PHI Redaction for Neuroimagery

Nakeisha Schimke[1], Alex Barclay[1,2], Peter Gehres[1,2], and John Hale[1]

[1]Institute of Bioinformatics and Computational Biology, University of Tulsa [2]Laureate Institute for Brain Research, Tulsa, Oklahoma

Objectives

The potential for large scale collaboration and data sharing is seriously undermined by concerns over the management and handling of personally identifiable information (PII) in neuroimagery data sets. We address this problem by augmenting XNAT [1] – a platform for capturing and managing neuroimages and associated data – with a toolset and integrated workflow for pseudonymous redaction of PII in DICOM image files [4]. The approach adopted by this effort gives researchers the maximum power and flexibility in sharing neuroimagery data sets while transparently coping with PII considerations in a standardized data curation process.

Neuroimage Data Stack

To ensure comprehensive redaction, we have logically deconstructed neuroimage data sets in a layered approach at the hardware, operating system, and data analysis levels [5]. At the simplest and lowest level on the stack, the raw neuroimage data exists on physical storage media (eg. hard drives) in a binary form as a stream of data. At the file system level, the neuroimage is represented as a file pointer to an address space on the physical disk, along with associated metadata. The DICOM file format is the middle layer, containing both metadata and pixel data in a single file. The image pixel data can be encapsulated for compression purposes, and the logical layer is the resulting neuroimage.

DICOM Redaction

This project focuses on the DICOM layer, removing PII metadata from stored images. This tool extends XNAT to seamlessly integrate redaction into the natural analysis and archive process, lessening the burden of data management on the end user by providing sanitization and verification along with persistent pseudonymous identifiers.

Workflow

The primary workflow is based on the XNAT workflow so that to the researcher, it remains unchanged. Figure 1 shows the redaction workflow from the end user's perspective. The process begins when the image is acquired. Raw data is collected and stored according to the local institution's policy. Once it is released from quarantine, the redaction process can be invoked by the user. This produces a new sanitized data set with privacy mapped pseudonymous identifiers that is sent to the quarantined pre-archive area, mandating a second sanity check before it is released to the XNAT archive. Once it enters the XNAT archive, it is treated as normal data and is available for analysis without modifications to XNAT itself. This new data set contains privacy mapped identifiers available only to a subset of users and can be publicly shared.

The logical view of the data path reflects the extensions in Figure 2. XNAT invokes our external redaction engine, which makes a sanitized a copy of the data and stores the original PII in the privacy map. The redacted data set is fed back into XNAT. The privacy mapped data is not made available to the end user but is accessible to administrators through a command-line tool.

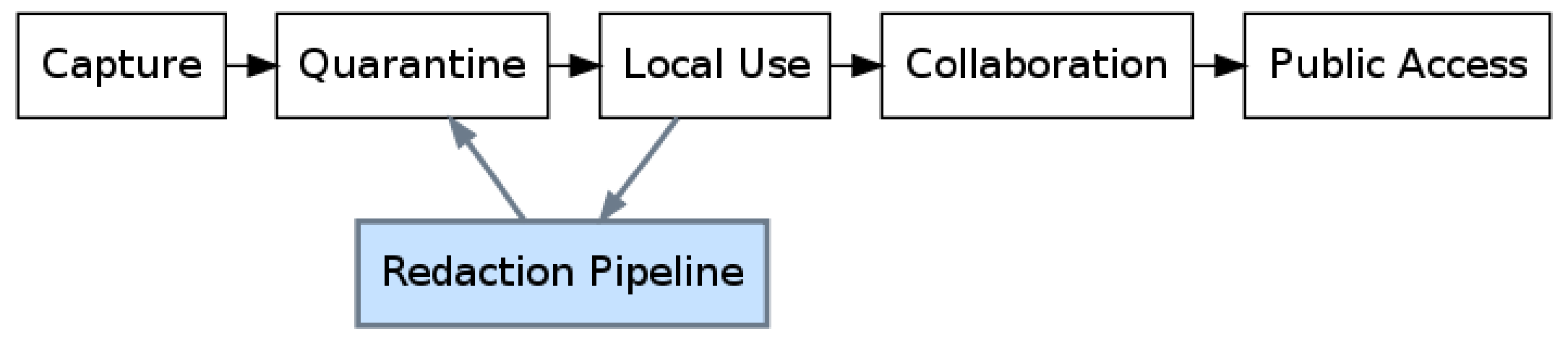

Figure 1: Neuroimage redaction data workflow.

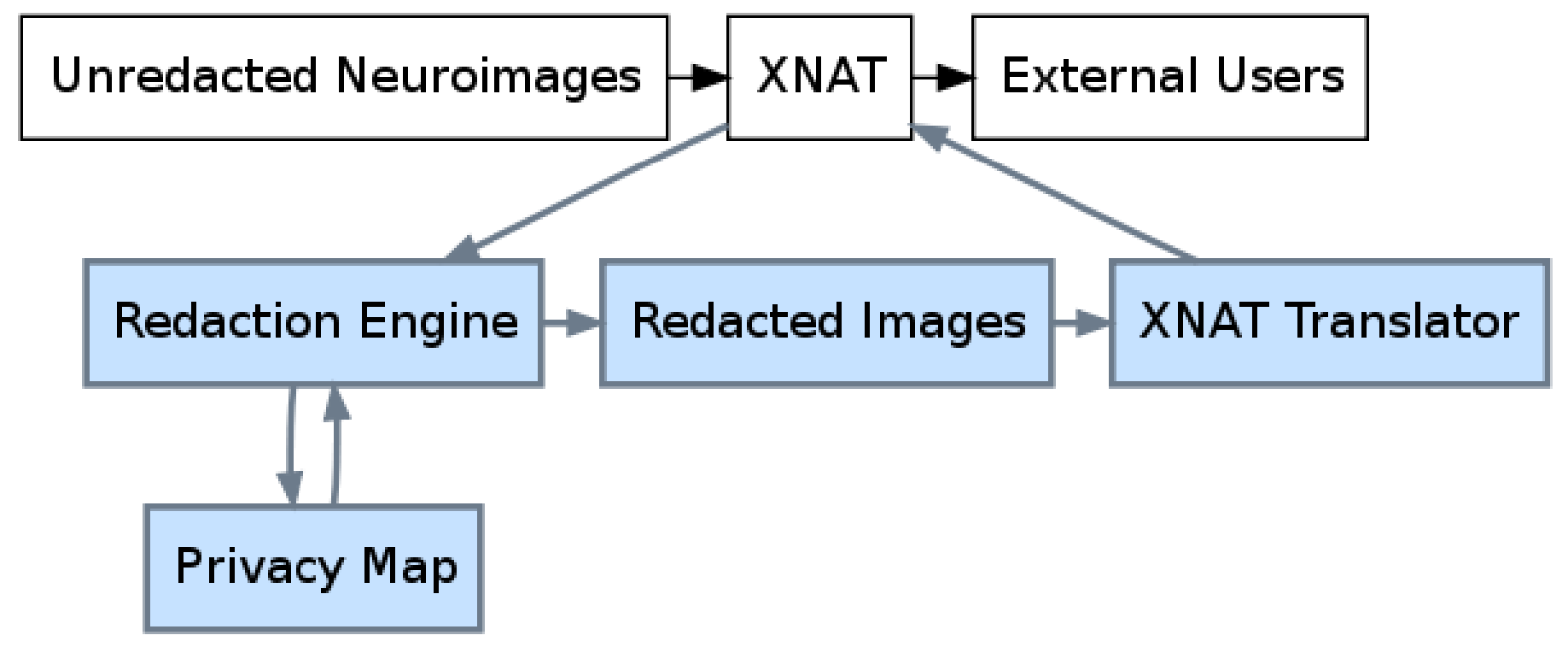

Figure 2: Logical workflow with modifications in blue.

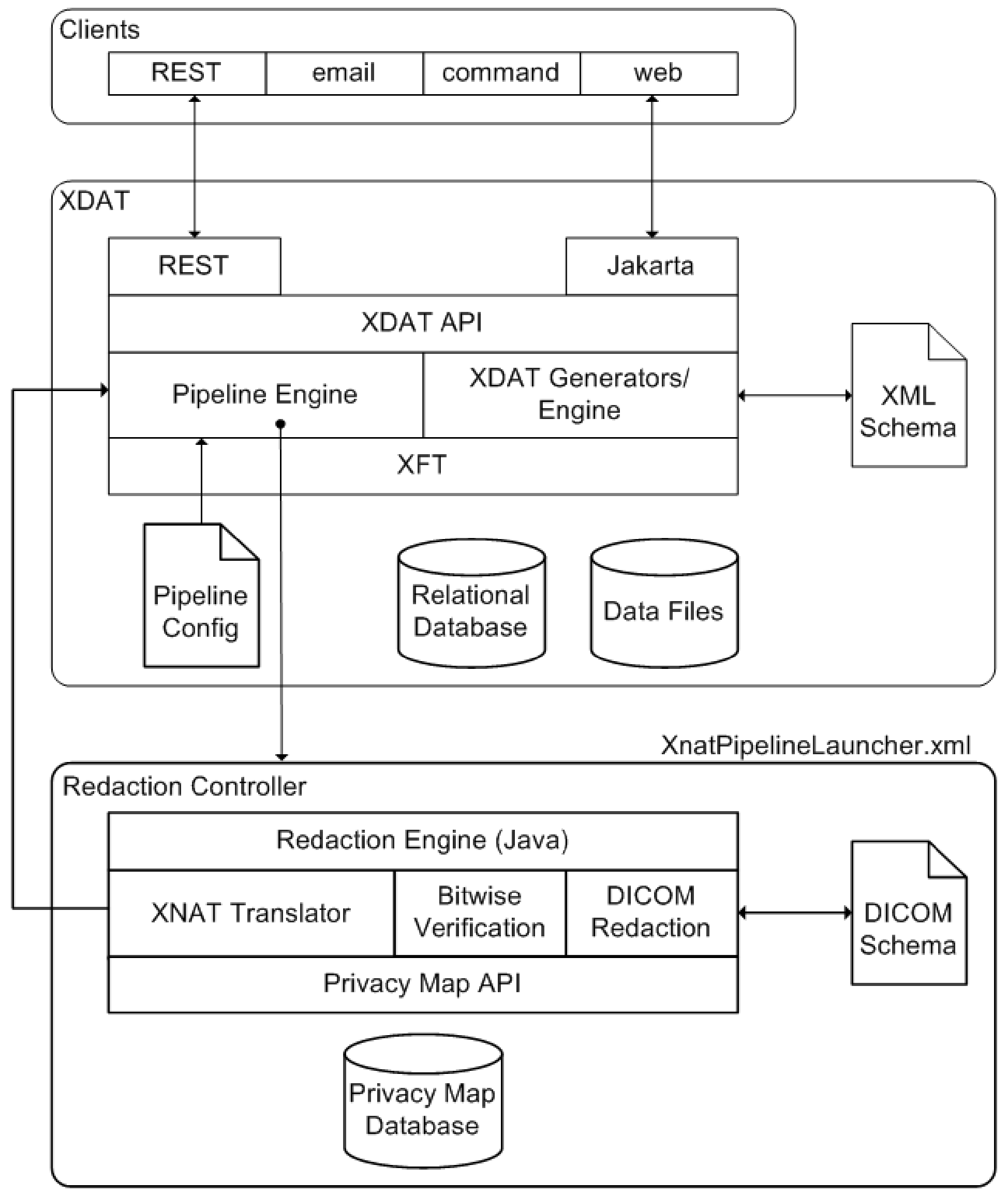

Figure 3: Modified architecture diagram.

Architecture

There are three primary logical components: (1) the redaction engine, (2) the privacy map, and (3) the XNAT translator. The redaction engine and privacy map integrate into the XNAT pipeline facility, enabling an XNAT-aware redaction process capable of providing feedback, status tracking and debugging features within XNAT.

The redaction engine serves as an expandable entry point and contains both DICOM image redaction procedures and a low-level bitwise verification mechanism to fully remove embedded PII. The privacy map is an encrypted database containing the link between the original images and their redacted counterparts, accessible only by an API to ensure the integrity of the databases by exposing a non-harmful subset of commands to the redaction engine. It provides consistency between research subjects of the redacted ID to subject PII, preventing statistical skew due to data duplication. The XNAT translator creates new XNAT identifiers and transmits redacted data in a format recognizable by XNAT.

Advantages

The privacy map provides a mechanism for managing subjects across institutional boundaries by removing PII and creating pseudonymous identifiers. This transparently maintains mapped identifiers throughout a study and ensures the persistence of the redacted identifier over multiple imaging sessions. As part of the XNAT processing pipeline, we leverage the existing framework and interface of XNAT, and redaction becomes a natural part of the existing workflow. The redaction process is forensically verifiable to ensure confidentially of subject data. This technique is accepted for handling digital evidence, providing researchers legally accepted protection and assurance that shared subject data does not fall under the breach notification requirements of HIPAA/HITECH.

Future Work

This effort will be expanded to integrate a comprehensive plan for neuroimage redaction, including the hard drive and file systems, as well as a defacing technique for structural images.

References

[1] D. W. Marcus, T. Olsen, M. Ramaratnam, and R. L. Buckner, Neuroinformatics, **5(1)** (2006), 11-34.

[2] NRG, Washington Univ. in St. Louis, Pipeline quick tutorial (2009).

[3] E. McCallister, T. Grance, and K. Scarfone, NIST Special Publication 800-122 (draft) (2009).

[4] NEMA, Digital Imaging and Communications in Medicine (2008).

[5] A. Barclay, N. Schimke, and J. Hale, USENIX Security (2009).

Acknowledgments

We gratefully acknowledge support from the William K. Warren Foundation.