Project Abstract/Summary

In neuroscience, 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. In particular, HIPAA and HITECH rules mandate substantial measures for the preservation of subject privacy. For the researcher, whose focus is on the science, the management of information privacy is a wholly burdensome task. The vast potential for sharing and meta analysis of neuroimagery in research goes unfulfilled for two reasons; (i) systematic sanitization is hampered by the dearth of tools to systematically expunge PII from DICOM images, and (ii) an established workflow that integrates sanitization into the data sharing process remains conspicuously absent.

This proposal seeks to address both of these challenges by augmenting an established and popular software framework for managing and sharing neuroimagery – the Extensible Neuroimaging Archive Toolkit (XNAT) – with a toolset and integrated workflow for redaction of PII in DICOM image files. The redaction process to be engaged by this effort differs from naive sanitization techniques in that it uses pseudonymous identifiers in a Privacy Mapping Database to disambiguate between subjects without tracking their identities. This approach gives researchers the maximum power and flexibility in sharing neuroimagery datasets, while transparently coping with PII considerations in a standardized data curation process.

The proposed redaction toolset complements the data curation and management tools within XNAT and folds neatly within existing XNAT operational workflows and lab processes. The toolset and workflow adhere to the secure design principles of psychological acceptability and least privilege, promoting broad user adoption and ensuring subject confidentiality. The sanitization process itself built upon principles of legal redaction and rules of evidence, providing heightened levels of assurance for scientists and investigators for HIPAA compliance.

The long term objectives of this project are to create a comprehensive and practical infrastructure for managing PII in neuroimagery datasets, and to relieve the burden of the investigator from the technical aspects of data sanitization and redcation. In so doing, this effort will remove substantial obstacles to large-scale collaboration and data sharing in neuroscience.

1 Project Narrative/Relevance

In neuroscience, 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. This proposal seeks to address this problem by augmenting the Extensible Neuroimaging Archive Toolkit (XNAT) with a toolset and integrated workflow for pseudonymous redaction of PII in DICOM image files. The approach adopted by this effort gives researchers the maximum power and flexibility in sharing neuroimagery datasets, while transparently coping with PII considerations in a standardized data curation process.

Facilities and Resources

Institute of Bioinformatics and Computational Biology, University of Tulsa - The Institute of Bioinformatics and Computational Biology (IBCB) spans a collection of research topics that include the development of new tools to support data analysis for the life sciences, and computational models that simulate biological processes. Significant energies of IBCB faculty and students are devoted to neuroinformatics in support of a large scale research collaboration between member institutions of the Tulsa Research Quadrangle (TRQ) Laureate Insitute for Brain Research (LIBR), The University of Tulsa, OU-Tulsa, and the Oklahoma Medical Research Foundation (OMRF).

Laboratory space will be dedicated to project software development and testing in the College of Engineering and Natural Science's Keplinger Hall at the University of Tulsa. The lab will house the equipment and personnel needed to complete the project.

Laureate Institute for Brain Research, Laureate Psychiatric Hospital and Clinic - The Laureate Institute of Brain Research (LIBR) is commencing longitudinal studies to discover the genetic basis of brain behavior. LIBR resources include a 40,000 sq. ft. neuroimaging facility dedicated to research, and containing a GE 750 3 Tesla scanner with plans underway to acquire additional scanners. LIBR presently has 34 full-time professional staff, 12 of whom hold doctoral degrees.

Drs. Jerzy Bodurka and Pat Bellgowan from LIBR will provide guidance and counsel on the development of the proposed solution from a neuroscientific and end-user perspective. Both are intimately familiar with the full spectrum of neuroimaging data formats and with the user and operator requirements as they pertain to established workflow patterns for the acquisition, maintenance and sharing of neuroimagery.

Neuroinformatics Research Group, Washington University School of Medicine - The Neuroinformatics Research Group (NRG) is focused on facilitating the integration, mining, and sharing of data from across the neuroscientific realm. The NRG has developed and maintains a number of open-source and open-access projects along these lines, including the Extensible Neuroimaging Archive Toolkit (XNAT). The NRG is part of the Neuroimaging Laboratory in the Mallinckrodt Institute of Radiology at the Washington University School of Medicine. It is also a member of the Biomedical Informatics Research Network (BIRN) and receive additional support from the Howard Hughes Medical Institute and the McDonnell Center for Higher Brain Function.

The XNAT platform developed by members of the NRG provides the underlying software framework for the proposed effort. Software developers of the neuroimage redaction solution will interact with computer scientists, system developers and researchers in the NRG regularly to ensure a smooth and successful integration of the technology produced by the effort.

Personnel Justification

John Hale, Ph.D. - Professor of Computer Science and Lead Research Scholar, Institute of Bioinformatics and Computational Biology, University of Tulsa (Principal Investigator): Dr. Hale will devote 11% of his time during the academic year and one month effort during the summer to supervise the software development effort.

Specific Aims

In neuroscience, 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. In particular, HIPAA and HITECH rules mandate substantial measures for the preservation of subject privacy. For the researcher, whose focus is on the science, the management of information privacy is a wholly burdensome task.

While multi-institutional collaborations and meta-analyses can make good use of sanitized datasets, the vast potential for sharing and meta analysis of neuroimagery in research goes unfulled for two reasons; (i) systematic sanitization is hampered by the dearth of tools to systematically expunge PII from DICOM images, and (ii) an established workow that integrates sanitization into the data sharing process remains conspicuously absent.

This proposal seeks to address both of these challenges by augmenting the Extensible Neuroimaging Archive Toolkit (XNAT) with a toolset and integrated workflow for redaction of PII in DICOM image files. The redaction process to be engaged by this effort differs from naive sanitization techniques in that it uses pseudonymous identiers in a Privacy Mapping Database to disambiguate between subjects without tracking their identities. This approach gives researchers the maximum power and flexibility in sharing neuroimagery datasets, while transparently coping with PII considerations in a standardized data curation process.

The specific aims of this project are:

1. Completion of a comprehensive redaction toolset for DICOM images: Dicom images are embedded with both sensitive personal health information (PHI) and PII data. Redaction of this information will be comprehensive across the varying DICOM fields utilized by the various imaging systems (e.g., GE, Seimens, and Phillips DICOM formats).
2. Construction of a Privacy Mapping Database: This database will map neuroimage datasets to pseudonymous subject identifiers, enabling large scale data sharing, while preserving subject privacy under HIPAA regulations.
3. Design and integration of a data curation/redaction workflow within XNAT: This aim will integrate the tools developed in aims 1 and 2 within an established neuroimage archival framework to enhance both security of the data and functional utility of the database for the neuroscientific community.

The proposed redaction toolset complements the data curation and management tools within XNAT and folds neatly within existing XNAT operational workflows and lab processes. The toolset and workflow adhere to the secure design principles of psychological acceptability and least privilege, promoting broad user adoption and ensuring subject confidentiality. The sanitization process itself built upon principles of legal redaction and rules of evidence, providing heightened levels of assurance for scientists and investigators for HIPAA compliance.

The long term objectives of this project are to create a comprehensive and practical infrastructure for managing PII in neuroimagery datasets, and to relieve the burden of the investigator from the technical aspects of data sanitization and redcation. In so doing, this effort will remove substantial obstacles to large-scale collaboration and data sharing in neuroscience.

Background and Significance

This section presents the dominant neuroimage repositories and archival frameworks, along with related efforts in data sanitization. It offers a gap analysis of the technologies and tools presented, and discusses the impact of closing these gaps with proposed effort.

Related Work

Researchers make use of online repositories for neuroimages to share data and their analyses. The promise held for these repositories to foster large collaborative endeavors remains immense, but is limited by concerns over the privacy of study subjects. Tools exist to sanitize neuroimage datasets, e.g., eliminating PII from DICOM images, but even these suffer from problems that limit their utility in fostering large-scale data sharing and collaboration.

fMRIDC

The fMRI Data Center is a repository for peer-reviewed neuroimages and related study data [3]. While no longer open for new data submissions, the subject privacy policy holds researchers responsible for santizing their own data prior to submission. The Data Center performs a scan of submissions to remove any potentially identifying data overlooked by the researchers. Researchers may also remove anatomical image volumes to prevent reconstruction of the subject's face; otherwise, the Data Center will strip the images before they are made available.

This approach adequately protects the privacy of the subject but requires the researcher to maintain a list of subjects identifiers if necessary. It also has the potential to strip data it determines to be uniquely identifying without notifying the researcher. This prevents the researcher from reusing the identifier if the subject participates in a future imaging session.

XNAT (Extensible Neuroimaging Archive Toolkit)

The Extensible Neuroimaging Archive Toolkit (XNAT) is an open software framework that provides data management and control for neuroimaging studies [5]. As such, XNAT is designed to facilitate multi-institutional research collaboration through data management and control. Data flows through XNAT in five phases: Data Acquisition, Quarantine, Local Use, Collaboration, and Public Access. In the Data Acquisition phase, data is uploaded to XNAT and is quarantined, where the user validates it before viewing and analyzing it. Users can then create subsets of the data that can be sent to collaborators or released for public viewing.

The XNAT platform consists of three architectural layers; a data archive, a user interface (including the standard web interface as well as desktop tools), and a middleware engine. XNAT uses XML to interface with existing neuroinformatics analysis tools and data repositories, underscoring the possibility of a federated solution for highly integrated collaborative research. XML allows implementations to extend the core XML schema for maximum customization, and supports easy translation between formats.

In its current state, XNAT provides basic support for sanitizing patient data. Investigators must perform this step manually using the DicomBrowser tool, which allows users to view or edit imaging session metadata. It is offered in both a graphical and command-line interface.

The flexibility of XNAT allows researchers to upload data in multiple formats with varying study parameters; however, this presents a challenge when standardizing a sanitization process, requiring the end user to tailor the anonymization process to a particular study or session through anonymization scripts. Anonymization scripts are based on the DICOM tag key-value pairs. The language allows specification of operations to be performed on the attribute pairs and constraints by which the operations are applied. Allowed operations are assignment to set a default value for the attribute and deletion.

LONI De-identification Debablet

The LONI De-identification Debablet is a software tool capable of reading medical images in a variety of formats and stripping them of an personally identifying information [6]. This tool is built on top of the LONI Debabeler execution engine, which facilitates translations between DICOM, ANALYZE, MINC and other medical image formats.

The De-identification Debablet It replaces a subject identifier with a user-defined identifier to permit anonymize tracking of datasets across subject populations. While it successfully de-identifies the images, this technique is suitable for localized use only; no solution is provided for decentralized neuroimage dataset tracking and management. Thus, a researcher must define, manage and maintain identifying links between original and sanitized images. The principal limitation herein is that this approach has no inherent method for identifying returning subjects or for recognizing datasets from the same subject.

Additional complications in dataset sharing and management are encumbered by this scheme. There may be other identifying features that an investigator wishes to alter or include, depending on the nature of the study and the sensitivity of the information. The investigator may also desire a more robust solution with expanded auditing capabilities to track the removal of data and to allow other authorized collaborators access to it.

Significance of Proposed Effort

The proposed solution satifies HIPAA requirements regarding subject privacy while maintaining ease of use for the researcher. We offer an automated approach to redacting patient data for use with XNAT as a neuroimage repository. This project has two primary advantages over existing measures: (1) it provides a privacy map for linking and managing redacted identifiers across institutional boundaries, and (2) it is part of the XNAT processing pipeline and thus a natural part of collaborative workflow.

Our solution seamlessly creates a privilege map of redacted subject data, creating pseudonymous identifiers to associate with subject datasets. Cross-linking and validation of these new identifiers is managed systematically by middleware, and integrated transparently into the data curation and resource sharing workflows. This permits persistent identifiers which, although they cannot be used to uniquely determine an individual's identity, can be used to track the datasets associated with an individual (anonymous) subject.

The primary responsibility for ensuring patient privacy still lies with the researcher, but the proposed solution minimizes the effort involved while enhancing the quality and utility of subject data. By integrating redaction within collaborative workflow, privacy management becomes a more systematic and intuitive process.

Preliminary Studies

Redaction is the process of removing privileged information from a document or set of documents before its presentation to other parties. The reasons for redaction are many and varied. The same concerns exist for privileged information residing on electronic storage devices and in electronic, but no standard method of digital redaction has been adopted by the legal community. Computerized methods that mimic the blackout process exist, as do those for mimicking the physical removal method. The latter typically involves the collection of all readable documents from a computer, placing them in a set, and selecting the items to redact. Yet, while electronic blackout and removal methods can sanitize a document or set of documents found on an electronic device, they do nothing to redact logical copies or copied fragments of the document that remain.

While tools for digital redaction have focused almost exclusively on electronic documents, a more general solution remains elusive. This can, in part, be attributed to the complexities associated with a single solution operating over the universe of possible data structures. [2] accounted for a number of obstacles to effective general purpose digital redaction, specifically:

- **The variety and disparities in electronic storage devices:** A multiplicity of electronic storage devices conceivably implies a multiplicity of data storage formats, each with its own nuances and peculiarities. Accounting for every existing and emerging format is a daunting task, and one that defies a singular solution.
- **Encrypted data:** The rising use of encryption has troubled law enforcement and intelligence communities for years. But the emergence and growing popularity of encrypted file systems has propelled this concern to forefront of cyber-law. One danger is in producing an encrypted data image that cannot be inspected, only to later learn that the receiving party was able to break the encryption and discover privileged information.
- **Deleted but recoverable files in slack or free space on a file system:** Conventional digital redaction tools, which largely focus on electronic documents, do not adopt a file-system centric mode of operation. Thus, for file systems that harbor hidden data in free or slack space, duplicate data (some possibly containing privileged information) can escape the redaction process.
- **Data fragmentation:** Data fragmentation is a ubiquitous phenomenon among file systems. Yet naive redaction tools overlook privileged data that spans fragments. An application-level redaction approach may avoid such issues, but not if the data exists in slack space.
- **Isolation of privilege by context for integrated data:** Privileged information that resides in fully integrated databases can be influenced or revealed by their context (e.g., schema) in the system. Again, the naive redaction process may overlook an important target for redaction if it lacks an understanding of the system in which the data resides.

The proposed effort addresses one aspect of the larger redaction problem, and is specifically intended to facilitate large-scale inter-organizational collaboration [1]. The requirements of HIPAA necessitate a higher level of understanding of the presentation of the underlying data structure for correct and complete data removal. Thus, data must be deconstructed and considered in each of the different layers in which it can reside – hardware, operating system, software, and the data analysis stack. The proposed solution offers such an expansive view of PII in neuroimagery files, and integrates the redaction process within established data acquisition and curation workflow patterns.

The XNAT platform has been studied as a suitable candidate framework within which to embed a comprehensive digital redaction solution. In particular, 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 [8]. It also afforded researchers at TU the opportunity to become familiar with the inner-workings of this popular neuroimage archival system. The risk assessment therefore provides an excellent basis on which to define operational and security requirements for an integrated digital redaction solution.

Research Design and Methods

Conceptual framework and architecture

The proposed effort can be divided into three primary components: (1) the processing pipeline, (2) the redaction engine, and (3) the privacy mapping database.

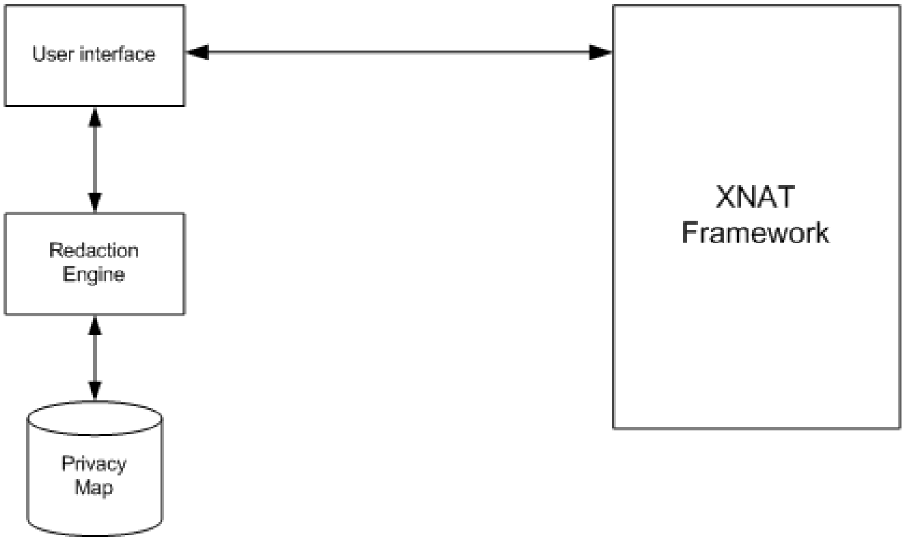

Figure 1: Preliminary architecture diagram.

ALEX: To do: Processing Pipeline and a note about REST

The redaction engine performs complete and verifiable redaction of DICOM files. The researcher will be able to edit and redact any DICOM fields through the use of configuration files, but the primary focus is the list of identifiers defined in HIPAA that can be used to uniquely identify the subject. The engine will assign a unique redacted identifier to the image and the subject, though users can enter this manually for legacy data sets. The primary function of the redaction engine is fully prepare neuroimages for external use by removing any identifiable patient data.

As the neuroimages are redacted, the XNAT translator will build the required XML files according to the XNAT schemas to create the new, de-identified sessions and subjects (if necessary) within XNAT. The translator will communicate with XNAT through the REST API. Redacted images are then transferred to XNAT through SFTP. The use of SFTP allows the redacted data set to be stored with both a logical and a physical separation of redacted data and unredacted data.

The privacy mapping database is an integral part of the system, containing the link between the original images and their redacted counterparts. It holds the unique identifiers (subject, session, image, etc.) for the data to create a linkage from a personally identifiable data set to a redacted data set that can be shared. The privacy map also offers a mechanism through which users may track a single subject throughout multiple scanning sessions. Unlike systems that only perform local data sanitization, the privacy map ensures that the subject identifier is consistent between multiple sessions and among researchers, so that multiple researchers can reference the same redacted identifier for a subject with minimal effort and maintenance.

Data workflow

The proposed method is builds upon the XNAT workflow, as seen in Figure 2. For our redaction engine, the workflow process begins at acquistion time. During a neuroimage scanning session, the raw data is collected and stored according to the institution's local policy. The redaction process is flexible and can be inserted into the established workflow, as long as it occurs prior to the collaboration phase. It may be desirable to include some of the embedded subject data (such as age, weight) in the analysis; in this case, the researcher may choose to include some of the data, as long as it cannot be used to uniquely identify the subject.

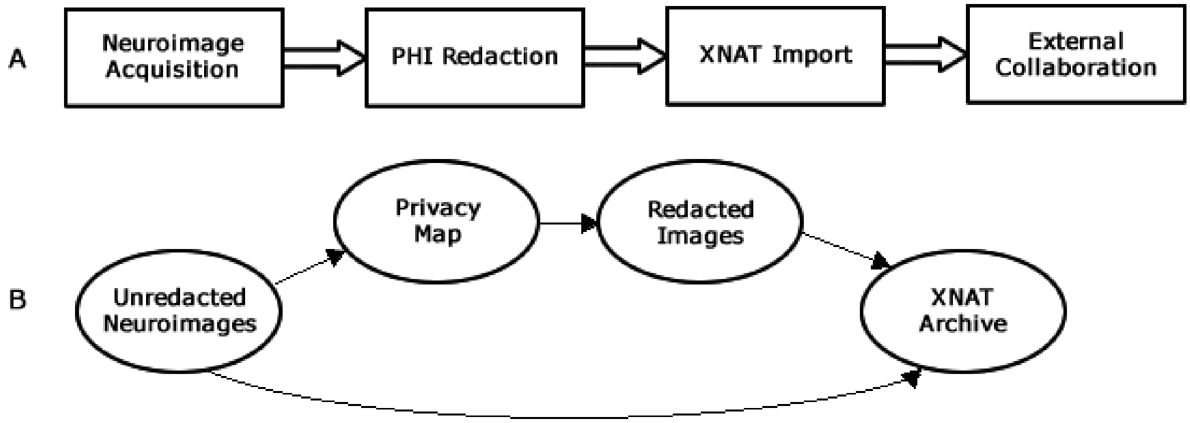

Figure 2: Neuroimage redaction workflow. (A) Process workflow as performed by the end users. (B) Data workflow as it follows the process in part A.

As the images are redacted, a privacy map is built to store the link between the original neuroimage and its redacted counterpart. The redacted images are then uploaded to XNAT as new images. If the subject exists in the privacy map, then the corresponding XML document is generated with the existing redacted identifier; otherwise, a new redacted identifier is assigned. These XML documents are sent to XNAT through the REST API, eliminating manual data entry.

ALEX TO DO: revisit this according to new architecture

Advantages over Current Solutions

The proposed effort has two primary objectives: (1) it provides a privacy map for linking and managing redacted identifiers across institutional boundaries, and (2) it is part of the XNAT processing pipeline, leveraging the existing framework and interface.

The benefit of the privacy map is apparent; it automatically assigns a redacted identifier to subject record and removes potentially identifable information. This relieves the researcher of the burden of maintaining the subject's new identifier throughout a study. The persistent sanitized identifier allows multiple sessions for the same subject to be tracked without the need for manual maintenance.

Researchers will not have to maintain an original image and a redacted image, and collaboration is facilitated through XNAT to ensure that external users can only view the public (redacted) data set.

Potential Challenges

Because this project focuses on redacting PII meta-data from DICOM files, we are only considering one layer in the data stack. This leaves redaction unresolved at the hard drive and file system levels, though with proper physical access controls to data centers and personal computers, the potential of a breach at this level is less immediate than at the logical DICOM level. Since we are focusing on the DICOM layer, though it is desirable to be able to redact data from other formats as well. The system will be flexible enough that future revisions can include other image formats. There is also concern over the possibility that a subject's full facial image can be reconstructed from a sufficient amount of structural data; we do not attempt to address this in our proposed effort.

Though this project will reduce the effort required for redaction, the ultimate responsibility for redaction still lies in the hands of the researcher. It is still possible to upload an unredacted data set to the XNAT repository and share it. This risk can be reduced by the use of default redaction profiles to automatically de-identify any data set uploaded through the user interface. Proper training and policy enforcement can reduce the likelihood of an accidental breach of privacy.

Timetable

References

[1] A. Barclay, N. Schimke, and J. Hale. Comprehensive neuroimage redaction. Poster at USENIX Security 2009, Montreal, Canada, August 2009.

[2] A. Barclay, L. Watson, D. Greer, J. Hale, and G. Manes. Redacting digital information from electronic devices. In P. Craiger and S. Shenoi, editors, *Advances in Digital Forensics III*, pages 205–214. Springer, 2007.

[3] J. D. Van Horn, J. S. Grethe, P. Kostelec, J. B. Woodward, J. A. Aslam, D. Rus, D. Rockmore, and M. S. Gazzaniga. The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. *Philosophical Transactions: Biological Sciences*, 356:1323–1339, 2001.

[4] S. H. Koslow. Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neuroscience*, 3(9):863–865, September 2000.

[5] D. Marcus, T. Olsen, M. Ramaratnam, and R. Buckner. The eXtensible Neuroimaging Archive Toolkit: An informatics platform for managing, exploring, and sharing neuroinformatics data. *Neuroinformatics*, 2007.

[6] S. C. Neu, D. J. Valentino, and A. W. Toga. The LONI Debabeler: a mediator for neuroimaging software. *NeuroImage*, 24:1170–1179, 2005.

[7] P. Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *SSRN eLibrary*, 2009.

[8] N. Schimke, N. Singleton, and J. Hale. XNAT security assessment. University of Tulsa, Technical Report, July 2009.

[9] G. Shepherd, J. Mirsky, M. Healy, M. Singer, E. Skoufos, M. Hines, P. Nadkarni, and P. Miller. The Human Brain Project: neuroinformatics tools for integrating, searching and modeling interdisciplinary neuroscience data. *Trends in Neuroscience*, 21(11):460–468, November 1998.

[10] Editorial Staff. A debate over fMRI data sharing. *Nature Neuroscience*, 3(9):845–846, September 2000.