# **Encodings for PDF Services**

The UTF-8 encoding, which is widely used is well handled by PyMuPDF library used in the project.

PyMuPDF can handle the PDF text extraction well, but can fail for other encodings and corrupted characters.

## **Handling**

For other encodings and edge cases, if the PyMuPDF default encoding decoder fails to handle those:

* It could be thrown as error, showing the PDF is not supported
* Or handled, using different encoding types.

## **Edge Cases**

**1.** PDFs with OCRed images or texts in form of Images: These would be more of character recognition, than encoding problem. (High demand technically)

**2.Encodings:**

ISO 8859-1 : Latin-1 or Western European  
ISO 8859-2 : Latin-1 or Eastern European  
ISO 8859-3 : Latin-3 or South European  
ISO 8859-4 : Latin-4 or North European  
ISO 8859-5 : Latin/Cyrillic  
ISO 8859-6 : Latin/Arabic  
ISO 8859-7 : Latin/Greek  
ISO 8859-8 : Latin/Hebrew  
ISO 8859-9 : Latin-5 or Turkish  
ISO 8859-10 : Latin-6 or Nordic languages  
ISO 8859-13 : Latin-7 or Baltic Rim  
ISO 8859-15 : Latin-9 or Western European

ISO 2022-JP : Japanese  
ISO 2022-JP-2 : Japanese  
ISO 2022-KR : Korean

**Unicode / ISO/IEC 10646**

UTF-7  
UTF-8  
UTF-16  
UTF-32  
UTF-32BE