

Chapter 2

Tutorial: Extracting tables from PDF files

Summary: Ideally, we want to work with tables in machine-readable formats that can be opened in a spreadsheet or database file. While the open data movement has forced public—and some private—institutions to release data in these formats, journalists must still deal with tables buried inside Portable Document Formats, otherwise known as PDFs.

As a first step, it's a good idea to ask the institution in question if you can have the table in machine-readable format. The philosophy of the open data movement that we discuss in Chapter 2 dictates that government information should be “open by default”, which should also mean providing it in a more user-friendly format such as a text file or Excel. You may succeed with moral suasion. If not, you'll have to extract the table from the PDF using a technique we discuss in this tutorial. First, a word about the PDFs themselves.

Essentially, there are three types of PDFs:

1. **Standard PDFs** retain the underlying attributes of the original document, which means they are easily converted. You can tell if a PDF is in standard format if you are able to conduct a word search.
2. **Secure PDFs** set the bar much higher because the creator, usually a government department, has enabled a security feature that prevents users from selecting text or extracting material. You should ask the institution in question to give you the user password, but don't be surprised if the answer

is no. If the information in question is in the public domain, or in the public interest, refuse to take no for an answer, or try making a formal access-to-information request, which is discussed in Chapter 3.

3. An **image PDF** is a picture of the file, which means that all the underlying attributes – the placement of letters and columns – are gone. Think of it as a digital photograph with very high resolution. You can tell if a file is an image PDF if you are unable to select words or conduct word searches.

The good news is that there are a number of online tools and software packages that allow journalists to convert tables into spreadsheet format. They are optical character recognition (OCR) programs that guess what the characters are supposed to be and then convert the words and numbers into readable characters, placing them in a tabular format of rows and columns.

Accuracy of the conversion depends on the quality of the original file. However, one of the advantages of PDFs, especially ones produced by governments, is that they are of excellent quality, meaning that the conversion is usually extremely accurate.

For this tutorial, we will use an online software tool called Cometdocs, a free, online document management system web application that converts documents at no cost. The supported conversion types include: PDF-to-Word, PDF-to-text and PDF-to-Excel. The free version of Cometdocs can handle PDFs with single- or multi-page tables. For more information about the size limitations and other important details, visit the website at: <http://www.cometdocs.com/> and read the “FAQ” section.

What you will learn:

1. How to sign up for Cometdocs.
2. How to upload, convert, and save a file.

Task 1: How to sign up for Cometdocs.

Downloadable data: We'll use the consolidated balance sheets for Blackberry's first quarter results. Click [here](#) to download the PDF (page 26).

BlackBerry Limited
 Incorporated under the Laws of Ontario
 (United States dollars, in millions)(unaudited)

Consolidated Balance Sheets

	As at	
	May 30, 2015	February 28, 2015
Assets		
Current		
Cash and cash equivalents	\$ 1,165	\$ 1,233
Short-term investments	1,799	1,658
Accounts receivable, net	470	503
Other receivables	93	97
Inventories	133	122
Income taxes receivable	16	169
Other current assets	258	375
Deferred income tax asset	8	10
	<u>3,942</u>	<u>4,167</u>
Long-term investments	293	316
Restricted cash	59	59
Property, plant and equipment, net	519	556
Goodwill	96	76
Intangible assets, net	1,281	1,375
	<u>\$ 6,190</u>	<u>\$ 6,549</u>
Liabilities		
Current		
Accounts payable	\$ 149	\$ 235
Accrued liabilities	466	658
Deferred revenue	464	470
	<u>1,079</u>	<u>1,363</u>
Long-term debt	1,550	1,707
Deferred income tax liability	48	48
	<u>2,677</u>	<u>3,118</u>
Shareholders' Equity		
Capital stock and additional paid-in capital		
Preferred shares: authorized unlimited number of non-voting, cumulative, redeemable and retractable	—	—
Common shares: authorized unlimited number of non-voting, redeemable, retractable Class A common shares and unlimited number of voting common shares		
Issued - 529,484,618 voting common shares (February 28, 2015 - 528,802,322)	2,459	2,444
Retained earnings	1,078	1,010
Accumulated other comprehensive loss	(24)	(23)
	<u>3,513</u>	<u>3,431</u>
	<u>\$ 6,190</u>	<u>\$ 6,549</u>

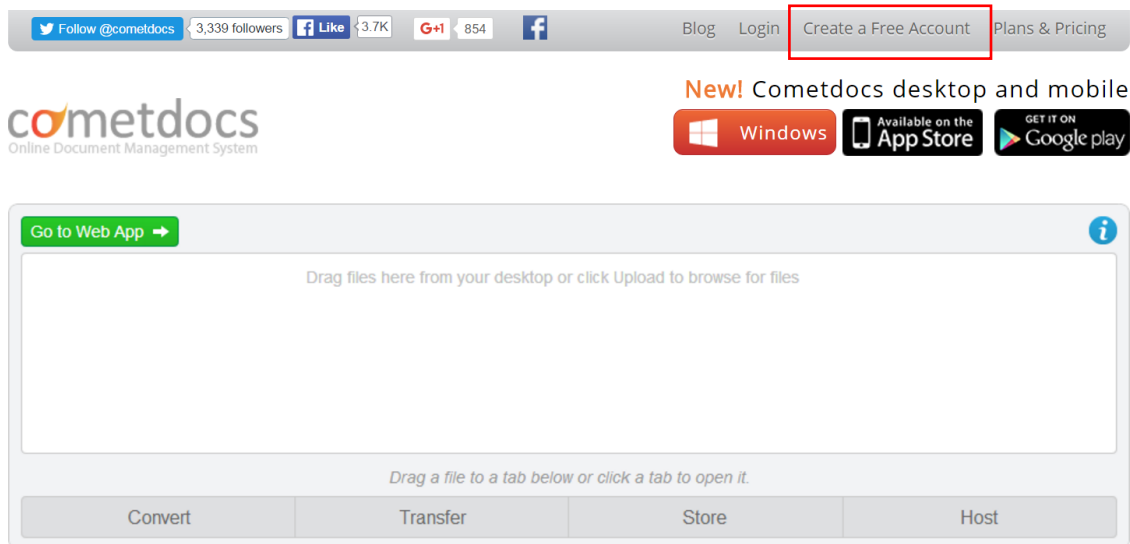
See notes to consolidated financial statements.

On behalf of the Board:

John S. Chen
 Director

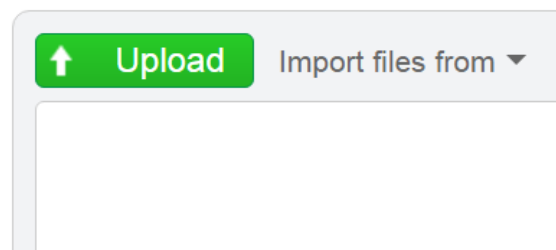
Barbara Stymiest
 Director

To sign up for the service, go the website and select the “Create a Free Account” tab on the top right-hand side.

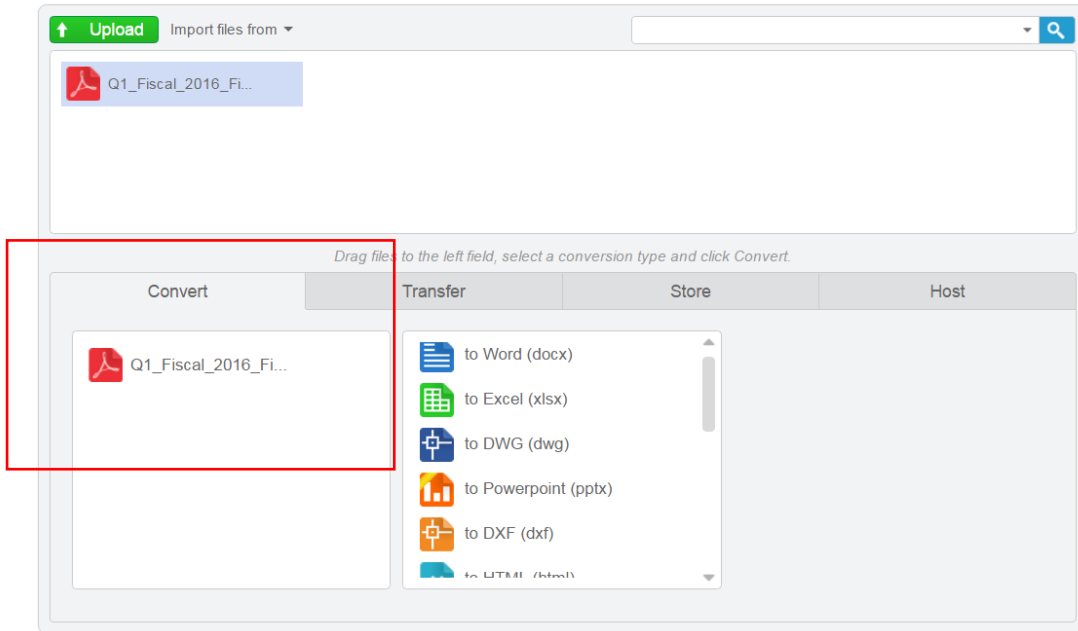


Task 2: How to upload, convert, and save a file.

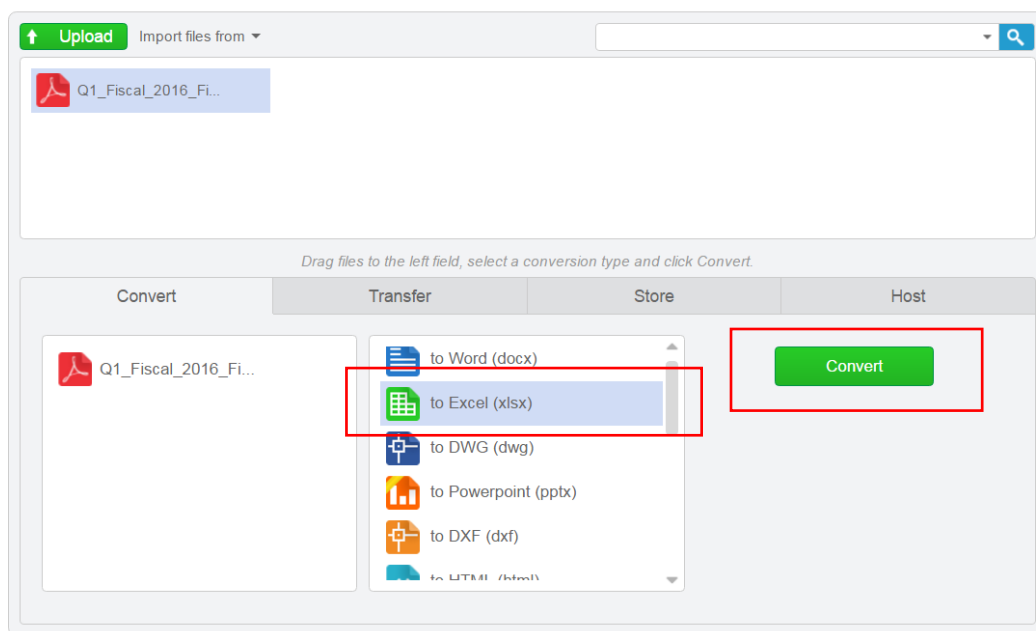
Once you’ve created your free account, click the “Upload” tab to browse your hard drive and locate the PDF.



Upload the document and drag it to Cometdoc’s “Convert” section.



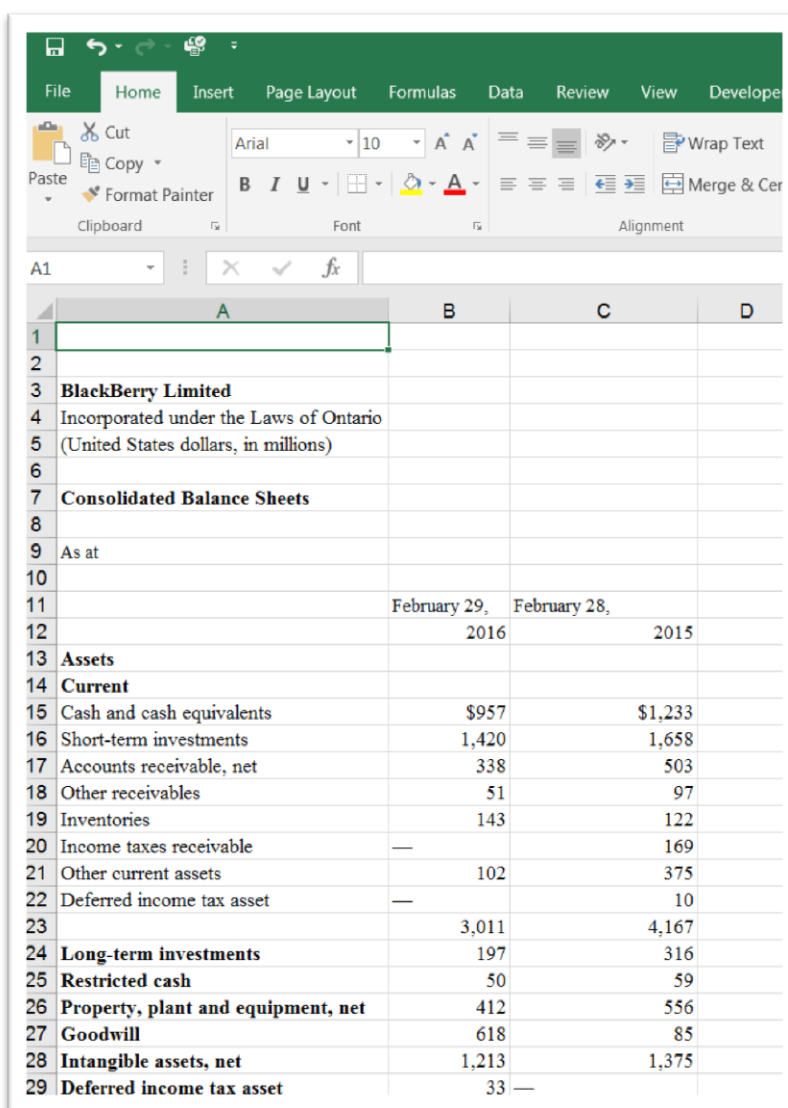
Doing so produces the options contained in the menu to your right. We want to convert the table in our PDF to Excel, so choose that option.



Click “Convert” to continue the process. Cometdocs will email you the Excel file, a process that typically takes a few minutes at the most. PDFs containing dozens of pages may take a little longer.

As with any free online tool, there are limits, which is how they convince you to pay for the more sophisticated versions that can handle larger PDFs containing dozens of tables. In this case, you can do the same. However, a better option is to become a member of [Investigative Reporters and Editors](#), a U.S.-based organization. Membership entitles you to a free account, which handles very large datasets and allows you to download the converted files directly to your hard drive.

Once you’ve downloaded your converted file, open it up and have a look.



	February 29, 2016	February 28, 2015
Assets		
Current		
Cash and cash equivalents	\$957	\$1,233
Short-term investments	1,420	1,658
Accounts receivable, net	338	503
Other receivables	51	97
Inventories	143	122
Income taxes receivable	—	169
Other current assets	102	375
Deferred income tax asset	—	10
	3,011	4,167
Long-term investments	197	316
Restricted cash	50	59
Property, plant and equipment, net	412	556
Goodwill	618	85
Intangible assets, net	1,213	1,375
Deferred income tax asset	33	—

You can see that the results are pretty good. Make a copy of this file and save it on your hard drive.

An important task is ensuring that the numbers are accurate, which involves double-checking the figures with those in the PDF. The task is easy when working with a small dataset such as this, but your integrity checks with a large one will take more time and patience. As is the case with doing any task that involves data, it's better to work with two screens, which in this case allows for quick integrity checks comparing the numbers in the Excel worksheet to those in the PDF.

With the integrity checks completed, you can analyze the data using many of the techniques we discuss in Chapter 4 and the accompanying online tutorials.