A System Architecture for Sharing De-Identified, Research-Ready Brain Scans and Health Information Across Clinical Imaging Centers

Ann L. Chervenak[a], Theo G.M. van Erp [b], Carl Kesselman[a], Mike D'Arcy[a], Janet Sobell[c], David Keator [b], Lisa Dahm[d], Jim Murry[e], Meng Law[f], Anton Hasso[g], Joseph Ames[b], Fabio Macciardi[b], Steven G. Potkin[b]

[a]*University of Southern California Information Sciences Institute*
[b]*University of California at Irvine Department of Psychiatry and Human Behavior*
[c]*U. of Southern California Department of Psychiatry and Behavioral Sciences*
[d]*U. of California at Irvine Institute for Clinical & Translational Sciences*
[e]*U. of California at Irvine Health Affairs Information Services*
[f]*U. Southern California Keck School of Medicine, Department of Neuroradiology*
[g]*U. of California at Irvine Department of Radiology*

Abstract. Progress in our understanding of brain disorders increasingly relies on the costly collection of large standardized brain magnetic resonance imaging (MRI) data sets. Moreover, the clinical interpretation of brain scans benefits from compare and contrast analyses of scans from patients with similar, and sometimes rare, demographic, diagnostic, and treatment status. A solution to both needs is to acquire standardized, research-ready clinical brain scans and to build the information technology infrastructure to share such scans, along with other pertinent information, across hospitals. This paper describes the design, deployment, and operation of a federated imaging system that captures and shares standardized, de-identified clinical brain images in a federation across multiple institutions. In addition to describing innovative aspects of the system architecture and our initial testing of the deployed infrastructure, we also describe the Standardized Imaging Protocol (SIP) developed for the project and our interactions with the Institutional Review Board (IRB) regarding handling patient data in the federated environment.

Keywords. Architecture, MRI, brain, imaging, standardization, federation, sharing, open source, de-identification, protocol, health, pilot, infrastructure, hospitals

Introduction

Understanding complex brain disorders increasingly relies on the costly collection of large standardized brain magnetic resonance imaging (MRI) data sets. Clinical MRI scanners collect, on average, approximately 2,500 clinical brain scans per year. Due to the use of different scanners and imaging protocols between imaging facilities, and the lack of optimization of scan protocols for research purposes, these scans are of limited use for research studies. Moreover, clinical interpretations of a patient's scan can benefit from compare and contrast analyses with scans from patients with similar, and sometimes rare, demographic, diagnostic, and treatment status as well as quantitative measures (e.g., brain volumes) that are commonly compared in research studies.

A solution to both clinical and research needs is to acquire standardized, research-ready clinical brain scans and to aggregate these scans across multiple imaging facilities. The resulting research-ready brain imaging resource would provide a wealth of accessible standardized brain imaging data relevant to patient care and research. Clinical applications would benefit from the ability to compare brain scans, while the availability of more extensive sets of images of sufficient quality for research would enable hypothesis testing that would not otherwise be possible.

To address these needs, we adopted imaging protocols for standardized, research-ready clinical brain scans, and designed, built and deployed a distributed infrastructure to share de-identified scans along with other pertinent clinical information across institutions. Informatics solutions taken in isolation of the clinical and policy environment in which they are utilized frequently fail. For this reason, we have taken a systematic approach with both technical and clinical stakeholders to achieve the following goals: (1) to define a standardized brain imaging protocol, applicable to a significant subset of the patients who receive clinical brain MRI scans in imaging facilities; (2) to develop general, open source, end-to-end software infrastructure to securely store and manage that imaging data for research purposes at each participating institution; (3) to develop general, open source, end-to-end software infrastructure to support data federation from image capture to distributed image query and download; (4) to deploy a pilot federation across two institutions, the University of Southern California (USC) and the University of California at Irvine (UCI), that supports cohort identification and retrieval of images based on subject characteristics; and (5) to explore models for patient consent with medical ethics teams from the pilot institutions.

We present a brief description of the issues that drove the design of our system, followed by descriptions of the image federation system architecture and the design and implementation of key software components. We conclude with a discussion of the current status of the pilot deployment, related work and plans for future work.

1. Design Issues for the SIP Pilot

Next, we describe several issues that drove the design of our system, including issues related to data federation and the design of the Standardized Imaging Protocol (SIP).

1.1. Federation Issues

Creating a data sharing federation across the USC and UCI hospitals required commitments from many people at each institution. The dean of each medical school committed to data sharing and regional cooperation. Institutional Review Boards (IRBs) at both institutions approved the pilot feasibility study. Radiologists and researchers at the UCI and USC medical schools worked together to define the Standardized Imaging Protocol. The agreed upon architecture took into account differences in hardware and software infrastructure at the two hospitals. We agreed on a data model for the collected information and on a strategy for de-identifying images by determining which patient attributes must be removed before sharing and federation of imaging data. We noted that most clinical scanners, Picture Archiving and Communication Systems (PACS), and Radiology Information Systems (RIS) use the Digital Imaging and Communications in Medicine (DICOM) protocols; we leveraged DICOM as the means for interfacing with clinical systems and defining the data model.

Finally, we obtained hardware, software and network resources and deployed the services for imaging data collection, query and access at both institutions.

1.2. Standardized Imaging Protocol

To maximize the potential for clinical and research use, we have defined a standardized imaging protocol that was implemented at each of the participating sites. Given that the neuroradiologists's primary responsibility is accurate patient diagnosis, standardization of clinical scans across imaging facilities is impossible without the expertise and cooperation of all neuroradiologists responsible for reviewing clinical brain scans.

Four factors went into the design of the protocol:

1. minimal time impact on most typical clinical brain MRI acquisitions; our target was to limit the acquisition time to between 5 to 10 minutes
2. portability so that sequences and parameters can easily be implemented across the most widely used scanner platforms (General Electric, Philips, Siemens)
3. acquisition of whole brain scans
4. use of high quality sequences that are likely to enable accurate, semi-automated image analyses.

The protocol includes a high-resolution structural scan and a Diffusion Tensor Imaging (DTI) scan because these sequences provide information relevant to most clinical and research applications. The high-resolution structural scan allows for detailed study of anatomical structure and quantification of gray matter, white matter, and cerebral spinal fluid volumes. The DTI scan allows for the study of anatomical microstructure based on the motion of water molecules in the brain and is becoming increasingly valuable in evaluating brain connectivity abnormalities (e.g., in traumatic brain injury). Clinically, the high-resolution scan is used in diagnosing a number of neurological diseases such as brain tumors and demyelinating diseases. The DTI sequence, in addition to producing fractional anistrophy (FA) measures useful for understanding the direction of water diffusion in the brain, produces apparent diffusion coefficient (ADC) and diffusion weighted images (DWI) used in clinical practice to aid in diagnosing acute ischemic stroke. Our initial protocol also includes two B1 calibration scans that can be used to correct for B1 field inhomogeneity.

In addition to the human scans, the SIP includes weekly MagPhan® phantom scans to track scanner stability. An imaging phantom is an object that is scanned for the purpose of analyzing and tuning the performance of an imaging device. Phantom scans can be used to correct spatial distortion and intensity inhomogeneity in images due to scanner variability over time and between sites. We track the sequence of phantom images and associate the immediate, previous, and subsequent phantom scan with each patient scan to facilitate troubleshooting problems that may develop in scan quality. The temporal linkage is done in a manner that prevents leakage of Personal Health Information (PHI). Our human studies are linked to phantom scans by appending the associated phantom Study UID to the DICOM file header. This metadata linkage allows us to identify the closest temporally related phantom study for a human study without the use of dates.

We continue to work on minimizing the time impact of using the SIP. For example, at UCI, the research-based high-resolution scan was adopted instead of the existing clinical sequence, resulting in a negligible addition of 1 minute and 30 seconds of scan time. USC chose to add the high resolution T1 scan to the existing clinical protocol to facilitate image comparison. The DTI/DWI scan, at 3 minutes and 50 seconds, runs in

parallel to the existing DWI scan to make sure it is of equal or greater diagnostic value than the existing scan. The B1 scans (1 minute and 22 seconds) are run when time allows and the patients tolerate the scans. Further development of standardized clinical imaging sequences across scanner vendors will likely require additional effort and funding. Improvements in MRI sequences and the sharing of new and improved sequences across vendors will in the future likely shorten the time to perform research-ready scans so that the benefits of such scans can be obtained without the cost of several minutes of extra time.

2. System Architecture

The system architecture consists of three main functional components: (1) the *DICOM Forwarder,* (2) the *DICOM Image Gateway,* and (3) the *Federated Query Engine/ Mediator* (see Figure 1). We provide a detailed description of each component, followed by a discussion of our approach to protecting personal health information.

2.1. The DICOM Forwarder

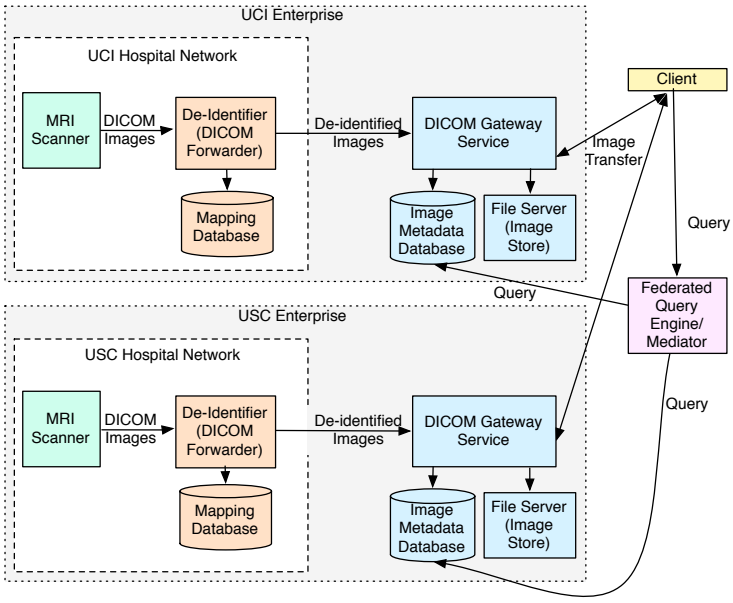

Figure 1: Overall architecture of federated imaging system

The DICOM Forwarder component acts as an interface between the local imaging system (the scanner) and the federated imaging repository. The Forwarder receives DICOM image files pushed from the local scanner, removes patient identifying information from those files in accordance with HIPAA guidelines for human subjects research, and forwards the files to the DICOM Image Gateway component. The de-identification of image data occurs within the institution that collected the data. By design, patient identifying information never leaves the secure network.

The DICOM forwarder component is implemented using a three-stage processing workflow, illustrated in Figure 2. We used the Apache Camel routing and mediation

engine as the basis for our implementation. Files processed by the DICOM Forwarder travel through a Camel routing workflow, where each stage of the workflow polls the directory associated with the previous stage to determine whether new files are ready to be processed.

The first stage of the workflow receives DICOM image files pushed from the imaging system via a standard DICOM C-STORE command.

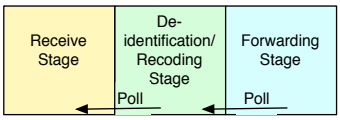

Figure 2: Three-stage Forwarder Workflow

The second stage of the Forwarder workflow is the De-identification/Recoding Stage, which applies a de-identification process to the DICOM header of the image files that removes non HIPAA-compliant fields, replaces other fields with generic "ANONYMOUS" tokens, and remaps identifier fields with newly minted unique identifiers. The Forwarder stores mappings between the original fields and the remapped attributes in a local Mapping Database, implemented with a Postgres relational database. Authorized staff within the secure hospital network may do a reverse mapping from the de-identified images to the patient identifying information if necessary. However, researchers and clinicians using the federated system only have access to the images via the Image Gateway Service and do not have access to any Personal Health Information (PHI). In addition, the Forwarder component removes all scan date attributes in accordance with HIPAA rules. Scan dates for phantom and human studies are preserved in the secure mapping database at each institution as a mechanism for resolving a temporal relationship between a human study and a phantom study. The medical record numbers and scan dates for the human studies stored in the secure mapping table are used to query for IRB approved health information associated with the scan from the Electronic Medical Records (EMR). Our pilot project has approval to obtain diagnosis (e.g., ICD-9), age, sex, race, and ethnicity information. Such data will be retrieved for each scan; the de-identified EMR data will be forwarded to the Gateway Service.

The final stage of the workflow acts as a DICOM client and forwards de-identified images to the DICOM Image Gateway Service. Both the Forwarder and the Image Gateway Service implementations use the PixelMed Java DICOM Toolkit [1].

2.2. DICOM Image Gateway Service

The DICOM Image Gateway Service stores the de-identified MRI images it receives from the Forwarder and supports queries on image attributes as well as retrieval of images for sharing within the federation. The Image Gateway Service consists of four components, as shown in Figure 3: a DICOM Protocol interface, a file server for image storage, an Image Metadata Database, and a web server.

First, the DICOM protocol interface presents a fully compliant DICOM network interface that allows clients to interact with the Image Gateway Service as they would with any DICOM device, such as a PACS system, to query and retrieve stored images. The Gateway supports standard DICOM query and retrieval methods, such as C-FIND and C-GET/C-MOVE. Any DICOM compliant software application (e.g., Osirix [2], eFilm [3]) can act as a client of the Gateway via this interface without modification.

Second, the gateway service stores complete image files it receives from the Forwarder in a file server, where they can be accessed via file transfer protocols.

Third, the Gateway service extracts metadata associated with MRI image files from the DICOM header fields of those files and stores the attributes in a relational Image Metadata Database, where they can be queried to identify image files with specified attributes. The metadata schema for the relational database is based on a subset of the DICOM header fields. The entity relationships of the database follow the DICOM patient->study->series->instance containment model. The metadata for each instance includes a link to the location of the image file on the file server.

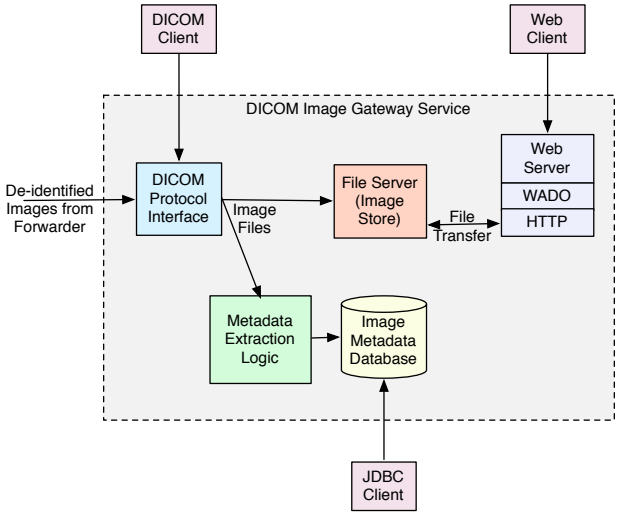

Figure 3: Image Gateway Service Components

Finally, the gateway service includes a web server that provides an alternate means to query and retrieve images via gateway web clients. Web clients query the Image Metadata Database via the Java Database Connectivity (JDBC) protocol and retrieve imaging files via the Web Access to DICOM Persistent Objects (WADO) protocol.

2.3. Federated Query Engine/Mediator

The Federated Query Engine/Mediator component is responsible for accepting user queries for images with certain metadata attributes and for distributing those queries across the institutions in the federation. The query engine submits those queries to the Image Metadata Database component in each DICOM Image Gateway service to identify matching images. The query engine returns the result of the distributed query to the client, which then interacts directly with the DICOM gateway services to download the desired image files. The implementation of the Federated Query Engine/Mediator (Figure 4) integrates three components: a mediator [4], the OGSA Data Access and Integration (OGSA-DAI) Web Service [5], and the Distributed Query Processing Engine (DQP) [6].

The federated query engine issues queries to the Gateway Image Metadata Databases in the federation, which currently have identical schemas. The mediator builds a global, virtual database view from backend databases in the federation; it

presents the federated data to the user as a single set of unified tables, each containing an additional column that indicates which data source in the federation each row came from. An advantage of using the mediator is that it hides some of the complexity of the DQP query interface, allowing users to issue simple SQL queries to the mediator's virtual database view. The Mediator also provides a schema mapping capability; if needed in the future, the mediator will support queries across Image Metadata Databases with different schema.

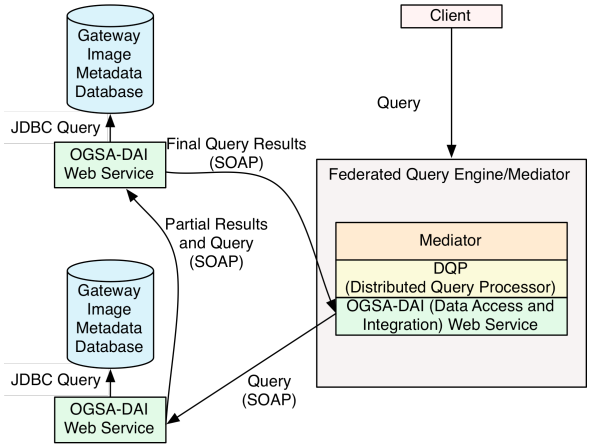

Figure 4: Implementation of Federated Query Engine/Mediator

The OGSA-DAI service provides the secure web service (HTTPS) interface to data resources, including relational databases and file servers. In our architecture, OGSA-DAI web service instances provide access to each gateway's Image Metadata Database. These web service instances are then federated using the DAI Distributed Query Processor (DQP). Client applications interact with a central Mediator service node that performs DQP queries on their behalf and returns a set of unified query results.

As illustrated in the figure, the default configuration of DQP issues a query first to one OGSA-DAI instance using the Simple Object Access Protocol (SOAP) [7], and then forwards the partial results of that query to the next OGSA-DAI instance, and so on, until the final query results are returned to the Distributed Query Processing engine.

2.4. Security Issues

As already noted, a key feature of the pilot system architecture is that patient-identifying information never leaves the secure hospital network. Only de-identified information is shared or queried by the nodes of the federation. This architecture allows for maximum flexibility across the federation as it leaves each site in the federation in control of the de-identified data it can release based on local approval.

A hospital's enterprise consists of the hospital's internal network and its gateway services. Each enterprise is protected by a firewall; clients outside the enterprise may only initiate connections to the gateway service node's web services port. Since DICOM provides very limited support for user authentication, the current deployment scenario limits the direct querying of a hospital's gateway via DICOM protocols to within that hospital's enterprise; it is not possible to issue a DICOM command via the wide area network, since the inbound traffic is blocked by the enterprise firewall.

Interactions between the federated query engine and the gateway web services are authenticated using standard SSL protocols and X.509 credentials and authorized using access control lists. Gateways may query each other's web services, and the federated query engine may query each gateway's web services. Enterprise firewalls prevent connections to gateway ports other than the web services port.

3. DEPLOYMENT AND TESTING

We installed the prototype software infrastructure at both USC and UCI and have begun collecting clinical imaging scans at both hospitals using the Standardized Imaging Protocol. The SIP sequences were installed on a 3T Trio Tim Siemens scanner at UCI and a 3T General Electric scanner at USC. Both sites conduct periodic Magphan® phantom scans (typically once a week) for scan protocol and de-identification procedure evaluation. As of this submission, the hospitals at UCI and USC have collected approximately 79 patient scans. Collection of this amount of imaging data for a typical research study would cost approximately $45,000.

Our initial testing conducted several phantom scans in which MR technicians entered information in comment fields on the scanner to determine fields through which PHI may leak and to verify the correctness of our de-identification algorithm. After the scans passed through the de-identification logic in the Forwarder, we downloaded images from the Gateway using the DICOM protocol interface and examined the downloaded image headers. Based on these reviews, we modified the de-identification algorithm to remove an additional field that lists the location of the imaging sequence that is based on user input during the setup of the scan sequence. In addition to removing publically available DICOM fields that may contain PHI, private vendor DICOM fields are removed such that no PHI can be leaked via those fields.

Using a simple query interface, we have run several queries on the federated imaging data, including queries for all patient identifiers, all studies, and all scans by institution. While we need to provide a more sophisticated and user-friendly query interface for the federated system, these simple test queries demonstrate that the federated query engine/mediator works as expected.

4. RELATED WORK

The federation of imaging databases developed for this pilot project is based in part on the Function BIRN (FBIRN) project [8], which provides each site autonomous control over its imaging data and a shared burden for participating sites of maintaining a local repository while making images available to researchers as if they came from a single resource [9]. We believe that such autonomous control over collected data is in the interest of both clinical imaging centers and patients.

In earlier work, the USC team developed Medicus [10], a wide area, distributed PACS system that stored DICOM images and associated metadata in the grid. Medicus presented a virtual data warehouse model to the user and stored one or more copies of DICOM images at different sites. By contrast, our system federates imaging files and metadata stored at their home institutions and supports federated query and retrieval.

Elger et al. describe the @neurIST project [11]. Like our pilot, their architecture federates data sets stored at multiple institutions by anonymizing and storing data in

repositories at each site that are accessible to researchers. @neurIST also supports direct data access from Clinical Information Systems and on-the-fly anonymization. They also discuss patient consent and data de-identification issues in detail.

The Cross-Enterprise Document Sharing (XDS) system [12] federates images and provides a document sharing interface based on web services. The XDS federation uses centralized metadata, while in our pilot, metadata are distributed in the federation and virtualized into a global data view at query time via the OGSA-DAI and mediator.

The Medical Image Resource Center (MIRC) [13] is a federated library of medical images. MIRC and the RSNA Clinical Trial Processor (CTP) [14] federate DICOM images and manage patient health data. Like our pilot, they include DICOM support and de-identification. MIRC issues XML schema queries against a document metadata index; our pilot issues SQL queries against the mediator's virtual relational schema.

The Extensible Neuroimaging Archive Toolkit (XNAT) [15] is an open source imaging informatics platform. XNAT facilitates common management, productivity, and quality assurance tasks for imaging data. XNAT uses XML schema for document query and metadata formats and RESTful services for query and retrieval. Projects can be federated into a centralized XNAT instance, but there is currently no built-in support for distributed query processing across multiple XNAT installations.

The Medical Imaging Informatics Bench to Bedside (Mi2B2) project [16] integrates DICOM based imaging systems into the I2B2 analytic data repository [17] for purposes of query and retrieval. Mi2B2 uses XNAT for its underlying image management. Our work is distinguished from Mi2B2 in that we provide a federated query mechanism for discovery.

5. CONCLUSIONS AND FUTURE WORK

We learned several important lessons from the development of the SIP prototype. First, deploying a federation across hospitals was challenging and required commitment from neuroradiologists and hospital IT staff at both institutions. A key infrastructure decision was the deployment of the Forwarder/De-identification component inside the hospital network; this eliminated the need to retrieve imaging files through a hospital's firewall.

Our experience with IRB approval for this pilot project at the two universities was relatively straightforward, given the project's status as a feasibility study. However, the next phase of our planned work will integrate the pilot system with clinical systems, which will involve a more extensive IRB approval process and require us to address questions regarding how and when to get written informed consent.

In developing the Standardized Imaging Protocol, we found that neuroradiologists at both institutions agreed in principle on the value of a standardized, research-quality imaging protocol and that there was significant alignment in the imaging protocols at the hospitals, which simplified the standardization effort. However, in practice, neuro-radiologists may be reluctant to perform additional scans unless they are convinced of their clinical utility and that the availability of additional scans does not create potential malpractice liabilities. A key SIP feature is the use of B1 and periodic phantom scans that allow corrections of spatial distortion and intensity inhomogeneity in images.

This pilot project reflects a commitment by both USC and UCI to create a Southern California alliance of Clinical Translational Science Institutes. Our goal is to expand this pilot, first to a small number of additional institutions, and eventually to

regional and national levels. We are in the process of linking our federated image repository to IRB approved data resources containing patient information.

Based on our experience, future activities will focus on: a) extending the SIP to include additional scan sequences (preferably as substitutes for sequences already in use), b) deploying the federation at other clinical imaging facilities with different hardware, software, and network infrastructure, and c) linking additional data resources containing patient information, subject to Institutional Review Board approval.

6. ACKNOWLEDGMENTS

This work was supported in part by the NIH NCRR grant Biomedical Informatics Research Network (NIH U24 RR025736-01); by the UCI Institute for Clinical and Translational Science (NIH RR031985-01); by a matching grant from UCI's Health Affairs Information Services; by the USC Clinical Translational Science Institute (NIH 1 UL1 RR031986-01); and by the BIRN Community Service Award (NIH U24 RR026057). We thank UCI MR Supervisor Dharmendra Patel for his assistance in implementing the SIP and Drs. Anton Hasso, Jason Handwerker, and Fred Greensite at UCI and Dr. Meng Law at USC for their help in developing the SIP.

References

[1] "PixelMed Publishing," http://www.pixelmed.com/.

[2] "OsiriX Imaging Software: Advanced Open-Source PACS Workstation DICOM Viewer," http://www.osirix-viewer.com.

[3] "eFilm Solutions," MERGE Healthcare, https://estore.merge.com/na/index.aspx.

[4] N. Ashish, J. L. Ambite, M. Muslea, and J. A. Turner, "Neuroscience Data Integration through Mediation: An (F)BIRN Case Study," *Front Neuroinform,* vol. 4, p. 118, Dec 28 2010.

[5] M. Antonioletti, et al., "The design and implementation of Grid database services in OGSA-DAI," *Concurrency and Computation: Practice and Experience,* vol. 17, pp. 357-376, 2005.

[6] M. N. Alpdemir, et al, "OGSA-DQP: A service for distributed querying on the grid," in *Advances in Database Technology, 9th Int'l Conf. on Extending Database Technology* Heraklion, Greece, 2004.

[7] "Simple Object Access Protocol (SOAP)," http://www.w3.org/TR/soap/.

[8] "Function BIRN," http://www.birncommunity.org/current-users/function-birn/.

[9] D. B. Keator, "Management of information in distributed biomedical collaboratories," *Methods Mol. Biol: Biomedical Informatics,* vol. 569, pp. 1-23, 2009.

[10] S. G. Erberich, et al., "Globus MEDICUS-Federation of DICOM Medical Imaging Devices into Healthcare Grids," *Studies in Health Technology and Informatics* vol. 126, pp. 269-278, 2007.

[11] B. S. Elger, et al., "Strategies for health data exchange for secondary, cross-institutional clinical research," *Computer Methods and Programs in Biomedicine,* vol. 99, pp. 230-251.

[12] "IT Infrastructure Technical Framework," Integrating the Healthcare Enterprise (IHE) http://www.ihe.net/Technical_Framework/index.cfm#IT.

[13] "Medical Imaging Resource Center," http://mircwiki.rsna.org/index.php?title=Main_Page.

[14] "CTP-The RSNA Clinical Trial Processor," http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Processor.

[15] D. S. Marcus, et al., "The extensible neuroimaging archive toolkit," *Neuroinformatics,* vol. 5, pp. 11-33, 2007.

[16] "Medical Imaging Informatics Bench to Bedside (mi2b2)," https://community.i2b2.org/wiki/display/mi2b2/mi2b2+Home.

[17] S. N. Murphy, et al., "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," *JAMIA,* vol. 17, pp. 124-130, 2010.