Contribution ID: **4** Type: **not specified**

# Ingestion, chunking, metadata and embeddings for RAG

*Tuesday, 7 July 2026 12:10 (25 minutes)*

This methodology session focuses on preparing knowledge so that it can be retrieved accurately by a RAG system. It explains why document preparation is not a minor preprocessing task but a core design decision that affects downstream retrieval and answer quality. Participants will learn how text can be extracted from PDFs, Markdown, HTML or structured files while preserving headings, sections, tables, article numbers, source identifiers and other traceable information. The session then introduces chunking strategies, including fixed-size, overlapping, semantic and structure-aware chunking, and discusses how chunk size influences retrieval precision and context completeness. It also covers metadata fields that support filtering, citation and source attribution. Finally, the session explains embeddings and vector indexing, with emphasis on multilingual or Greek-aware representations, index choice, latency, scaling, filtering and maintainability in prototype and production RAG settings. These foundations prepare participants for the following indexing and retrieval implementation exercises in Colab, including practical debugging.

**Presenter:** DROSATOS, George (ATHENA RC)