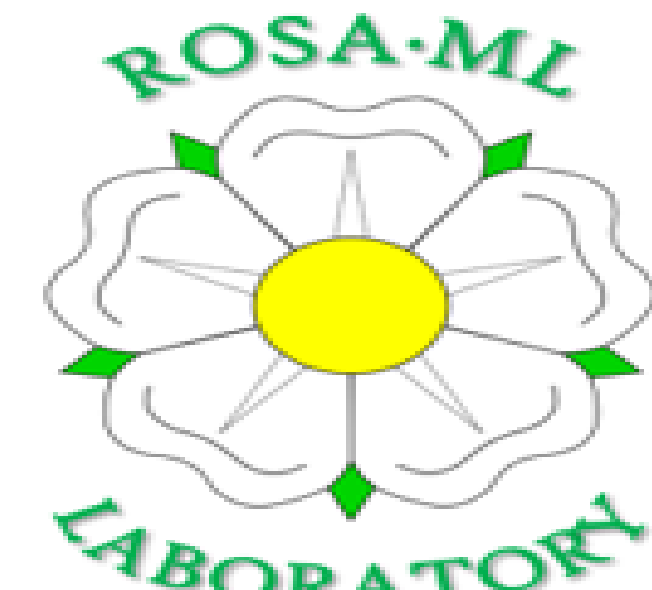




1 Background / Purpose

- Large-scale radiotherapy research depends on heterogeneous DICOM datasets with complex cross-references.
- Downstream analysis and machine learning require consistent organization, validation, and transformation of data.
- rosamllib is an open-source Python framework for scalable radiotherapy DICOM processing, visualization, and preprocessing.

2 Core Capabilities

- In-memory and database-backed ingestion pathways optimized for different data scales.
- Explicit resolution of radiotherapy-specific relationships using DICOM identifiers and frame-of-reference associations.
- Built-in querying and visualization for cohort filtering, inspection, and validation.
- Preprocessing support for structure masks, dose alignment, normalization, and export of model-ready datasets.

3 Innovation / Impact

- 1 Ingests data from DICOM-compliant clinical systems into SQL-backed workflows.
- 2 Indexes metadata to explicitly link objects that are only implicitly linked in native DICOM.
- 3 Provides visualization tools for validation and troubleshooting of dataset structure and inter-object dependencies.

Expected impact

Reduce technical and software-programming barriers for physicists performing radiotherapy planning-data analyses at the scale of an entire clinic or multi-institution level.

4 Methods / Architecture

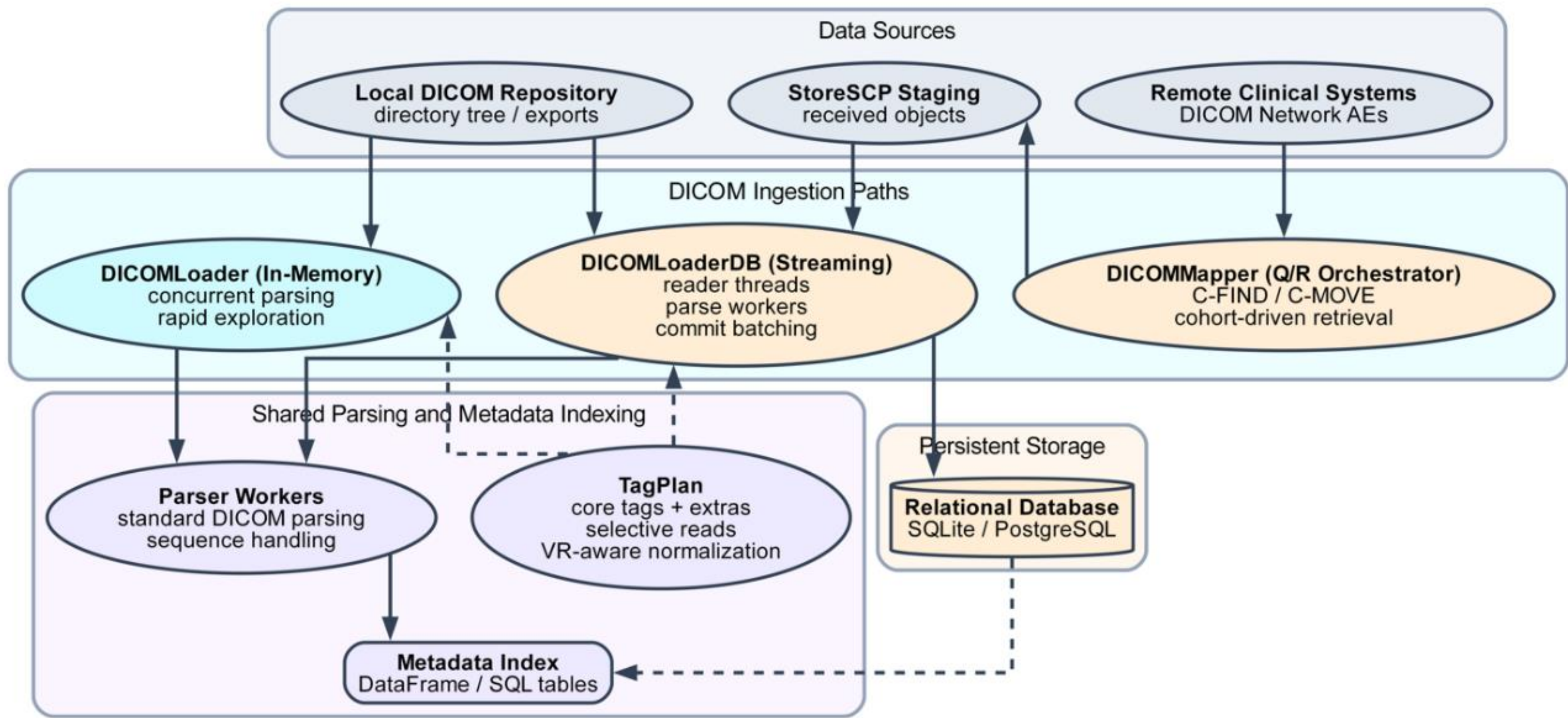


Figure 1. Overview of the rosamllib ingestion and metadata indexing architecture. Radiotherapy DICOM data are ingested from local repositories or remote clinical systems using complementary in-memory and database-backed workflows. A shared parsing layer with selective tag planning and value-representation-aware normalization enables metadata extraction and indexing for preprocessing.

5 Implementation Highlights

- ✓ Local filesystem, StoreSCP staging, and remote clinical system support.
- ✓ Shared parsing and metadata indexing layer.
- ✓ Streaming producer–consumer database workflow.
- ✓ Selective tag extraction and VR-aware normalization.
- ✓ Persistent relational storage (SQLite / PostgreSQL).
- ✓ Built-in cohort querying and graph-based inspection.

6 Preprocessing Outputs

- 1 **Structure masks**
Generate consistent voxelized masks from RTSTRUCT contours.
- 2 **Dose alignment**
Align RTDOSE with images and frame-of-reference relationships.
- 3 **Normalized data**
Apply value-representation-aware metadata normalization.
- 4 **Model-ready export**
Prepare reproducible datasets for analytics and deep learning.

7 Relationship Visualization

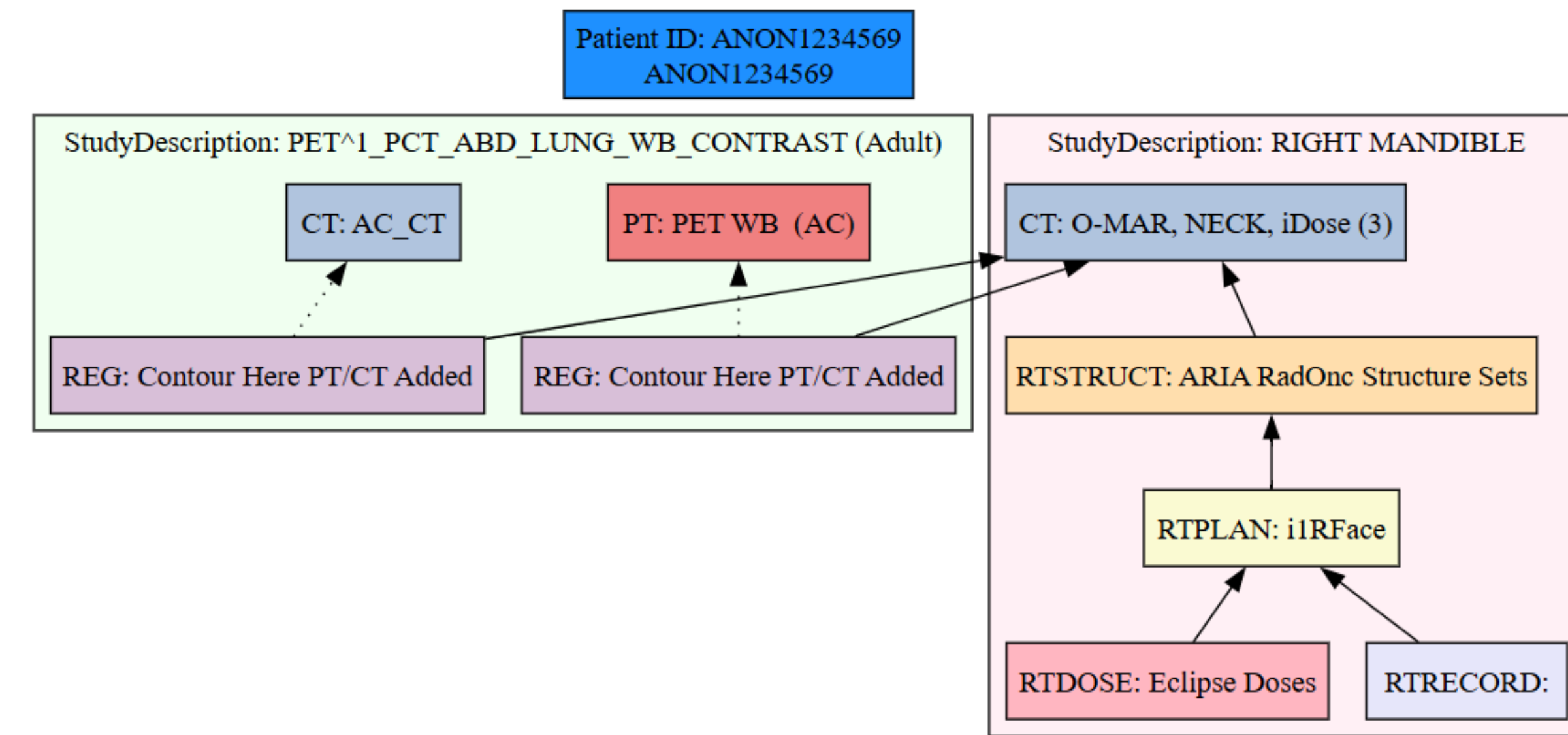


Figure 2. Example series-relationship visualization for an anonymized radiotherapy patient. Nodes represent CT, PET, REG, RTSTRUCT, RTPLAN, RTDOSE, and RTRECORD series; directed edges represent resolved DICOM references across studies.

8 Results

1,869
Patients

12,406
Studies

349,246
Series

6,370,243
Instances

Applied to approximately one year of institutional radiotherapy DICOM data, rosamllib enabled scalable ingestion, relationship validation, visualization, and automated preprocessing across large cohorts.

9 Conclusion

- rosamllib unifies ingestion, metadata indexing, relationship resolution, visualization, and preprocessing within a single framework.
- The framework provides a scalable and extensible foundation for big-data radiotherapy research.
- It supports efficient downstream analytics and machine learning workflows in radiation oncology.



Availability

Open-source repository:

github.com/YAAbdulkadir/rosamllib

Scan to access the code and documentation.