ASU Center for Evolution and Medicine
Arizona State University

STARFISH

How ASU's Center for Evolution and Medicine Used Starfish to Build a Searchable DICOM Catalog for Global Health Research

Written by:
Suhail Ghafoor, Information Technology Manager at the Center for Evolution and Medicine at Arizona State University.

Matt Hutchison, Senior Solutions Engineer, Starfish Storage

April 2026

Abstract

Arizona State University's Center for Evolution and Medicine, in partnership with Starfish Storage, built a metadata-driven catalog of CT scan data for the Tsimane Health and Life History Project—transforming previously disorganized DICOM (Digital Imaging and Communications in Medicine) files into a structured, searchable, and researcher-friendly system. The collaboration between ASU and Starfish tackled challenges in metadata normalization, data governance, and researcher access—ultimately delivering a solution using Starfish that makes complex and voluminous medical imaging data instantly and easily explorable by researchers, via familiar directory structures.

Introduction

The Tsimane Health and Life History Project (THLHP) is a longitudinal biomedical study focused on understanding healthy aging in a non-industrialized population, specifically among the Tsimane, an indigenous group in Bolivia's lowlands who exhibit exceptional cardiovascular health, vitality, and activity well into old age. Nearly all biomedical research today is conducted in sedentary urban populations, so the project examines what healthy aging was like prior to industrialization. Since its inception in 2001, the project has provided medical care and collected extensive demographic, genetic, biochemical, and medical imaging data to investigate chronic diseases of aging, such as Alzheimer's and cardiovascular disease, resulting in a unique dataset spanning over two decades.

The project collaborates with a local hospital in Trinidad, Bolivia where CT-scans are performed. Most of the imaging data is in the form of DICOM files, a standard format for medical imaging. DICOMs contain the pixel information of the image in binary format, as well as metadata describing the machinery, the protocols, the patient, the body part, the hospital, and other information. Each scan can produce hundreds of DICOM files, each containing information about a single slice. As the project has continued to grow, efficient management and analysis of this data became increasingly important, and difficult.

This paper describes how Suhail Ghafoor, Manager of Information Technology at Arizona State University's Center for Evolution and Medicine, tackled a longstanding challenge in global health research: making DICOM data from CT scans easily searchable and accessible to dozens of collaborating researchers around the world. Working with Starfish Storage and other partners, Ghafoor developed a metadata-driven, directory-based DICOM catalog for the Tsimane Health and Life History Project that overcame legacy metadata inconsistencies and addressed institutional constraints around security and cloud storage, ultimately providing researchers a vastly more efficient, self-serve system to access critical imaging data.

Background

When Suhail Ghafoor joined the project in 2022, the DICOM imaging process was still largely manual in terms of logistics: After images were captured at the hospital in Bolivia, U.S.-based team members physically transported hard drives to researchers—a process that often took months between image acquisition and researcher access.

Today, CT scans are copied over to the Bolivian hospital's Picture Archiving and Communication System (PACS) as DICOMs, and from there the relevant research scans are copied to the project's data repository at ASU via a series of scripts Ghafoor wrote for that purpose.

The Problem

DICOM files contain a wealth of metadata, and the PACS provides a high degree of searchability. But once the files are exported from the PACS—you're on your own.

Historically, the data was manually curated and distributed on request after it was in ASU's repository. But this approach was time-consuming, inconsistent, and became unsustainable as more researchers requested increasingly specific subsets of data—by organ type, scan quality, patient characteristics, or the reconstruction protocol for specific parts of the scan.

The imaging data arriving from Bolivia was rich, but inconsistent. In addition, DICOM metadata varied widely across time, scanner versions, and even individual technicians. Changes in CT scanner models, technician practices, and protocols introduced variations in how data was labeled and stored. File names were also unreliable as indicators of the contents of the file.

Most importantly, there was no centralized, searchable catalog of DICOM metadata. Each new request required manual sorting, investigation, and curation, which slowed down research and introduced opportunities for error. Specific scan types researchers might request—like "all high-quality lung scans for male participants over 50"— were impossible to automate, and researchers had to wait days or weeks for manual filtering.

Complicating the problem were a few additional constraints:

- **Data governance:** Every image is tied to a unique consent trail, verified by both ASU and local leadership in Bolivia. Access is restricted and must be auditable.
- **Infrastructure limitations:** Cloud services like Amazon S3 were unusable due to several factors. Stringent consent agreements with the Tsimane community required all data, after transfer from Bolivia, to be stored on ASU-controlled infrastructure and have high security standards. Additionally, maintaining local infrastructure ensured that researchers would always have access to direct, local support. If a researcher encountered issues with data access, downloads, or needed technical assistance, they could reach out to ASU personnel for immediate help rather than navigating third party support channels.

- **Data complexity:** Scans include multiple passes of varying quality—smaller and lower quality scans for machine calibration, and larger, higher quality scans for research and medical treatment. There are also multiple image reconstruction types for different body parts or tissue types, for example bone, soft tissue, or organs. Metadata fields are often cryptic, not human-readable, and embedded in machine-specific tags.
- **Privacy and Anonymization**: ASU's workflows had to support privacy-preserving data access. Datasets needed to be either *de-identified* (with removable placeholder IDs) or *fully anonymized* (with no re-linkable identifiers).

The Objective

To address these problems, Ghafoor set out to create an organized DICOM catalog that would:

- Allow researchers to find scans based on specific metadata: body part, quality level, reconstruction type, and more.
- Display the scans in a familiar file system interface that researchers could access themselves, without requiring direct involvement from Ghafoor or his team to curate desired scans.
- Be maintained locally at ASU without the use of cloud resources.
- Respect the strict data access controls agreed upon with Tsimane leadership.

Solution: Building a Metadata-Driven Catalog with Starfish

Ghafoor turned to Starfish Storage to solve the cataloging and metadata query challenges. Starfish had the capacity to:

- Ingest file-level metadata.
- Allow custom tagging and enrichment.
- Support metadata query logic and automation.
- Reflect query results as customized symlinked directory structures that mirrored researcher needs.

Implementation Details: From Metadata Normalization to Directory Trees

Implementing Ghafoor's vision took place in four phases:

1. Transferring the data from the PACS in Bolivia to ASU's servers
2. Normalizing the metadata
3. Importing the metadata into Starfish
4. Displaying the results in a researcher-accessible directory structure

Data Transfer from PACS

Ghafoor first wrote and installed scripts to copy the DICOM data from the PACS in Bolivia to servers at ASU, eliminating the need for physical transfer of hard drives. This data transfer pipeline consists of multiple scripts that query, download, validate, extract metadata, compress, and transfer the data. They interact with each other using a Postgres database as a job queue to check for pending tasks. These scripts are now open source and available at GitHub.

Metadata Normalization

Ghafoor then set about the task of normalizing the DICOM metadata—mapping inconsistent legacy metadata fields to a normalized vocabulary of categories across multiple generations of imaging protocols. For example, body part data often appeared in different fields depending on the year, or the specific technician who performed the scan: Ghafoor collaborated with radiologists to determine which machine-generated fields (like scanner position) could reliably infer anatomical regions. He then created mappings that unified this information into standard categories: body part, scan quality, reconstruction type, and others.

Import Metadata into Starfish

Once the metadata was normalized, it could now be imported into Starfish. Ghafoor wrote a Python script that extracts metadata from the DICOM images that the research team would need, like scan parameters, positioning data, and patient characteristics. The script automatically determines the organ that was scanned based on window center and width values by referencing a JSON lookup table. It then imports the metadata into Starfish.

The script also transforms the exported datasets depending on the requirements for being *de-identified* (with removable placeholder IDs), or *fully anonymized* (with no re-linkable identifiers). A mapping table is securely stored and accessible only at ASU for cases where future re-identification might be necessary.

Ghafoor has generously open-sourced this script to make it available to other research enterprises facing similar challenges with DICOM data: The script is now available at GitHub.

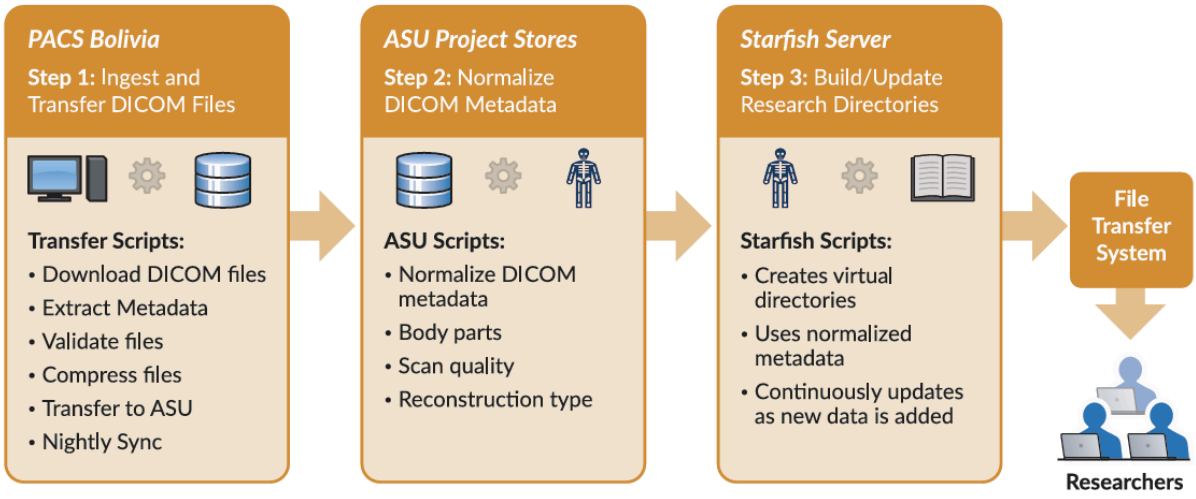


Diagram: An overview of the flow of DICOM files from Bolivia to ASU researchers show ASU and Starfish workflow

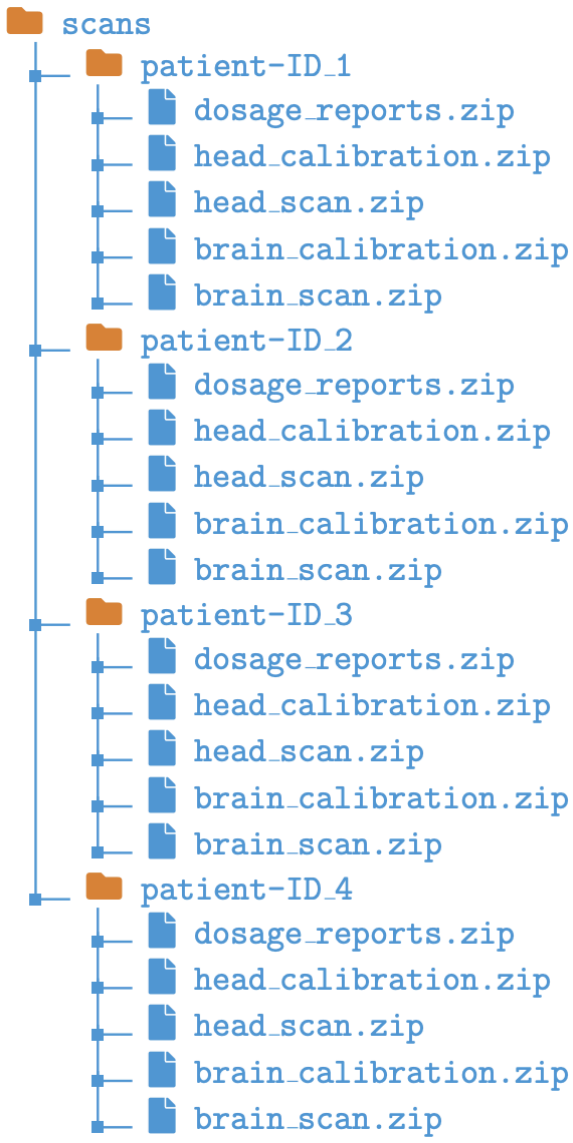
Displaying the Results in Directory Structures

Once the metadata was in Starfish, Ghafoor then collaborated with Starfish's Customer Success team to realize his vision of presenting the DICOM data to researchers in the form of a familiar directory structure. Starfish's Matt Hutchison wrote scripts to create virtual, symlinked directory structures based on specified combinations of the normalized metadata values. Since these structures are composed of symlinks, not copies, they require no additional storage space. They are quasi-permanent and self-serviceable—researchers can browse them like a normal file system and retrieve exactly what they need without technical assistance. This approach empowered researchers for the first time to work independently while preserving strict control over the underlying data.

Initial directory structure

This directory organizes scans in zip files by patient ID. For each patient ID it contains dosage reports, as well as head and brain scans of high quality, and lower-quality scans used for machine calibration. This directory is most useful as new data is being collected, since any new directories appear in the top level directory, and the ID or date modified of the directory can be used to filter out recent files. This is used most by people doing quality assurance at early stages, and who may want to see both high and low-quality images used for calibration.

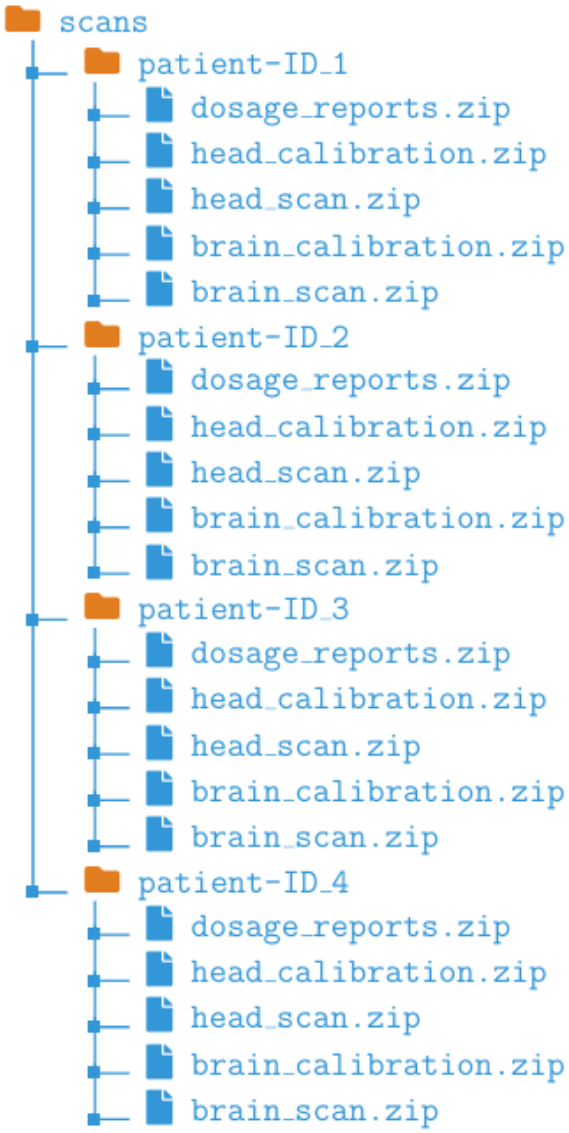
Directory structure Example 1:



Structure for sharing with researchers

This structure organizes scans by body part, allowing researchers to access scans for whatever body part they are interested in. This structure is useful for sharing data with external researchers who are typically only interested in a single organ and want to separate high-quality scans from low-quality calibration scans.

Directory structure Example 2:



Results and Outcomes

With the new Starfish-powered catalog and directory structures in place, the research team achieved:

- **Dramatic acceleration in access**: Turnaround time for imaging requests from researchers went from months to minutes for most requests.
- **Elimination of bottlenecks**: Researchers can browse and select data independently, without the need for manual curation by the ASU team.
- **Improved data governance**: Every dataset is traceable, compliant, and reproducible.
- **Sustainability**: The system requires minimal ongoing intervention and can be managed by others if needed. Using Starfish's REBUILD command and scripting tools, Ghafoor can update the virtual directories with the latest changes to the dataset.

Crucially, the solution respects all ethical, legal, and technical constraints—enabling cutting-edge research while honoring community commitments.

A Replicable Model for Research Imaging Data

The collaboration between ASU and Starfish Storage transformed a fragmented DICOM repository into a structured, metadata-driven resource tailored for researcher needs. By combining deep domain knowledge, technical creativity, and platform flexibility, Ghafoor and his collaborators created a solution that not only serves the Tsimane Health and Life History Project, but also offers a blueprint for other research institutions facing similar challenges.

For IT managers and data stewards grappling with unstructured medical imaging data, this project demonstrates how Starfish Storage can be a foundational part of a scalable, researcher-ready solution—one that turns raw scans into accessible scientific assets.

Suhail Ghafoor is the Information Technology Manager at the Center for Evolution and Medicine at Arizona State University, where he develops open-source data management solutions, maintains research databases, and builds computational pipelines for complex biomedical research spanning genomics, medical imaging, and population health studies. He specializes in creating secure, scalable infrastructure for petabyte-scale datasets across multiple NIH-funded projects, developing automated tools and systems that enable efficient data collection, transfer, and analysis.

Matt Hutchison is a Senior Solutions Engineer with **Starfish Storage**, the unstructured data management platform for high performance computing (HPC), AI/ML, and other demanding file-based workloads. For more information about how your organization can unlock the power of its data, please visit Starfish Storage at https://starfishstorage.com/

The authors would like to acknowledge **Daniel Cummings** of **Chapman University** for his essential contributions to this work. Dr. Cummings played a key role in establishing the CT scanning system in Bolivia, including the configuration of scanner protocols. He also provided critical guidance on data management and served as the primary resource for troubleshooting and workflow development. His efforts in testing and refining the workflow were instrumental to the success of this project.