# Making inaccessible text accessible

*Books are mostly available in print or in different digital formats such as PDF or publisher-specific formats like Kindle. In this course you will learn how to extract the book text from these types of files and transfer it to Microsoft Word for initiating the creation of accessible books.*

## Introduction

The standard printed book is not accessible to many persons with disabilities. They need it in other accessible formats like EPUB or Braille. Let us now see what is the standard process for the conversion of a printed book to a digital format.

Make sure that for any of the reproduction and transformation techniques discussed in this course you (or your organization) respects the relevant copyright laws in your jurisdiction. These legal conditions may vary from country to country, and they will depend notably on whether the provisions of the Marrakesh Treaty are implemented in your country or whether national law permits the reproduction, distribution, and making available to the public of Accessible Format Copies of Works for the benefit of people who are print disabled. Also note that some countries only allow the reproduction and transformation of printed works that are not commercially available in accessible formats.

If the national law in your jurisdiction does not have provisions allowing the reproduction, distribution, and making available to the public of Accessible Format Copies of Works for the benefit of people who are print disabled, you will need to request permission from the copyright owner before being able to produce an Accessible Format Copy of a Work.

The books are generally scanned and then converted to digital text using OCR (Optical Character Recognition) applications. The text is retained in Microsoft Word or similar word processors.

If the OCR is not available for a particular language, then the only way to convert the book to digital format is to type the whole book.

Fortunately, OCR software for the most commonly used languages including those used in low and middle-income countries has been released. Support for these languages has also been added to the Braille translation software such as Duxbury. Thus, the need for manual typing or direct Braille input has reduced considerably. Digital files of books are likely to be available from other sources as well and scanning or typing may not be necessary.

There is some movement towards inclusive practices in the publishing industry. Inclusive Publishing involves adopting those digital standards which are accessible for everyone and conversion to alternate format if required is simple and quick.

Until such standards are fully adopted, you will continue to get files in many formats and in multiple non-Unicode fonts from publishers. You are likely to get publications in the following formats that will need to be made accessible:

* Printed book (hard copy)
* PDF files with an image of every page of book
* Untagged PDF file that does not copy properly to Braille translation software

There are tools and methodologies that can help in converting source file in any format to a format useful for accessible books production or Braille printing.

A word processing tool such as Microsoft Word, LibreOffice Writer, Google Docs, or Apple Pages is required for preparing the document for Braille production. In this course, you will learn to use Microsoft Word for Braille production. However, note that the concepts and techniques explained in this course can be applied to other word processors as well.

After preparing the document in Microsoft Word, an EPUB version can be created using the WordToEPUB tool or the same document can be imported into the Braille production tool. All the text formatting done in Microsoft Word is useful for EPUB creation as well as Braille production.

In this course, we will concentrate on methods and tools to create an editable Microsoft Word document from hard copy print, PDF, and EPUB files so that it is ready for both braille and accessible EPUB production.

Thus, the key objectives of this course are:

* To introduce various sources of digital files that can be utilized for accessible book production
* To introduce tools and methods that will help in the conversion of digital files available from online libraries, ebook stores, and publishers to Microsoft Word and Unicode-based font.

## Tools required for this course

* You need only Microsoft Word to practice the concepts and assignments given in this course. Microsoft Word is a part of the Microsoft Office package and its various versions such as Office 2019, Office 2016, Office 2013, Office 2010 are in wide use. Most likely your computer will be having one of the Office versions. If not, you need to install it, the newer versions are preferred although you can complete this assignment using any version of Microsoft Word. To know more you can visit [Microsoft website](https://www.microsoft.com/en-in/microsoft-365/microsoft-office).
* Scanner: You need to have a scanning device to convert print material to digital format. Information about various types of scanners is included in this course. Any kind of scanner or even a Smart Phone can be used to perform the exercises given in this course.
* OCR and other conversion tools: Depending upon the requirements of your organization you can use any of the tools described later in this course such as ABBYY Fine Reader, Codex, DAISY Pipeline and Google Docs.

## Sources of digital files of publications

When you receive a request for braille conversion of a particular book, it is highly recommended that you check if a digital version of the book is available on any digital platform. It will help if you search online libraries like ABC Global Book Service and Bookshare, commercial platforms like Amazon Kindle, or request a digital copy directly from the publisher.

The possible sources of digital files are summarized below.

1. All publishers prepare a digital version of the book first and then send it to be printed. This digital version can often be utilized directly to create an accessible format version.
2. Another possible source is online book stores selling digital books for reading on computers, smartphones, and tablets.
3. Government textbook development agencies are increasingly sharing digital versions of their books on their websites in PDF or other formats.
4. You can also get digital files of books from special online libraries such as the ABC Global Book Service and Bookshare.

Millions of digital files can be used to prepare accessible books. This can considerably reduce the time and effort required in accessible book production. Manual typing of complete books or scanning and then proofreading is very time-consuming. When a very time-consuming process is adopted, it reduces the capacity and efficiency of braille production but also affects the learners as they do not get their required books on time. Therefore, it is always advisable to check on the availability of a book in digital format from online resources or from the publishers before starting manual typing or scanning.

If the book to be converted into braille is already available in digital format from any source, then you just have to download, convert, format, and emboss the book. This process takes very little time in comparison to manual typing, scanning, and editing of the book. For example, you can prepare a 200-page book for braille production in one day if a digital version is available.

## Converting a print book to a Word document

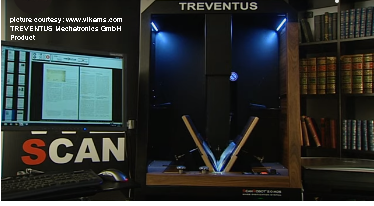
There are two steps involved in the conversion of a hard copy book into the Microsoft Word format:

* Scanning (creating a digital image of every page of the book)
* OCR – Optical Character Recognition (converting an image of the page to editable text and exporting it to Microsoft Word format)

Now let us have a look at these steps in detail.

### Scanning:

Scanning is the process of creating a digital image of a printed page. This is performed using a device called a scanner, which is attached to a computer. The digital image of the page is saved on the computer. Some scanners can also be attached to a mobile phone through a cable, Bluetooth, or over Wi-Fi. There is a wide variety of scanners available in the market. Such as:

* **Flatbed scanners** for individuals or small organizations, typically costing USD 50 or more. In such scanners, each page of the book has to be manually turned. It takes about 30 to 40 seconds to scan one page. Typically, it would take about 2 to 3 hours to scan a 200 to 250 pages book. One disadvantage of some flatbed scanners is the inability to flatten the page on the scanning surface when the book is thick. You may prefer flatbed scanners with scanning glass which continues until the edge as shown in the picture below.With a scanning surface that reaches the edge, the spine of the open book can be easily placed over it. This flattens the page to be scanned on the scanning surface.
* **Scanners with automatic document feeders (ADF)** can be priced from USD 400 to USD 1000. When using ADF scanners the book spine needs to be cut, with all pages placed in the scanner paper tray. The ADF scanner takes one page at a time and scans the whole stack by itself. A full-duplex ADF scanner will scan both sides of the page simultaneously. A 200-page book could be scanned in 5 to 7 minutes.Note that when you use an ADF scanner, you will need to cut the spine of the book and then after scanning is complete bind/file/stitch the book otherwise it will be unusable for reading.
* **Camera Scanners** consist of a digital camera placed on a stand. The open book is placed under the camera. The camera takes photos of the open page. The curvature that gets created in the center of an open book (between two pages) often creates problems in capturing good quality photos of the pages.
* **High-Speed Scanners:** If the need is to scan a large number of books, as in a university library, without cutting the book spine, then a high-speed professional book scanner can be used.These professional scanners have the capability of turning pages themselves and can scan between 1500 and 2000 pages per hour. The cost of such professional scanners is in the range of USD 100,000 to USD 120,000 and well-trained technical human resources are also required to operate and maintain them.
* **Smartphone apps:** On the other end of the spectrum, a smartphone could also work as a scanner. There are dedicated apps for scanning using the phone camera. Additionally, cloud storage apps such as Google Drive, Dropbox, and OneDrive have OCR features.

**The video below is about different types of scanners.**

[**https://player.vimeo.com/video/469605388**](https://player.vimeo.com/video/469605388)

[Video transcript for Scanners](https://dl.daisy.org/projects/training/VideoTranscripts/Scanners_EN.txt)

### Optical Character Recognition (OCR)

Once a scanner is attached to the computer, the scanning process is done using OCR software such as ABBYY Fine Reader, Omnipage, Kurzweil, or Open-book. After the image is captured, we get the option to extract text from the image. Instead of extracting text from every page, we can scan multiple pages and then begin the text extraction process in one go. If an ADF scanner is used, then images of the whole book are captured before running the text extraction process. Text extraction is possible only if the language of the book is supported by the OCR software being used. After the text is extracted, we can export the text as a Microsoft Word document.

**Watch this video to see the demonstration of the whole process.**

[**https://player.vimeo.com/video/469605454**](https://player.vimeo.com/video/469605454)

[Video transcript for OCR](https://dl.daisy.org/projects/training/VideoTranscripts/OCR_EN.txt)

Although ABBYY FineReader OCR supports more than 200 languages, many  Asian and African languages are not yet supported. In such a case, you should check if your language is supported in other OCR solutions such as the Google OCR. If it is, you can scan the book using FineReader or any other scanning utility and save the book as image PDF files. See the next section on how to extract text from image PDF files using Google OCR, an online OCR tool.

### Dealing with Image PDF files

Documents like agreements and meeting notes are generally scanned and saved as image PDFs. Often, image files are also received from publishers. One of the methods of converting such an image file to a Word file is to use ABBYY FineReader OCR, a popular tool with which you may be familiar. However, ABBYY Fine Reader doesn’t support many Asian or African languages such as Hindi or Amharic. Another way to extract text is to use Google OCR for converting image files. Google Drive is a file storage and synchronization service created by Google. It allows users to store files in the cloud, share and edit documents. With two simple steps it converts PDF files to Microsoft Word by performing OCR as follows:

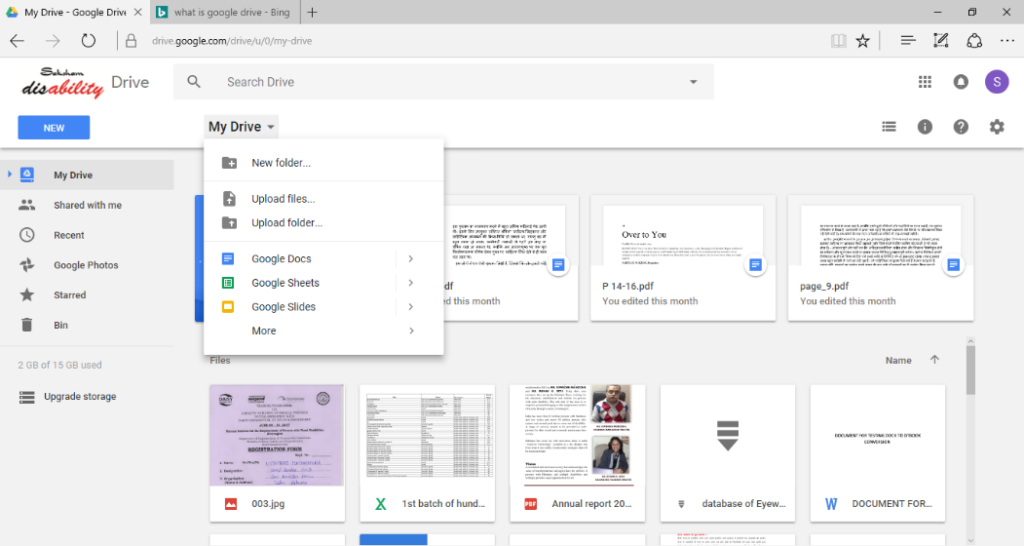
1. Open Google Drive by going to [https://drive.google.com](https://drive.google.com/) and sign in using your Google/Gmail account. Click on the My Drive button, then click on Upload Files and select the PDF files you want to convert. Note that there is a file size limit of 2 MB or 10 pages and you will have to choose files accordingly.

Fig: Google Drive Upload

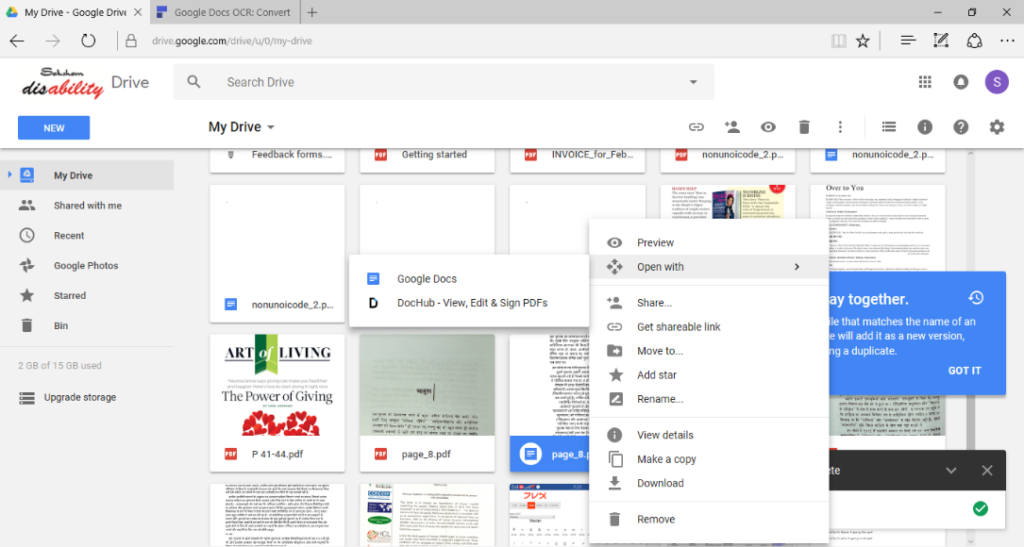
1. Once the file is uploaded, you will be able to locate it on your Google Drive page. Right-click on the file and select Open With Google Docs. Your PDF file will then be converted to an editable text document after using the Google OCR service.

Fig: Google Drive OCR

1. The converted Google Docs document can now be downloaded as a Microsoft Word document and saved on your local drive. This process is explained in the video below.

[**https://player.vimeo.com/video/469605517**](https://player.vimeo.com/video/469605517)

[Video transcript for Google Docs OCR](https://dl.daisy.org/projects/training/VideoTranscripts/Google_Drive_EN.txt)

Google Drive OCR can also handle multi-lingual and multi-font documents. If a page has text in multiple languages, the OCR will often detect those languages and convert the text to become editable. A restriction of 10 pages or 2 MB in file size is applied for using the free of cost service, which is offered to all Google account holders. Google also offers its OCR service for businesses, enabling them to create products which embed the Google OCR process without limitations. An example of such a product is called Accessital, which offers accessible book production workflow management, where many volunteers are involved in proofreading scanned books.  Accessital offers Google OCR service without any page or size restrictions as part of their subscription. More information is available at  
<https://iaccessible.net/accessital>

## Conversion of documents from different formats to Microsoft Word

Digital version of book may be available from different sources in the following formats:

1. PDF
2. InDesign or PageMaker
3. Digital book store specific proprietary format such as Kindle
4. EPUB
5. DAISY text only

You need to convert these documents to Microsoft Word format before it can be made suitable for braille translation software like Duxbury or to create accessible EPUB files. Various conversion processes and the required tools are explained below.

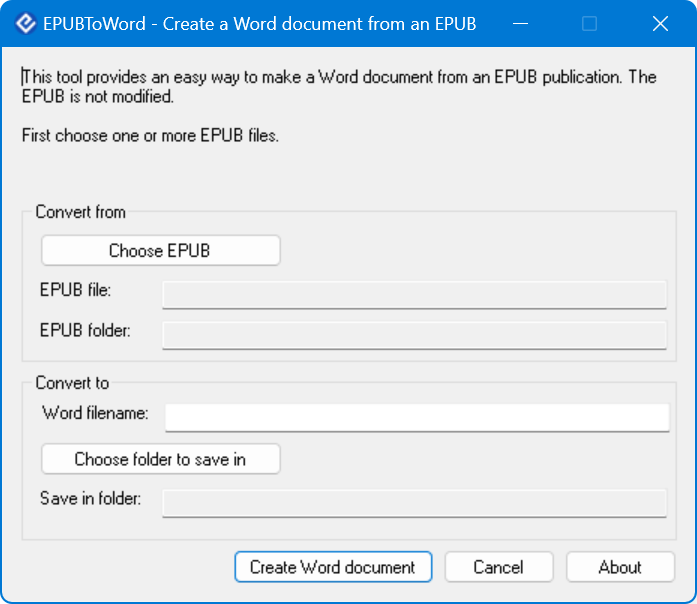
### EPUB to Word

WordToEPUB, the free tool from DAISY Consortium which is used to convert Word documents to EPUB contains a easy to use converter for the reverse conversion, EPUB to Word.

[Download and install WordToEPUB](http://www.daisy.org/wordtoepub) from the DAISY Consortium website. Then you can launch EPUB to Word converter in two ways:

1. Open the WordToEPUB installation directory which generally is at C:\Program Files (x86)\DAISY\WordToEPUB. Here locate the file named “EPUBToWord.exe” and double click to launch it.
2. Alternatively, you can right click on any EPUB file and choose “Convert with EPUBToWord”.

The following EPUBToWord window will be displayed.



Here you just need to make sure that the EPUB file to be converted is selected. You can also provide a name for the Word document that will be created and a folder to save it. When the convertion is complete, a Word document is created with all content from the EPUB. It retains the formatting and images with alt text.

### DAISY Text to Word

A DAISY text-only book can be converted to RTF using DAISY Pipeline.  DAISY Pipeline is an open-source software tool provided by DAISY Consortium and contains scripts for different file format conversions.

The RTF(Rich text format) retains much of the formatting including heading styles, which is very useful for the subsequent creation of a structured Microsoft Word document. Below is an explanation for using the DAISY Pipeline:

1. Open DAISY Pipeline
2. Go to file – new – pipeline job
3. Go to Create and modify – dtbook to RTF
4. Browse and select the source file and destination folder and click on the finish button
5. Go to the Run menu, click on Run

This will initiate the conversion process which is generally very quick. Now you can check your output folder for the RTF file.

To know more about the DAISY Pipeline software and to download it, visit: [www.daisy.org/pipeline](http://www.daisy.org/pipeline).

### PDF to Word conversion

When a PDF file with editable text is available then the best and easiest way to convert it to Microsoft Word format is to directly open the PDF file in the latest version of Microsoft Word. Newer versions of Microsoft Word allow the opening of PDF files. A confirmation message is displayed prior to conversion.

However, if you are using an earlier version of Microsoft Word which does not support the opening of PDF files, you can use the processes described below.

### PDF to Word conversion using Adobe Acrobat Pro

Acrobat Pro is a commercial software application from Adobe for creating and editing PDF documents. This tool also provides import and export facilities for several other formats, including Microsoft Word. The steps listed below need to be performed for this conversion:

1. Open Adobe Acrobat Pro
2. Open the PDF file which is to be converted
3. Use the Export option and choose Microsoft Word

You will now get a Word document that retains much of the formatting and structure contained in the PDF file.

### Online conversion of PDF to Word

Smaller PDF documents can be converted using online conversion services provided by many websites like pdf2doc.com, pdfonline.com, and freepdfconvert.com. You need to upload the files to be converted and very quickly these websites will allow you to download a Microsoft Word version of the content.

### InDesign or Page Maker files to Word conversion

Follow these steps to convert InDesign files

1. Open Adobe InDesign
2. Go to the File menu, click Open, and choose the file to be converted. You can also select PageMaker files for conversion.
3. Go to the File menu and click on Export
4. When the conversion is complete, you will get the PDF file.
5. Now open the newly created PDF file in Adobe Acrobat.
6. Go to the File menu and select Export and select Microsoft Word.

## Assignment

Download the image PDF file using the link below and convert it to Microsoft Word document with editable text using Google Drive and Google Docs.

[Image\_PDF\_to\_word\_conversion\_assignment](https://dl.daisy.org/projects/training/Image_PDF_to_word_conversion_assignment.pdf)

Check if the resulting Word document contains all text from the image PDF file. Try to format the Word document in such a way that it resembles the PDF file.

## Questions for Making Inaccessible Text Accessible module Quiz

1. **An ADF scanner is very good for scanning bound books[Points: 1]**
   * True
   * False

1. **OCR software such as ABBYY Fine Reader, and OmniPage can be used for:[Points: 1]**
   * Converting text to braille
   * Extracting text from an image
   * Capturing image of page from scanner
   * Converting text to audio

1. **Which tool is used to convert EPUB documents to Microsoft word?[Points: 1]**
   * EPUB2docx
   * Codex
   * Convertex
   * Adobe Acrobat Pro

1. **In Office 365, we can directly open PDF file in Microsoft Word.[Points: 1]**
   * False
   * True

1. **You can download a Google Docs document in the Microsoft Word format[Points: 1]**
   * True
   * False

1. **Codex is used for[Points: 1]**
   * Converting EPUB to Word
   * Creating DAISY format books
   * All of these
   * Converting EPUB to DAISY

1. **What is the use of DAISY Pipeline?[Points: 1]**
   * Create RTF from DAISY text books
   * Convert Word files to DAISY format
   * Convert DAISY books to PDF
   * Record DAISY books

## Quiz Answers for Review questions for making Inaccessible Text Accessible module

**Question 1:** An ADF scanner is very good for scanning bound books

**Correct answer:** False

**Question 2:** OCR software such as ABBYY Fine Reader, and OmniPage can be used for:

**Correct answers:**

* Extracting text from an image
* Capturing image of page from scanner

**Question 3:** Which tool is used to convert EPUB documents to Microsoft Word?

**Correct answer:** Codex

**Question 4:** In Office 365, we can directly open PDF file in Microsoft Word.

**Correct answer:** True

**Question 5:** You can download a Google Docs document in the Microsoft Word format

**Correct answer:** True

**Question 6:** Codex is used for

**Correct answer:** Converting EPUB to Word

**Question 7:** What is the use of DAISY Pipeline?

**Correct answer:** Create RTF from DAISY text books