To structure data from PDF files into a format suitable for machine learning training, follow these steps:

1. **Extract Text from PDFs**: Use a library like PyMuPDF (also known as fitz) or PyPDF2 to extract text from each PDF file.
2. **Clean and Preprocess Text**: Clean the extracted text to remove any unwanted characters, normalize whitespace, and handle punctuation.
3. **Split Text into Snippets**: Depending on the use case, split the text into manageable snippets. This could be by sentences, paragraphs, or fixed-size chunks.
4. **Label the Data**: If your task is supervised, ensure each snippet is labeled appropriately.
5. **Store in a Structured Format**: Save the structured data in a format like CSV, JSON, or a database, where each entry is a text snippet with its corresponding metadata (e.g., source, label).