

Enabling Scalable Multimodal AI Research in STARR: Integrating Whole Slide Imaging Pathology Data with Clinical Information

Hannah Morgan-Cooper^{1,2}, Aaron Fanoush^{1,2}, Priya Desai^{1,2}

¹ Stanford Health Care , ²Stanford School of Medicine

1. Background

The increasing use of multimodal artificial intelligence in healthcare has created demand for large-scale integration of digital pathology images with electronic health record (EHR) data. Whole slide imaging (WSI) presents challenges for observational research due to proprietary file formats, vendor-specific metadata structures, and limited support for integrating pathology images with standardized clinical data models.

As part of the STAnford Medicine Research Data Repository (STARR), we developed infrastructure to integrate pathology WSI data into an OMOP-based research environment, supporting scalable linkage between pathology images, their associated metadata, and electronic health record data for downstream AI and observational research. As a first-of-its-kind proof of concept, 35,000 iSyntax files and 1,000 SVS files were converted into standardized pyramidal TIFF files, paired with PHI scrubbed JSON metadata and linked to relevant EHR data in the OMOP CDM. Associated slide metadata was extracted and loaded into an OMOP extension table, including URIs to the corresponding image files, enabling linkage between whole-slide images and clinical data in the OMOP CDM.

2. Methods

2.1 Converting Proprietary WSI Files into Standardized Formats

Whole slide imaging (WSI) scanners generate extremely high-resolution digital representations of pathology slides. To enable efficient viewing and storage, WSI files are typically stored as multi-resolution image pyramids, where lower-resolution versions of the slide are embedded alongside the full-resolution image. This allows researchers to zoom in on cellular details at high magnification without overloading computer memory.

A typical WSI file contains:

- The slide image stored as a multi-resolution image pyramid
- Slide identifiers and accession information
- Scanner and acquisition metadata
- Image dimensions and resolution information
- Thumbnail and label images

- Additional vendor-specific metadata

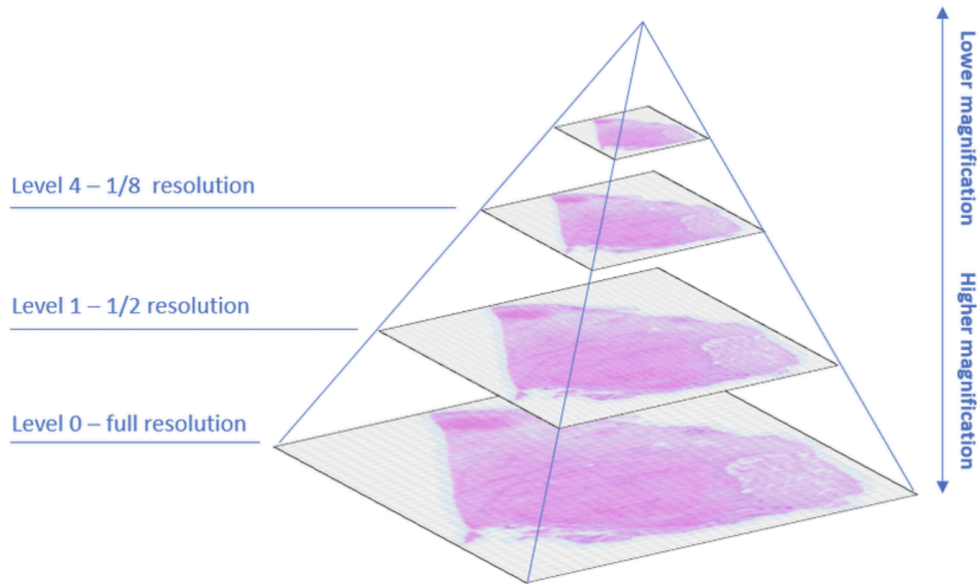


Figure 1: Whole slide imaging file structure source:

https://www.researchgate.net/figure/Whole-Slide-Image-pyramid-structure_fig2_361374092

At Stanford, pathology WSI data originated from two vendor-specific formats:

- **iSyntax** files generated by Philips scanners
- **SVS** files generated by Aperio Leica scanners

Both formats contain image data and associated metadata, but differ substantially in their internal structure and encoding. While SVS files are based on the standard TIFF format, iSyntax uses a proprietary format that requires specialized software libraries for access and processing.

In order to make this data AI ready, we needed to convert the files into a standardized format, and harmonize the metadata. We converted both the SVS files, and iSyntax files into pyramidal TIFFs for the image data, and JSON files for the metadata. Conversion required decoding proprietary vendor formats and re-encoding image pyramids into a common representation suitable for downstream image analysis and machine learning workflows. Existing open-source libraries, including libisyntax and libvips, were extended to support metadata extraction, preservation of clinically relevant attributes, and re-embedding of standardized metadata into the final research-ready TIFF and JSON files.

2.2 De-identification of WSI Files

Patient-identifiable information was removed at both the image and metadata levels during the conversion process. WSI files typically contain embedded label and header images that may include patient names, medical record numbers, or other identifiers printed on the physical slide label. These images were identified and removed from the output pyramidal TIFF files to prevent inadvertent disclosure of PHI.

At the metadata level, vendor-specific metadata fields containing PHI were identified and stripped from the extracted JSON files. The resulting PHI-scrubbed TIFF and JSON files retain only clinically relevant attributes needed for downstream research, such as scanner information, image dimensions, and resolution, while excluding all patient-identifiable content.

2.3 Structuring Pathology Source Data and Clinical Data

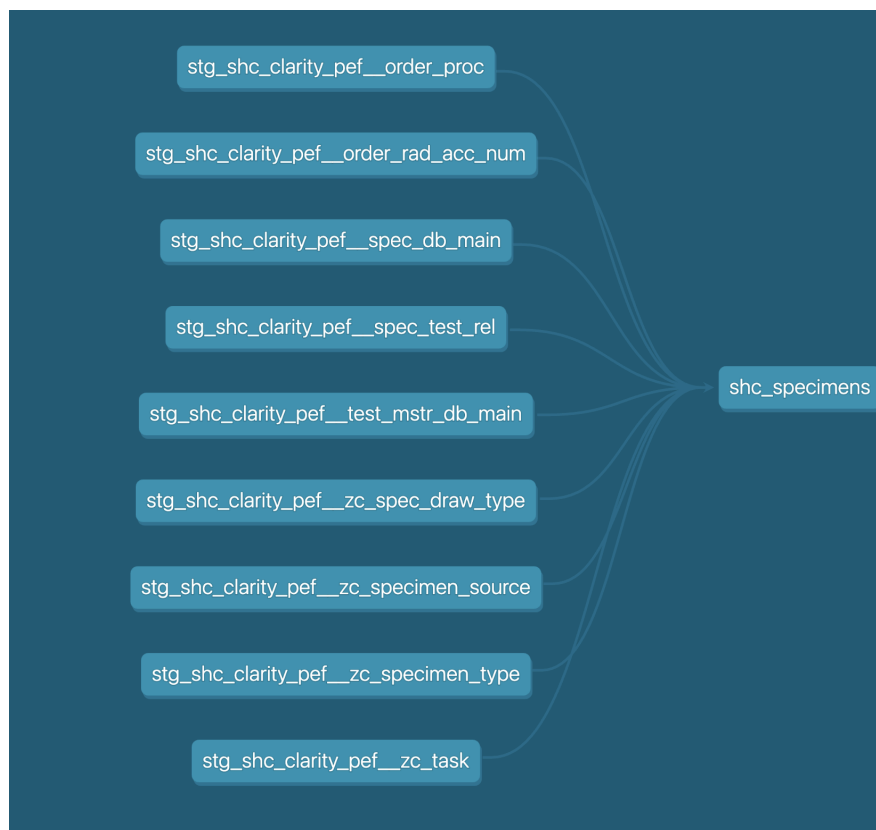


Figure 2: Source tables compiled to create the STARR Common Specimens table

Pathology source data is housed across multiple Clarity tables representing specimen, block, and slide information. These entities form a hierarchical one-to-many relationship in which a single specimen can produce multiple blocks and each block can produce multiple slides. Because clinically important information exists at different levels of the hierarchy—for example, procedure information is held at the specimen level and stainer information at the block level—a

major design consideration was how to represent these relationships without creating unnecessary data bloat.

We decided to compile the twenty underlying Clarity tables into three STARR Common tables, pulling out key clinical information:

- Specimens
 - Accession number
 - Procedure information
 - Specimen source (e.g. lung)
 - Specimen type (e.g. Tissue)
 - Tests ordered
- Specimen_blocks
 - Block ID (accession number + block letter)
 - Stainer
 - Tasks performed on specimen block
 - Container type
 - Media type
- Specimen_slides
 - Slide ID (accession number + block letter + slide number)

Figure 2 shows an example of the simplification of the source Clarity tables into a single specimens table.

We made the design decision to use structural arrays for these tables, allowing us to include multiple lines of information (e.g. multiple stainers) whilst still keeping the tables at the specimen, block or slide level.

2.4 OMOP Extension Table Design and Clinical Linkage

To support pathology image discovery, cohort creation and multimodal analysis, we developed a whole_slide_imaging extension table for inclusion within STARR OMOP deliveries. This table combines the EHR data from the three STARR Common tables, key metadata from the WSI file conversion and URIs to the WSI files.

The table was designed at the slide level, with one row representing a single pathology slide and whole side image. Records were linked to OMOP clinical data using foreign key relationships to person_id and note_id. Linkage through note_id enabled direct association between slide images and pathology reports stored in the OMOP NOTE table.

The table also included harmonized slide metadata extracted from PHI-scrubbed JSON files and secure Google Cloud Platform (GCP) URIs that reference associated PHI-scrubbed TIFF and JSON files. This design allows researchers to identify image cohorts using structured OMOP queries while maintaining access to linked image data and narrative pathology interpretation.

Separate from the WSI file-level PHI scrubbing described in Section 2.2, the OMOP extension table itself undergoes standard STARR de-identification operations before delivery to researchers. These operations include: offsetting integer identifiers (whole_slide_image_id, note_id), substituting person_id with a research group specific new identifier, hashing the accession_number, applying date jitter to all date fields (acquisition_date, procedure_start_time, specimen_received_date), and restricting access to GCP URIs through bucket-level permissions. Table 1 summarizes the extension table schema and the de-identification operation applied to each field.

Table 1: PHI-scrubbed Whole_Slide_imaging table details

Field	Type	Required	Description	phi scrubbing operation
whole_slide_image_id (PK)	INTEGER	Yes	The unique key that is given to an imaging study record.	Offset
person_id (FK)	INTEGER	Yes	A foreign key identifier to the person in the person table	Substitute
accession_number	STRING	Yes	Case accession number	Hash
specimen	STRING	No	Specimen letter e.g. 'A'	
block	STRING	No	Block identifier e.g. A1	
slide	STRING	No	Slide identifier e.g. A1-1	
acquisition_date	DATE	Yes	Date and time when the slide was scanned from the wsi file metadata.	Jitter
procedure_start_time	DATE	Yes	Date and time when the associated procedure started.	Jitter
specimen_received_date	DATE	Yes	Date and time when the specimen was received by the lab	Jitter
Brand	STRING	No	Brand of the scanner used to create the WSI file	Jitter
Scanner	STRING	No	Scanner used to create the WSI file	Jitter
MPP	STRING	No	Microns per pixel	Hash

JSON_URI	STRING	No	GCP URI to de-identified JSON	Restricted access to GCP bucket
TIFF_URI	STRING	No	GCP URI to de-identified TIFF	Restricted access to GCP bucket
Specimen_source	STRING	No	Source of the specimen e.g. lung	
Specimen_type	STRING	No	Type of specimen e.g. blood	
Spec_task_list	ARRAY STRING	No	Lists the tasks performed on the block, including stainer used	
note_id (STARR construct for linkability)	INTEGER	No	A foreign key identifier to the OMOP NOTE table. In cases where more than one note can be linked, the latest note date is used.	Offset

3. Results and Conclusion

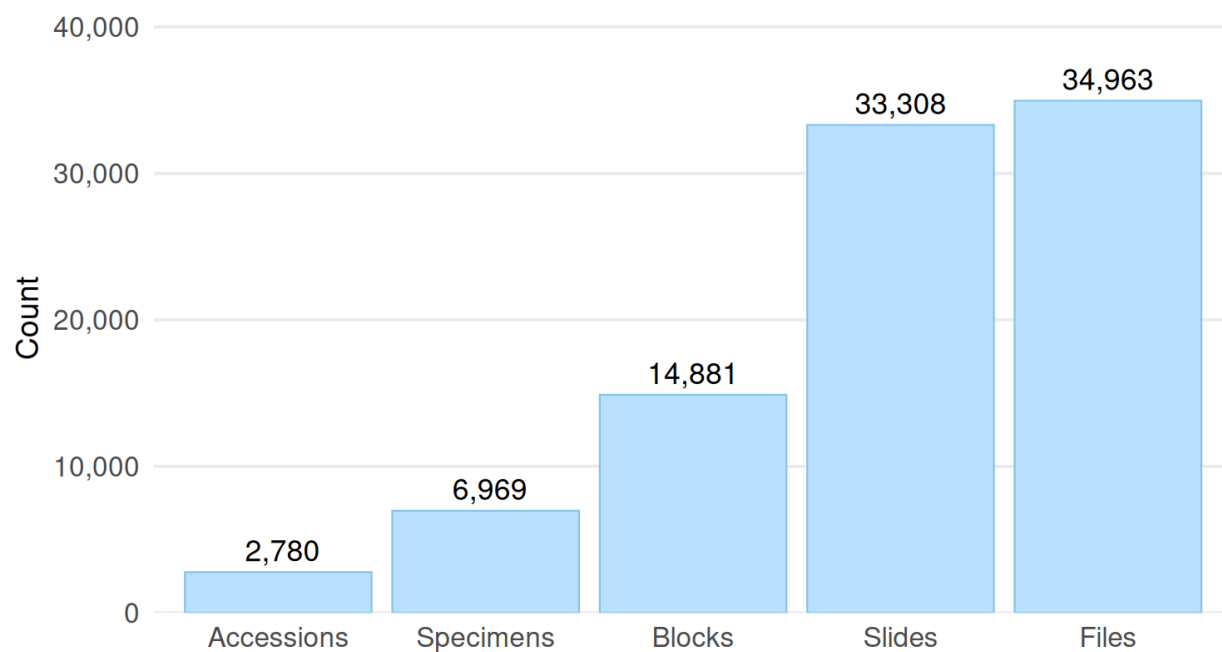


Figure 3: Number of accessions, specimens, slides and WSI files available in STARR OMOP

This implementation established a scalable framework for integrating pathology WSI data within STARR while reducing barriers imposed by proprietary imaging formats and heterogeneous metadata standards. 35,000 iSyntax files and 1,000 SVS files were converted into standardized pyramidal TIFF files paired with deidentified JSON metadata and linked to relevant EHR data.

By converting images into machine-learning-ready formats and linking pathology slides with structured EHR data and pathology reports, researchers can perform multimodal analyses combining image data with clinical context and narrative interpretation. This approach creates a foundation for pathology-focused AI development and demonstrates a practical strategy for extending OMOP-based environments to support large-scale digital pathology research.

References:

- 1) Masjoodi S, Anbardar MH, Shokripour M, Omidifar N. Whole Slide Imaging (WSI) in Pathology: Emerging Trends and Future Applications in Clinical Diagnostics, Medical Education, and Pathology. Iran J Pathol. 2025 Summer;20(3):257-265. doi: 10.30699/ijp.2025.2044210.3367. Epub 2025 Jul 1. PMID: 40746923; PMCID: PMC12308189.
- 2) <https://www.leicabiosystems.com/us/life-sciences-and-research-solutions/application/how-whole-slide-imaging-is-changing-pathology/>