link to download the converted Excel file.



Convert PDF to Excel: Exercise

10. Open *Hospital and Beds Private and Government 2013.pdf*², and try converting to Excel but these time use <http://www.zamzar.com/>. It is another program similar to cometdoc.

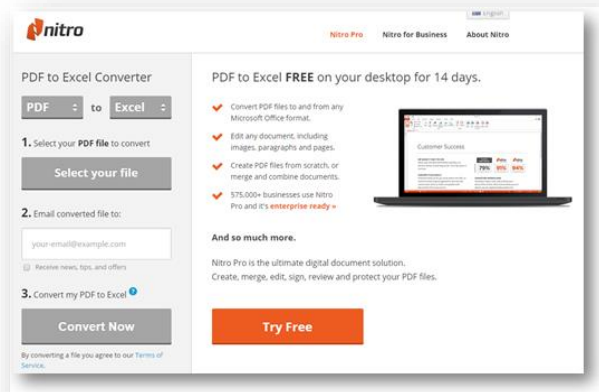


² Document from http://kpbos.gov.pk/prd_images/1399372381.pdf

More Online Tools to Convert PDFs to Excel

Apart from the two online tools that you tried so far, here are two more websites that you can use to convert PDFs to Excel:

PDF to Excel Online



The screenshot shows the Nitro PDF to Excel Converter website. The header includes the Nitro logo and navigation links for Nitro Pro, Nitro for Business, and About Nitro. The main content area is titled "PDF to Excel Converter" and features a dropdown menu to select the file type (PDF to Excel). Below this, there are three steps: 1. Select your PDF file to convert (with a "Select your file" button), 2. Email converted file to: (with an email input field and a "Receive news, tips, and offers" checkbox), and 3. Convert my PDF to Excel (with a "Convert Now" button). To the right, there is a promotional section titled "PDF to Excel FREE on your desktop for 14 days." which lists benefits like converting PDF files to and from any Microsoft Office format, editing any document, creating PDF files from scratch, and a testimonial from a customer. A "Try Free" button is prominently displayed at the bottom right.

To access, go to: <https://www.pdfexcelonline.com/#>

PDF to Excel Org



The screenshot shows the PDF to Excel Org website. The header features the "PDF EXCEL" logo, a "PDF XLS" icon, and social media links for Google+, Twitter, Facebook, and LinkedIn. The main heading is "The Best PDF to Excel Spreadsheet Conversion Available" followed by "Convert PDF to Excel Free Online". The form is divided into two steps: STEP 1 "Upload File" with a "Choose File" button and "No file chosen" text, and STEP 2 "Email File" with an email input field and a "Send" button. Below the email field, there is a checkbox labeled "I'd like to receive more information about PDF news, tips, and offers".

To access, go to: <http://www.pdfexcel.org/>

Task 2: Converting PDFs to Excel using Tabula



Tabula is a tool that you can install on your computer to extract data from PDF files. It works well for most PDFs with black and white data tables. The interface for the tool is in your browser but it does not need an internet connection to work.

Installing Tabula

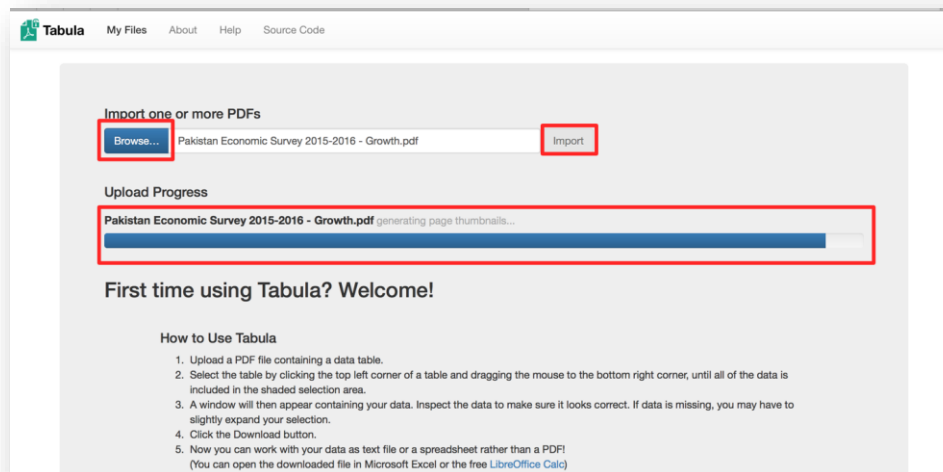
Here are the instructions to download and install Tabula:

11. Ensure Java is installed on your computer. You can download Java here: <https://www.java.com/en/download/>
12. Open the Tabula website: <http://tabula.technology/>
13. Download the version of Tabula for your operating system
14. Tabula downloads as a zip file on your computer. Extract the downloaded **zip** file – this creates a folder called “tabula” on your computer.
15. Go into the “tabula” folder. Run the **tabula.exe** program inside. A control window may open; allow this window to run.
16. Next, a web browser will open – this is Tabula. If your web browser does not open, use your web browser to go to: <http://localhost:34555>

Using Tabula

Here is an example, let's use Tabula to extract the data table that's included in a PDF file. The PDF file is called **Pakistan Economic Survey 2015-2016 – Growth**.

17. Once Tabula is open in your browser window, click the **Browse** button to find and select **Pakistan Economic Survey 2015-2016 - Growth.pdf**. The file name is displayed in Tabula.
18. Now, click **Import**. Tabula processes the PDF file, and shows a preview of the data table included in the PDF **Pakistan Economic Survey 2015-2016 - Growth.pdf** within the Tabula window.



19. Tabula works by enabling you to select the table(s) in a pdf document. You can use 'Autoselect tables' but it does not usually work for large documents as it takes a long time. 'Clear all sections' helps you when you make a mistake selecting the table. You will know the table(s) selected by the red fill square on the table, for example:

Tabula My Files About Help Source Code

Pakistan Economic Survey 2015-2016 - ...

Autodetect Tables Clear All Selections Preview & Export Extracted Data

investment in economic diversification. Particularly, Qatar's economy performed better with double-digit growth in most of the years, this improvement will continue as the country's hosting of the football world Cup may play the role to boost economic activities. Kuwait's economic recovery will continue, it grew at 1.3 percent in 2014 and forecasted to grow at 1.7 percent in 2015 and accelerate further at 1.8 percent in 2016. Saudi Arabia is expected to perform at stabilized growth rate and continue investment in economic diversification and infrastructure. Exporters' earnings have been IMF has forecasted the better growth prospects for the African and South Asian countries for the year 2015 which will further accelerate in coming year. The country wise detail of projected GDP growth is presented in the Table 1.1. The improving performance of Europe is good for the world and it will also have better impact on the economy of Pakistan due to GSP plus status to Pakistan, which will certainly further enhance exports and the industrial performance of Pakistan as the country has a significant volume of trade with Europe.

Table 1.1: Comparative Real GDP Growth Rates (%)

Region/Country	2009	2010	2011	2012	2013	2014	2015	2016P
World GDP	0.0	5.4	4.2	3.4	3.4	3.4	3.5	3.8
Euro Area	-4.5	2.0	1.6	-0.8	-0.5	0.9	1.5	1.6
United States	-2.8	2.5	1.6	2.3	2.2	2.4	3.1	3.1
Japan	-4.0	2.9	0.4	2.6	1.6	-0.1	1.0	1.2
Germany	-5.6	3.9	3.7	0.6	0.2	1.6	1.6	1.7
Canada	-2.7	3.4	1.0	1.9	2.0	2.3	2.2	2.0
Developing Countries	7.7	9.6	7.7	6.8	7.0	6.8	6.6	6.6
China	9.2	10.4	9.3	7.8	7.8	7.4	6.8	6.3
Hong Kong SAR	-2.5	6.8	4.8	1.7	2.9	2.3	2.8	3.1
Korea	0.3	6.9	1.7	2.3	1.0	3.3	3.3	3.5
Singapore	-0.6	15.2	6.2	3.4	4.4	2.9	3.0	3.0
Vietnam	5.4	6.4	6.2	5.2	3.4	6.0	6.0	5.8
ASEAN								
Indonesia	4.7	6.4	6.2	6.0	5.6	5.0	5.2	5.5
Malaysia	-1.5	7.4	5.2	5.6	4.7	6.0	4.8	4.9
Thailand	-2.3	7.8	0.1	6.5	2.9	0.7	3.7	4.0
Philippines	1.1	7.6	3.7	6.8	7.2	6.1	6.7	6.3

Repeat this Selection

20. Go to page 6, there you will find table 1.2 Growth Rate by Sector from 2007 to 2015. As you see the table runs across two separate pages so is necessary to make two sections. **IMPORTANT:** Select the full area of the table but do not expand too much to include text that is not part of the table. Usually, do not include table titles or notes because they do not follow the table format. But make sure you do select the area around the entire table.

Tabula My Files About Help Source Code

Pakistan Economic Survey 2015-2016 - ... Autodetect Tables Clear All Selections Preview & Export Extracted Data

Investment, forestry and fishing. Industry is also

Table 1.2: Growth Rate (%)

Sectors/Sub-Sectors	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15 P
A. Agriculture	1.8	3.5	6.2	2.0	3.6	2.7	2.7	2.9
Crops	-1.0	5.2	-4.2	1.0	3.2	1.5	3.2	1.0
Important Crops	-4.1	8.4	-3.7	1.5	7.9	0.2	8.0	0.3
Other Crops	6.0	0.5	-7.2	2.3	-7.5	5.6	-5.4	1.1
Cotton Ginning -7.0	1.3	7.3	-8.5	13.8	-2.9	-1.3	7.4	
-Livestock 3.6	2.2	3.8	3.4	4.0	3.5	2.8	4.1	
-Forestry 8.9	2.6	-0.1	4.8	1.8	6.6	-6.7	3.2	
-Fishing 8.5	2.6	1.4	-15.2	3.8	0.7	1.0	5.8	

Repeat this Selection

Growth and Investment

Table 1.2: Growth Rate (%)

Sectors/Sub-Sectors	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15 P
B. Industrial Sector	8.5	-5.2	3.4	4.5	2.6	0.6	4.5	3.6
1. Mining & Quarrying	5.2	-2.5	2.8	-4.4	5.2	3.9	1.7	3.8
2. Manufacturing	6.1	-4.2	1.4	2.5	2.1	4.6	4.5	3.2
-Large Scale	6.1	-6.0	0.4	1.7	1.1	4.2	4.0	2.4
-Small Scale	8.3	8.0	8.5	8.5	8.4	8.3	8.3	8.2
-Slaughtering	3.3	3.8	3.2	3.7	3.5	3.6	3.4	3.3
Electricity Generation & Distribution & Gas Dist.	37.2	-12.1	16.7	63.9	1.4	-26.4	5.6	1.9
Construction	15.4	-9.9	8.3	-8.6	3.1	1.1	7.3	7.1
Commodity Producing Sector (A+B)	5.1	-6.9	1.8	3.3	3.1	1.7	3.6	3.2
Services Sector	4.9	1.3	3.2	3.9	4.4	5.1	4.4	5.0
1. Wholesale & Retail Trade	5.7	-3.0	1.8	2.1	1.7	3.5	4.0	3.4
2. Transport, Storage and Communication	5.3	5.0	3.0	2.4	4.6	4.0	4.6	4.2

21. Click **Preview & Export Extracted Data**. A window appears that displays the preview of the extracted data in a structured, machine readable format. Inspect the data to make sure it looks correct. If any data is missing, you may have to slightly expand your selection. Sometimes, if headers are formatted strangely (vertically or in merged cells), you have to select the data tables without the headers and type in the column headers manually after.

Tabula My Files About Help Source Code

Pakistan Economic Survey 2015-2016 - ... Export Format: CSV Export Copy to Clipboard

Is the extracted data incorrect? You can revise your selected cells or try an alternate extraction method.

Revise Selected Cells Data has been extracted from the cells you selected in the previous step. You can revise your selection(s) to add or remove cells.

Choose Alternate Extraction Method The current preview uses the Stream extraction method. If the data is not mapped to the correct cells, try the Lattice method instead.

Stream Lattice Stream looks for whitespace between columns, while Lattice looks for boundary lines between columns.

Still look wrong? Contact the developers and tell us what you tried to do that didn't work.

Preview of Extracted Tabular Data

Table 1.2: Growth Rate (%)

Sectors/Sub-Sectors 2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15 P
A. Agriculture 1.8	3.5	6.2	2.0	3.6	2.7	2.7	2.9
Crops -1.0	5.2	-4.2	1.0	3.2	1.5	3.2	1.0
Important Crops -4.1	8.4	-3.7	1.5	7.9	0.2	8.0	0.3
Other Crops 6.0	0.5	-7.2	2.3	-7.5	5.6	-5.4	1.1
Cotton Ginning -7.0	1.3	7.3	-8.5	13.8	-2.9	-1.3	7.4
-Livestock 3.6	2.2	3.8	3.4	4.0	3.5	2.8	4.1
-Forestry 8.9	2.6	-0.1	4.8	1.8	6.6	-6.7	3.2
-Fishing 8.5	2.6	1.4	-15.2	3.8	0.7	1.0	5.8

Table 1.2: Growth Rate (%)

Sectors/Sub-Sectors 2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15 P
B. Industrial Sector 8.5	-5.2	3.4	4.5	2.6	0.6	4.5	3.6
1. Mining & Quarrying 5.2	-2.5	2.8	-4.4	5.2	3.9	1.7	3.8
2. Manufacturing 6.1	-4.2	1.4	2.5	2.1	4.6	4.5	3.2

22. From the Export Format drop-down, you can select a file format to download the extracted data in. Choose a file format to work with – including the CSV format. Keep the selection as **CSV**, and click the **Export** button.
23. A CSV file called **tabula-Pakistan Economic Survey 2015-2016 - Growth.csv** downloads on your computer. Save this file at a suitable location.
24. Now you can work with your data using any spreadsheet software – including Excel, rather than a PDF. To open this file in Excel, first launch Microsoft Excel.

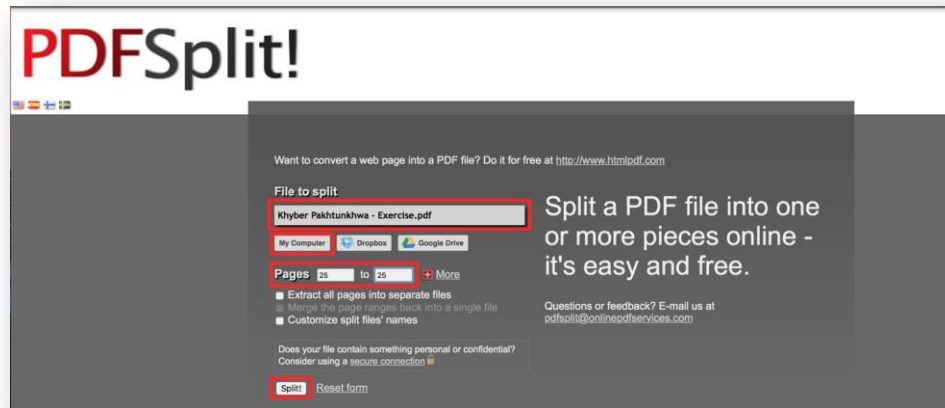
Using Tabula: Exercise

25. In the same document **Pakistan Economic Survey 2015-2016 - Growth.pdf** in Table 1.1 Comparative Real GDP Growth Rates, extract the data with Tabula
26. Any difficulty with the exercise? Remember to limit the selection and also remember that the idea of Tabula is to extract the information then clean it in Excel.

Exercise: Downloading and Extracting a Data Table from PDF

Sometime you find a data set with many tables and you need just one or some pages from a pdf document. Try the following exercise to practice how to extract a specific data table from a PDF you find online. In this example, you will download the [health institutions & their bed strength in Khyber Pakhtunkhwa](#) and work in Table No.129 district wide malaria control activities in Khyber Pakhtunkhwa.

27. Download [health institutions & their bed strength in Khyber Pakhtunkhwa](#) and rename as **Khyber Pakhtunkhwa - Exercise**
28. To extract Page 25 from this large file, go to <http://www.splitpdf.com/>
29. Upload the **Khyber Pakhtunkhwa - Exercise** to this website from your computer.
30. Select page **25** to **25**.
31. Click **Split!** The selected page downloads automatically. Save to your desktop



Extract Data to Excel Format

32. Navigate to one of these online scraping services:
 - a. www.cometdocs.com
 - b. www.zamzar.com
33. Upload the PDF file that you saved after extracting page 25, and convert it to the Excel format.
34. Review the converted file to see how complete and clean it is

Extracting Data from a Table in Image Format

Here is a scenario – The Bureau of Investigative Journalists in the UK obtained a huge number of documents detailed the Pakistan government's knowledge of US drone strikes. This is an excerpt of what the original documents looked like: <https://www.thebureauinvestigates.com/2014/01/29/get-the-data-pakistani-governments-secret-report-on-drone-strikes/>. As you can see, they are image files. The challenge for the journalists was to convert image files into structure, machine-readable data. The problem right now is that there is text, but the computer has no idea that there is text. We have to use software called Optical Character Recognition to train the computer to convert the image of text into actual computer-generated characters.

Commented [EC1]: Update this example.

Monday, May 21, 2012

HEALTHCARE FINANCING IN SOUTH SUDAN: DRILL, BABY, DRILL

While browsing through CNN or the NYTimes recently, you might've heard that there's tension brewing on the border between South Sudan and Sudan. As you probably already know, if conflict does eventually break out, it unfortunately won't be anything new for folks in this region. Ever since Sudan gained independence from Britain in the '50s, they've enjoyed only 18 years of temporary peace from civil war. That's worth repeating. In the only years since gaining sovereignty, there has only been around one decade (between the 70s and 80s) not plagued by regional strife. Recently, there seems to have been progress, though. In 2005, the Second Sudanese civil war ended with the signing of the Comprehensive Peace Agreement – and by 2011, citizens of Southern Sudan overwhelmingly indicated a desire to form their own nation in a historic referendum.

Recovering from any kind of conflict is tough for any region – but such prolonged exposure to war has had a devastating effect on the development of basic infrastructure here. Over time, this has impacted the one resource that no country can afford to sacrifice: health.

Doesn't believe me? There's a lot of evidence connecting conflict to deterioration of health systems, and consequently of health outcomes (thank you, Prof. Stephen Smith!). But, let's make this concrete. In 2007, USAID conducted a health system assessment of Southern Sudan. Here's what they found:

TABLE 3. KEY HEALTH INDICATORS FOR SOUTHERN SUDAN, SUDAN AND SUB-SAHARAN AFRICA

	Southern Sudan	Sudan	Sub-Saharan Africa
Total population (million)	Estimate range from 8 to 12 million	38	13
Life expectancy at birth (years)	47	57	48-51
Physicians (per 100,000)	2	2	2
OPV coverage	13%	78%	67%*
Under-5 mortality rate (per 1,000)	230	90	151
Infant mortality rate (per 1,000)	130	62	93
Children under 5 sleeping with mosquito nets	<1	<1	n.a.
Polio vaccination (% 12-23 months)	25	95	n.a.
Maternal mortality rate (per 100,000 live births)	1,037	190	655
Fertility rate	6.3*	4.2	5.19
Contraceptive prevalence (%)	<1%	21-4%	21-4%
Births attended by skilled attendance	6%	27%	31.7%

*Source: UNICEF, UNFPA, WHO, DHS, Demographic Yearbook 2008, UNICEF 2007, WHO 2008 and National Health Authority preliminary 2009 household survey. UNICEF, WHO, DHS, Demographic Yearbook 2008, UNICEF 2007, WHO 2008 and National Health Authority preliminary 2009 household survey. UNICEF, WHO, DHS, Demographic Yearbook 2008, UNICEF 2007, WHO 2008 and National Health Authority preliminary 2009 household survey.

Let's use Google Docs to extract data from this image. Here are the steps:

Download Image File

35. Open <http://medvocacy.blogspot.com.tr/2012/05/health-financing-in-south-sudan-drill.html>
36. The blog post opens in your web browser. Right-click on Table 2 and select **Save image as** to save this table on your computer's Desktop.

medvacy.blogspot.com.tr/2012/05/health-financing-in-south-sudan-drill.html

systems, and consequently, of health outcomes (thank you, Prof. Stephen Smith!). But, let's make this concrete. In 2007, USAID conducted a [health system assessment](#) of Southern Sudan. Here's what they found:

TABLE 2: KEY HEALTH INDICATORS FOR SOUTHERN SUDAN, SUDAN AND SUB-SAHARA AFRICA

		Sudan	Sub-Saharan Africa
Total population (millions)		8	15
Life expectancy at birth (years)		7	48.45
Physicians (per 100,000 population)			2
DPT3 coverage (%)		8%	67%**
Under 5 mortality rate (per 1,000 live births)		0	151
Infant mortality rate (per 1,000 live births)		2	93
Children under 5 under insecticide-treated bednets (%)		1	n.a.
Measles immunization coverage (per 100,000 live births)		590	855
Maternal mortality rate (per 100,000 live births)	2,037		
Fertility rate (per 1,000 live births)	6.7*	4.2	5.19
Contraceptive prevalence (%)	<1%	7%	21.4%
Births assisted by skilled attendant (%)	6%	57%	51.7%

Sources: (MCHGoSS, UNICEF 2006, World Bank 2006, UNICEF 2007, WHO 2005a) and National Health Assembly presentations *SP, MCHGoSS, USAID, JSI (2006), **WHOAFRO

37. Notice that the table saves as an image file called **USAID.png**

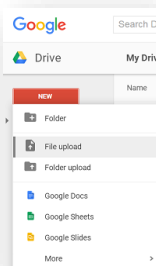
Extract Data Table using Google Docs

38. Now, open Google Drive: <https://www.google.com/drive/>

39. Click **Go to Google Drive** if needed, and log in to Google using your Gmail login credentials.

40. The Google Drive screen opens. Here, click the red button on the left that says **NEW**.

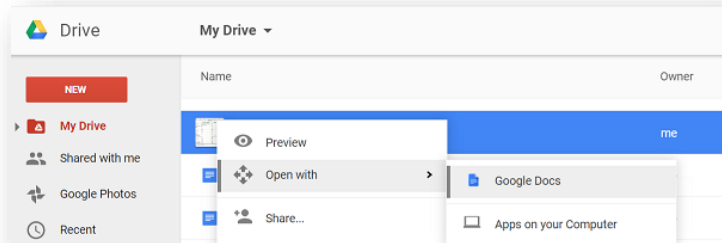
41. From the drop-down menu, select **File upload**



42. In the window that opens, browse and select **USAID.png** and then click **Open**.

43. Once the image uploads, it should appear in the list the Google Drive list

44. Right-click on **USAID.png** icon or name in the list, and select **Open with > Google Docs**



45. A Google Doc with the table and the text extracted from the table opens in a new window
46. Compare the image and the text to see how the text was extracted