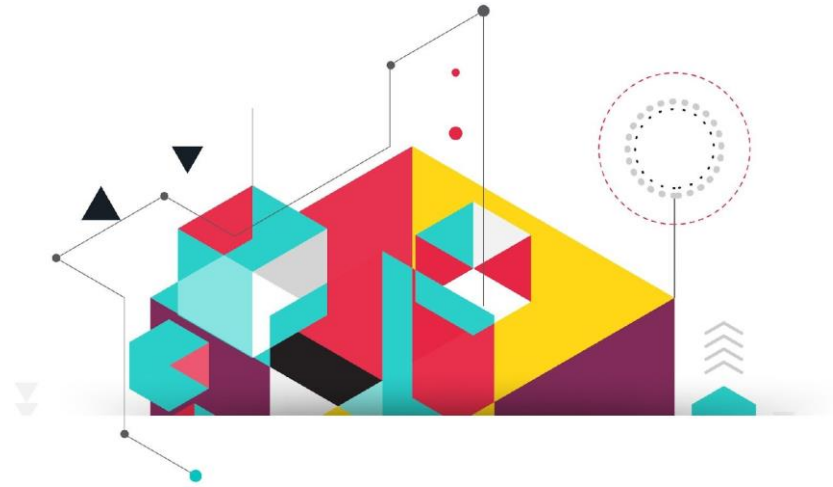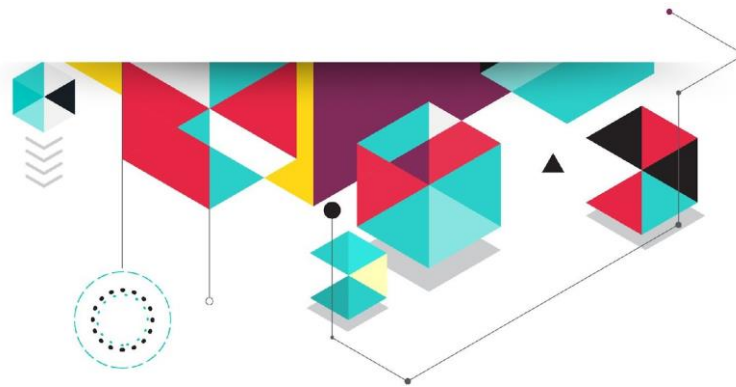LESSON 10 – PDF Automation – RECAP

Overview

In this lesson you have learnt about the types of PDF documents and the available methods for extracting data from such files. We also looked into anchors, a way to deal with unstable selectors.

Takeaways

- You can place PDF activities into 2 categories: one for when processing large chunks of text or whole documents and one for when focusing on specific text elements.

- When looking to extract data from PDF, depending on your file you should choose one of these 2 activities: **Read PDF Text** and **Read PDF With OCR.**

- Both activities can run in the background.

- Another method of grabbing blocks of text is the **Screen Scraping** tool.

- When looking to extract a certain value from PDF files, you can also use **Anchor Base.**

Best practices

- Use **Read PDF Text** instead of **Read PDF With OCR** when possible since OCR is error prone.

- The **Anchor Base** method can be more reliable than the others since it can handle major structural changes in the file.

Useful links

PDF Data Extraction