\section\*{Preliminary Studies}

This effort builds upon the investigators and associated personnel’s experience in security compliance (HIPAA/HITECH), research data flow, and digital forensics. A background in all these areas provides the cross-discipline foundation for this effort to succeed.

\subsection{Data Management and Compliance}

At the Laureate Institute for Brain Research, HIPAA compliance has typically been handled by the parent hospital system, however incorporating the neuroimaging component has shifted organizational boundaries, creating a new institute responsible for all IT and compliance aspects of the longitudinal study and associated research projects. In the course of the institutional creation, a comprehensive examination of research data flows and compliance measures has been taken, paying close attention to the needs of core researchers and the nature of their external collaborations.

This review revealed that scientific personnel use haphazard methods for data management, often relying on arcane naming systems and individualized workflows. As a new institution, with a shared information infrastructure, individualized flows prevent cross investigator collaboration. This barrier to growth mandated the need for centralized data management and analysis facilities, with a tentative plan to build upon the XNAT architecture \cite{barclay09} \cite{xnat}.

All personnel are trained in the theory of HIPAA compliance for research subjects, however there is a disconnect between what a researcher sees presented from his tools on screen to what is hidden in file metadata. This abstraction of PHI leads to data breeches of well meaning staff. This problem is not restricted to DICOM files; it is a bigger problem when sending datasets over physical media, where PHI can hide in low-level file system structures, file system meta-data, and operating system indices\cite{gavin paper}.

HIPAA regulations and supporting documents were designed for clinical care and the assumption of well-regulated centralized storage and access control. This model is typically not used in dispersed neuroimaging studies where each researcher requires access to raw imaging data. However, in this model data security compliance is almost impossible. Under HIPAA; PHI data must be protected by certain mechanisms, in ways often unknown and impractical for individual researchers \cite{nisthipaa}.

By using workflow tools such as the XNAT framework, an institution meets the HIPAA data security requirement, and by performing redaction against the centralized storage an organization meets the privacy regulations as well \cite{nistpii}.

XNAT is the leading candidate framework within which to embed a comprehensive digital redaction solution, due to the data management facilities and previous experience.

Recently, the PI led a team (with cooperation from the Neuroinformatics Research Group at the Washington University School of Medicine) that conducted an information security risk assessment for XNAT, yielding a deeper understanding of threats to and vulnerabilities in the platform \cite{schimke09}.

This afforded student developers and graduate students, who would implement this proposal, the opportunity to become familiar with the design and operation of XNAT. The risk assessment therefore provided an excellent basis on which to define operational and security requirements for an integrated digital redaction solution.

\subsection{Redaction}

Redaction is the process of removing privileged information from a document or set of documents before its presentation to other parties. This is not a trivial process, in that redaction must include 100% of targeted data to be removed, and must be scientifically sound and verifiable to combat legal opposition to methods used or produced datasets.

When a group from the University of Tulsa (including the PI) started addressing redaction for legal production in 2005, the recommendation from the NSA was to copy files to remove metadata\cite{nsaredact}. This approach was seen as crude since it was not scalable for many files, nor did it map the original data to the redacted version or provide any verification mechanisms. Borrowing techniques from cryptography, digital forensics, and operating system design, a redaction method was developed that (1) examined the same set of data at different abstraction layers of the file/media, (2) created a mapping of redacted content to original, and (3) verifies redaction of data \cite{redactieee} \cite{Barclay2007} \cite{redaction3}.

This above project was developed as a general framework for redaction, allowing the inclusion of domain and file type specific information to increase completeness. The partnership with LIBR and this proposal will provide a test bed to adapt a limited version of the previous effort. There are eighteen PHI identifiers, which provide context when searching data; this computationally simple search space reduces complexity of proposed redaction engine code.

This proposal is built upon comments received at the 2009 USENIX security conference, where a poster was presented which explored the issue of comprehensive neuroimage redaction. Feedback reduced the scope of this proposal, concentrating on the DICOM layer first, with redaction of the full file system stack as a later target \cite{Barclay2009}.

The cross institution nature of this effort provides an ideal setting for the development of an integrated neuroredaction solution. The cross discipline makeup and experience of computer security professionals, information technology managers, and neuroimagers enable a tailored redaction solution developed to address ease of use, compliance, and of confidentiality of all stakeholders.