

Hello

ParsedDocument:

- doc_id (стабильный хэш от содержимого)
- source: имя файла, путь, размер, даты (если доступны)
- content_units: список блоков (text/table/image_caption)
- metadata: язык, автор, заголовок, количество страниц/листов, ошибки/предупреждения
- chunks: готовые к RAG куски (или отдельный шаг chunker'a)