# Multimodal Data Management

Course: Algorithms, Data Structures and Databases

Authors:

Edwin Santiago Delgado Rospigliosi

Toni

Project Title: Medical Visual Assistant (Educational)

Date: October 2025

## Introduction and Context

Modern data systems increasingly need to handle heterogeneous data types — images, text, audio, and video — in a unified way. Managing such multimodal data requires robust data pipelines capable of ingestion, transformation, validation, and embedding generation for analytics and AI applications.  
  
This project implements a multimodal data management pipeline that integrates textual and visual medical data (e.g., X-ray images and short radiology notes). The goal is to build a scalable data backbone that supports DataOps principles (traceability, quality, automation), cross-modal analytics (text ↔ image search), and generative AI capabilities (image captioning or summarization).  
  
The solution is developed as part of the Algorithms, Data Structures and Databases course, focusing on designing efficient data flows and retrieval structures rather than diagnostic accuracy.

## Project Objectives

1. Design a multimodal data management system supporting at least two data modalities.  
2. Implement a data pipeline structured into four zones (Landing, Formatted, Trusted, Exploitation).  
3. Integrate MinIO for storage and ChromaDB for semantic search.  
4. Execute three analytical tasks:  
 - Same-modality similarity search (image → image)  
 - Multi-modality similarity search (text → image)  
 - Generative captioning (image → text)  
5. Ensure full automation and reproducibility of the process via orchestration scripts.  
6. Produce a concise report describing design choices, quality validation, and architecture.

## Data Context and Sources

Domain: The project simulates a medical imaging assistant that can manage, query, and interpret chest X-ray images and their corresponding textual reports. It is explicitly non-diagnostic — meant only for educational and technical exploration.  
  
Data Modalities:  
- Image: X-ray chest scans (.png, .jpg)  
- Text: Short anonymized radiology findings (.txt)  
  
Data Sources:  
- Chest X-Ray Pneumonia Dataset (Kaggle)  
- Sample educational notes manually written for demonstration  
  
Value: Enables similarity-based retrieval of medical images, cross-modal exploration (text → image), and automatic generation of descriptive captions.

## Data Management Backbone

The pipeline follows the DataOps zone architecture, ensuring clear data lineage and reproducibility.  
  
Zones Overview:  
- Landing Zone: Raw ingestion (Python + MinIO)  
- Formatted Zone: Format homogenization (Pillow, text normalization)  
- Trusted Zone: Data quality and cleaning (imagehash, regex)  
- Exploitation Zone: Embeddings and search (CLIP + ChromaDB)

## Implementation Summary

Tools & Environment:  
- Storage: MinIO (Docker)  
- Processing: Python 3.11, Pandas, Pillow, TQDM  
- Embeddings: ChromaDB + OpenCLIP  
- Generative Model: HuggingFace BLIP Image Captioning  
- Versioning: GitHub, Docker Compose, CI workflow  
  
Pipeline Execution:  
python run\_pipeline.py  
Executes all phases automatically (Landing → Formatted → Trusted → Exploitation).

## Analytical Tasks

1. Same-Modality Search: Find visually similar X-rays (CLIP + ChromaDB).  
2. Multi-Modality Search: Text query retrieves related images (CLIP multimodal embeddings).  
3. Generative Captioning: Generate image description using BLIP image captioning model.

## Data Quality Report (Trusted Zone)

Metrics (example placeholders):  
- Total images processed: N  
- Low-resolution images dropped: N  
- Duplicates removed: N  
- Text files cleaned: N  
- Identifiers replaced: N  
  
All final data converted to .png or .txt, no sensitive information retained, and all transformations logged.

## Operations and Automation

- Orchestration: run\_pipeline.py executes all stages sequentially  
- CI/CD: GitHub workflow validates code format  
- Containerization: MinIO via Docker Compose  
- Configuration: .env file for credentials  
- Reproducibility: deterministic transformations and embeddings

## Results

- All four pipeline zones operational  
- Embeddings successfully generated  
- Multimodal search and image captioning tested successfully  
Example caption result: “A grayscale X-ray showing lung opacity consistent with pneumonia.”

## Conclusions and Next Steps

The project demonstrates a working multimodal data pipeline integrating text and images under a unified structure with three analytical tasks. It applies DataOps principles (reproducibility, quality control, automation) within an extendable framework.  
  
Next Steps for P2:  
- Integrate Streamlit UI  
- Automate ingestion with schedulers  
- Implement advanced analytics (RAG or classification)  
- Add monitoring and evaluation dashboards