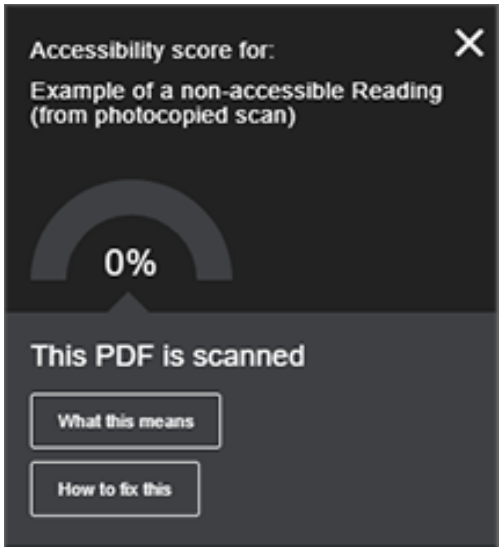
Creating PDFs that are Scanned for OCR [*Optical Character Recognition*]

A common accessibility-related issue in online course content at the college is the prevalence of **scanned PDFs**. Many faculty who upload photocopied PDFs of readings to their D2L courses receive feedback from **Ally (BB Ally)** that tells them their PDFs are "scanned" and 0% accessible.

Figure 1: BB Ally's feedback on this PDF advises that the file is scanned and 0% accessible



In this tutorial, you will learn what a "scanned" PDF means and how you can provide **more accessible*** versions of course readings either by sourcing a better digital version OR by creating "OCR'ed" PDFs instead.

(***Note**: OCR'ed PDFs are not necessarily 100% accessible; additional formatting is probably required to add structure to the document, descriptions to images, etc. The goal of this tutorial is to help you significantly improve the accessibility of your PDFs, even if more work could still be done.)

Contents

WHAT'S THE DIFFERENCE BETWEEN *SCANNED*, *OCR'ED*, *TAGGED AND UNTAGGED* PDFS? 2
OPTION 1: SEARCH LIBRARY DATABASES FOR EXISTING FILES 2
OPTION 2: REQUEST OCR'ED PDFS THROUGH PRINTSHOP SERVICES 3
Essential Practices for Good Quality OCR-Scans *3*
OPTION 3 (*IN EMERGENCIES*): USE ALLY IN D2L TO GENERATE OCR'ED PDFS FROM SCANNED 3

WHAT'S THE DIFFERENCE BETWEEN *SCANNED*, *OCR'ED*, *TAGGED AND UNTAGGED* PDFS?

- A **scanned PDF** refers to a file that was created through photocopying a source document into PDF format and in which the photocopy is effectively just a picture of the original document. In a photocopy-scanned PDF, the text cannot be extracted by text-to-speech tools and is therefore completely unavailable to any student who requires that support.
- An **OCR'ed PDF** refers to a file that has been scanned for "Optical Character Recognition" before being saved as PDF. In other words, text has been extracted from the scanned source material and it can be understood by text-to-speech tools.
- An **untagged PDF** is likely also an OCR'ed PDF. The OCR scanning process is able to extract the text from a document but it does not re-create the structure of the original document. "Tags" refer to the structural components of a document, such as headings, paragraphs, tables, and links.

  An untagged PDF will likely be readable by a text-to-speech programs (e.g. **docReader**), but it is missing important information that would make it truly accessible for all students. *For example*, the lack of document structure will mean that a student who relies on a screen-reader (e.g. JAWS), or a student who navigates the screen using keyboard shortcuts will not be able to navigate through the PDF, jump to a specific section of the document, or access context about any images in the document.
- A **tagged PDF** is the most accessible type of a PDF.
  If you create a PDF from your own Word or PowerPoint files, much of the structure you add to the original file will become the heading, paragraph, image description, etc. tags in the PDF version. (For more guidance on document structure, see: Accessibility Checkpoints for WORD DOCUMENTS).

  Tagging an untagged PDF re-creates the missing document structure and will do a lot to improve the accessibility of the file. It is possible to use **Adobe Pro DC** to edit an OCR'ed PDF and add tags to improve its accessibility.
  (The steps for tagging a PDF using *Adobe Pro DC* are out of scope this tutorial.)

OPTION 1: SEARCH LIBRARY DATABASES FOR EXISTING FILES

If you have not already done this, try searching the college's library databases to determine if there is already a good-quality digital version of the article available through these resources that you could provide to your students.

Note: PDFs that are available through the databases may already be Tagged PDFs, which are the most accessible type of PDF file.

OPTION 2: REQUEST OCR'ED PDFS THROUGH PRINTSHOP SERVICES

Before you spend hours standing at a photocopy machine, scanning your course readings (e.g. articles and book chapters) to create inaccessible PDFs, consider contacting the Printshop for help with this task.

While the mandate for Camosun's Printshop is to scan materials for printing purposes, they can also scan your course readings for OCR (Optical Character Recognition) – **upon request**.

"We can do it quite easily, just ask!" (*Printshop Services*)

Note: the Printshop is **not set up to do any tagging** of OCR'ed PDFs.

ESSENTIAL PRACTICES FOR GOOD QUALITY OCR-SCANS

When submitting original materials to the Printshop to be OCR-scanned:

1. **Select clean copies** of all source materials (i.e. unmarked originals, without ANY handwritten or marginalia notes, or underlining and highlighting).
2. Select source materials with **high-contrast** between background colour and foreground text.

If the source materials are shadowy or grainy, or the pages are wrinkled, or they include lots of underlining, highlighting, or margin notes, the quality of the OCR-scanned PDF will suffer. The OCR-scanning process simply interprets the picture it sees and converts it to text; if the empty spaces on the original page are not clean, the OCR process may try to interpret shadows and graininess and marginalia as text and you will be left with some gibberish mixed into your actual content.

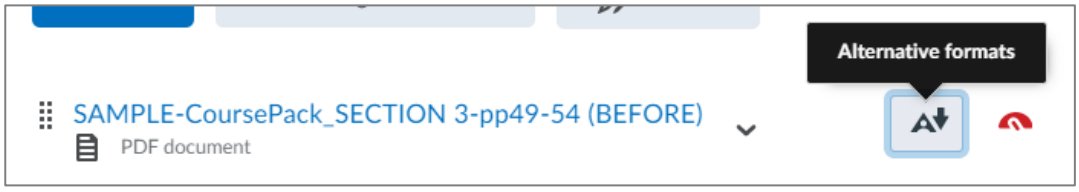
Rule of thumb: The cleaner your source documents are, the better the OCR'ed PDFs will be.

OPTION 3 (*IN EMERGENCIES*): USE ALLY IN D2L TO GENERATE OCR'ED PDFS FROM SCANNED

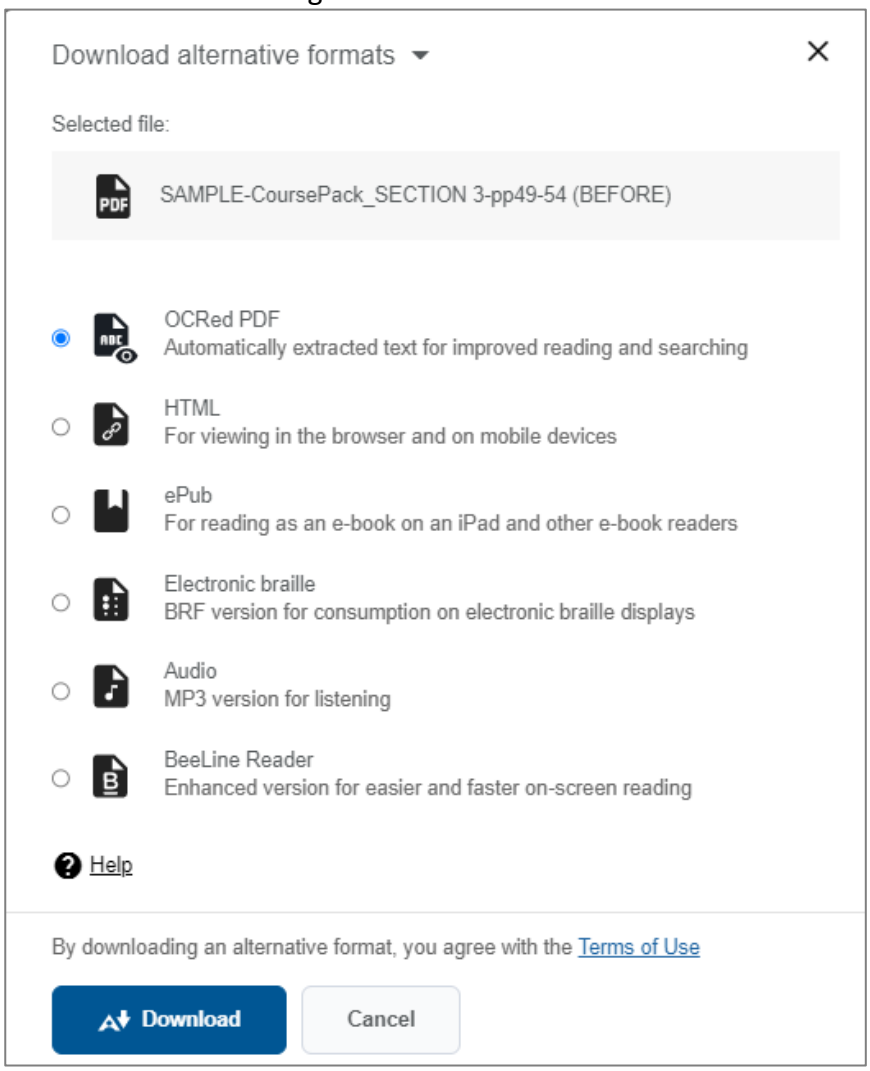
Option 2 will walk you through the steps of using ***Ally*** to generate an OCR'ed PDF from a scanned PDF in your course that you can post for students instead. If you have already created and uploaded PDFs to your course and do not have time to reconnect with the Printshop to turn a ***scanned*** PDF into an ***OCR'ed*** PDF, this option might work in a pinch.

Same rule of thumb: the cleaner the original scanned PDF is, the better the OCR'ed PDF will be. If the scanned PDFs on your course site are "dirty", the OCR'ed versions will probably include some gibberish-text where the OCR process interprets shadows and static as characters of text.

1. In your D2L course **Content**, locate the file and the ***Ally*** **Alternative Formats** menu.

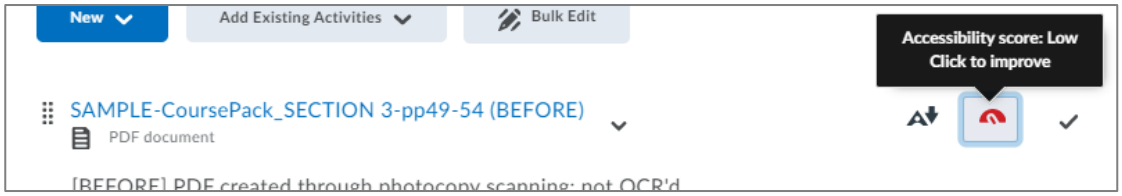


2. Scanned PDFs will include the option to download an **OCRed PDF**; select this option and click Download. This might take a minute or two.

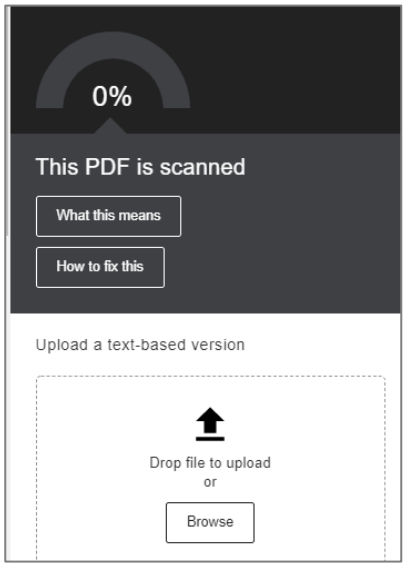


3. **Quality review**. After OCR'ed PDF has finished downloading, open the file and review the pages to ensure the quality of the text is good and that the OCR-process has not misidentified text from the original.
   If you are satisfied with the quality of the OCR'ed PDF, proceed to the next step.

4. Replace the scanned PDF with the new OCR'ed PDF. In your D2L course **Content**, locate the original file again and this time click on the ***Ally*** **Feedback** icon beside the Alternative Formats menu.



5. In the Feedback window, locate the "Browse" button and use it to locate and upload the new OCR'ed PDF.



6. Once the new file has been uploaded and re-reviewed by ***Ally***, you will see a new and improved Accessibility Score. (Now that the text is readable, you will probably also see new accessibility feedback from ***Ally*** and recommendations for other fixes you can make at some point. ☺)