Exercise 17: Forms of Production and Cost

GOALS: The goals of this exercise are for the student to:

1. Convert evidence to PDF and TIFF with text; and

2. Assess impact of alternate forms of production in terms of impact on cost of ingestion and hosting.

OUTLINE: Students will convert a Microsoft Word document to PDF, TIFF and text formats, compare file sizes and calculate the projected cost of ingestion and monthly hosting for alternate forms of production when the cost of services is assessed on a per-gigabyte pricing model.

Producing parties frequently seek to convert native file formats used by and collected from custodian into static image formats like PDF or more commonly, TIFF images plus load files holding extracted text or text generated through use of optical character recognition. Proponents of static image productions assert claims of superior document security and point to the ability to emboss page numbers and other identifiers on page images. Too, page images can be viewed using any browser application, affording users ready accessibility to some content, albeit sacrificing other content and utility.

Often overlooked in the debate over forms of production is the impact on ingestion, processing, storage and export costs engendered by use of static image formats. Most e-discovery service providers charge to ingest, process, host (store) and export electronically stored information on a per-gigabyte basis. As a result, when items produced occupy more space (measured in bytes), they cost the recipient more to use. This exercise invites students to consider what, if any, increase in cost may flow from the use of static imaged formats as forms of production.

The Myth of Page Equivalency

It's comforting to quantify electronically stored information as some number of pieces of paper or bankers' boxes. Paper and lawyers are old friends. But you can't reliably equate a volume of data with a number of pages unless you know the composition of the data. Even then, it's a leap of faith.

If you troll the Internet for page equivalency claims, you'll be astounded by how widely they vary, though each is offered with utter certitude. A gigabyte of data is variously equated to an absurd 500 million typewritten pages, a naively accepted 500,000 pages, the popularly cited 75,000 pages and a laggardly 15,000 pages. The other striking aspect of page equivalency claims is that they're blithely accepted by lawyers and judges who wouldn't concede the sky is blue without a supporting string citation.

In testimony before the committee drafting the federal e-discovery rules, Exxon Mobil representatives twice asserted that one gigabyte yields 500,000 typewritten pages. The National Conference of Commissioners on Uniform State Laws proposes to include that value in its "Uniform Rules Relating to Discovery of Electronically Stored Information." The Conference of Chief Justices cites the same equivalency in its "Guidelines for State Trial Courts Regarding Discovery of Electronically-Stored Information." Scholarly articles and reported decisions pass around the 500,000 pages per gigabyte value like a bad cold. Yet, 500,000 pages per gigabyte isn't right. It's not even particularly close to right.

Years ago, Kenneth Withers, Deputy Executive Director of The Sedona Conference and then e-discovery guru for the Federal Judicial Center, wrote a section of the fourth edition of "The Manual on Complex Litigation" that equated a terabyte of data to 500 billion typewritten pages. It was supposed to say million, not billion. Eventually, the typo was noticed and corrected; but, the echoes of that innocent thousand-fold mistake still reverberate today. Anointed by the prestige of the manual, the 500-billion-page equivalency was embraced as gospel. Even when the value was "corrected" to 500 million pages per terabyte—equal to 500,000 pages per gigabyte—we're still talking about equivalency with all the credibility of an Elvis sighting.

So, how many pages are there in a gigabyte? It's the answer lawyers love: "*It depends.*"

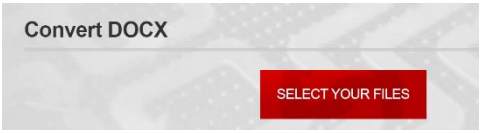
Page equivalency is a myth. One must always look at individual file types and quantities to gauge page equivalency, and there is no reliable rule of thumb geared to how many files of each type a typical user stores. It varies by industry, by user and even by the life span of the media and the evolution of particular applications. A reliable page equivalency must be expressed with reference to both the quantity and form of the data, *e.g.,* "*a gigabyte of single page TIF images of 8-1/2-inch x 11- inch documents scanned at 300 dots per inch equals approximately 18,000 pages*."

Exercise 17a: Convert Word Document to Imaged Formats

For this exercise, you will download an exemplar Word document and use free, online tools to convert the file to PDF, TIFF and plain text formats.

Step 1: Download the File. Download the file http://www.craigball.com/Always_and_Never.docx and save it to your Desktop or some other location where you can easily find it for this exercise. Should your system not permit download of Word files, you can download the file as a compressed .Zip file from here. Be sure to extract the .DOCX form of the file to your Desktop before proceeding. *You must undertake the conversion exercise using the .DOCX form of the file.*

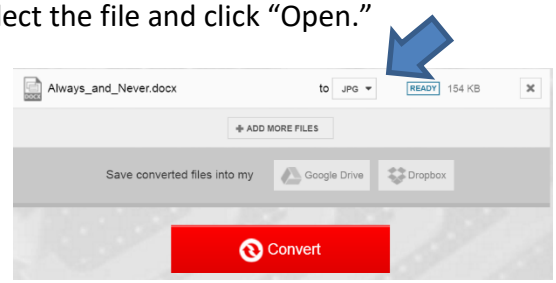
Step 2. Convert the .DOCX file to a PDF. Though there are many ways to convert a Word document to a PDF format, including by using Word itself to Save As a PDF or Print to PDF, we will use an online file converter here for consistency and simplicity.



Using your browser, go to https://convertio.co/convert-docx/ and click on the red SELECT YOUR FILES button.

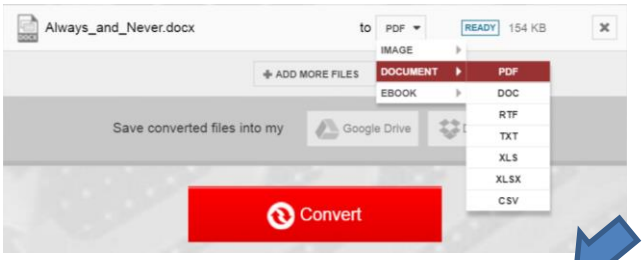
From the Select Files to Convert screen, select "Choose from Computer" then navigate to the file just downloaded called **Always_and_Never.docx.** Select the file and click "Open."

You should see the following screen:

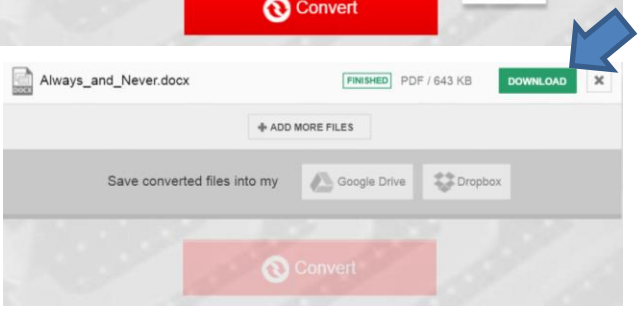


Note the pulldown menu where you may select the format for conversion (JPG in the figure at right) and select the down arrow to view options.

Select DOCUMENT and PDF from the menu and submenu (see figure at right).

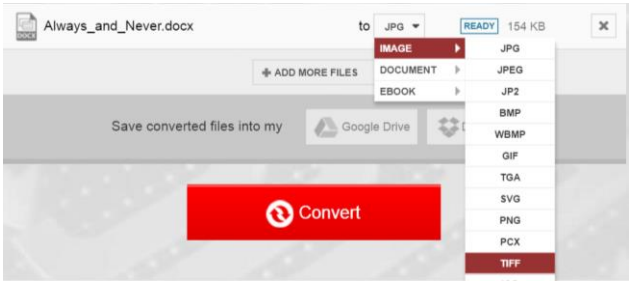


Click the red Convert button.



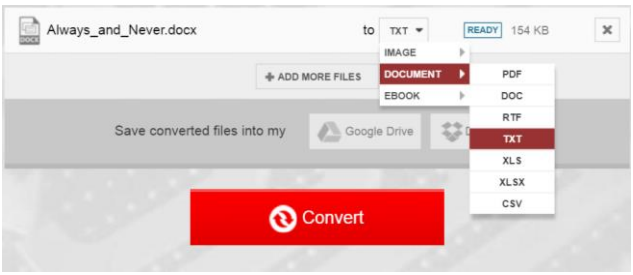
In the next screen, click the green DOWNLOAD button and save the Always_and_Never.PDF file to the same location where you saved the .DOCX file.

Step 3: **Convert the .DOCX file to TIFF images.** Follow the same steps as above, but this time select IMAGE>TIFF using the drop down menu (see image below) before clicking the red "CONVERT" button.



Click the green DOWNLOAD button again and save the file Always_and_Never.tiff to the same location where you placed the .DOCX and .PDF files.

Step 4: Convert the .DOCX file to plain text. Follow the same steps, but now select DOCUMENT>TXT from the drop down menu (see image below) before clicking the red "CONVERT" button.



Click the green DOWNLOAD button again and save the file Always_and_Never.txt to the same location where you placed the .DOCX and .PDF files.

Step 5: Record the file sizes. Navigate to the location where you downloaded the files and record their file sizes in the blanks below***. Be sure to note if the size value is expressed in units of bytes, kilobytes, megabytes or gigabytes.***

Always_and_Never.DOCX: 18.3 KB

Always_and_Never.PDF: 627 KB

Always_and_Never.TIFF: _____ MB

Always_and_Never.TXT: _____ KB

Exercise 17b: Calculate the Cost Difference Flowing from Alternate Forms of Production

There may be many variables that go into computing the cost of vendor services for e-discovery, and the charges for ingestion, processing, hosting and export are just parts of a more complicated puzzle. The purpose of this exercise is to gauge the difference that forms of production may make as a component of overall cost.

Problem: ***You are a requesting party in a federal case, and you have made a timely, compliant and unambiguous written request for production of responsive information in native and near-native forms. You have expressly requested that Microsoft Word documents be produced in their native .DOC or .DOCX formats. Your opponent instead produces Word documents to you***

as multiple .TIFF image files accompanied by a load file containing the extracted text from each document. When you object, your opponent counters that "this is what they always do" and that "TIFF plus load file is reasonably usable, so the Rules gave them the right to substitute TIFFs for natives."

Assume that your opponent has produced 1,000 different Word documents which (for ease in making the calculation) are all exactly the same size as the native and converted file sizes for the file Always_and_Never.DOCX. Assume that none of the documents are privileged or required redaction. None are hash-matching duplicates of any other items produced.

You've contracted with an e-discovery service provider to load and host the documents produced so you can review and tag the documents for use in the case. The service provider charges by the gigabyte to ingest, process and host the data month-to-month. This is the applicable fee schedule:

To Ingest and Process Data Supplied:

0 to 300 GB: $75.00 per GB

301 GB to 1 TB: $55.00 per GB

Greater than 1 TB: $40.00 per GB

Monthly Hosting Fee:

0 to 300 GB: $23.00 per GB

301 GB to 1 TB: $20.00 per GB

Greater than 1 TB: $17.00 per GB

Any fraction of a gigabyte will be rounded up to a full gigabyte when calculating charges

You intend to approach the Court to compel your opponent to produce the documents in the form you designated, and in addition to raising issues of utility, completeness and integrity, you want to determine whether the form produced to you will prove more expensive to ingest, process and host for the one-year period you expect to have the data online.

Question: If you accept the production in TIFF and load file, approximately how much more will it cost you over twelve months versus the same production in native forms?

<u>How to Solve this Problem</u>:

Step 1: Normalize the file sizes. Because the prices are quoted in gigabytes, you will want to express all data volumes in gigabytes, rather than as kilobytes or megabytes.

Remember: *A kilobyte is one thousand bytes. A megabyte is one thousand kilobytes. A gigabyte is one thousand megabytes and a terabyte is one thousand gigabytes.*

Step 2: Calculate the cost of Native Production using normalized values:

Native Production: One thousand files, each 18.3KB in zize, is 18,300KB or 18.3MB. Because the service provider's minimum charge is one gigabyte. The cost to ingest and host for one year would be:

Ingest and Process (1GB at $75.00/GB) + Hosting (1GB at $23.00/GB/month x 12 months) = $351.00

Step 3: Calculate the cost of TIFF andText Load File Production using normalized values:

TIFF Plus Production: One thousand files, each *(X) MB in size, is (X) GB,* where (X) is the size of the file Always_and_Never.TIFF*. We must also add the extracted text in the load file, which will be one thousand times (Y) where (Y) is the size of the the* Always_and_Never.TXT file. Any fraction of a gigabyte should be rounded up to the next whole gigabyte. Consequently, the value (Z) is the sum of X plus Y rounded up to the next whole gigabyte.

Ingest and Process (ZGB at *$75.00/GB) + Hosting* (Z GB *at $23.00/GB/month x 12 months) = $_____*

Examplar calculation using hypothetical values:

For example, if Always_and_Never.TIFF was 19MB in size and Always_and_Never.TXT file was 57KB in size, the calculation would be:

X = 1,000 (files) times 19MB = 19GB

Y = 1,000 (text extractions) times 57KB = 57MB = .057GB

Z = (19GB + .057GB) = 20GB (rounded up)

These values would make the calculation of the cost to ingest, process and host the TIFF Plus production:

Ingest and Process (20GB at *$75.00/GB) + Hosting* (20 GB *at $23.00/GB/month x 12 months) = $7,020.00*

The cost difference would be ($7,020.00 less $351.00) = $6,669.00.

Step 4: Calculate the difference using the <u>actual</u> file sizes obtained by your conversion of the file Always_and_Never.DOCX to TIFF and TXT.

What is the actual difference in cost comparing the native production to the TIFF plus TXT load file production?

Enter the actual difference here: $ _____