The Application of a Service-Oriented Infrastructure to Support Medical Research in Mammography

Christopher Tromans[1], Sir Michael Brady[1], David Power[2],
Mark Slaymaker[2], Douglas Russell[2] and Andrew Simpson[2]

[1]Wolfson Medical Vision Laboratory, Department of Engineering Science, University of Oxford, Parks Road, Oxford, UK, OX1 3PJ.
[2]Computing Laboratory, University of Oxford, Parks Road, Oxford, UK, OX1 3QD.

Abstract

We describe the application of a framework for secure access to, and sharing of, data from disparate data sources. The approach ensures that data is shared only in circumstances permitted by the data owner. In a healthcare context, such circumstances are often limited by legal and ethical concerns. Moreover, our approach allows for the aggregation of data from disparate data sources. The approach offers the potential to contribute to the ideal of fully connected healthcare. The specific use case of Digital Mammography is discussed.

Contents

Since May 2006, the authors (and other colleagues) have been working on a collaborative research project, GIMI (Generic Infrastructure for Medical Informatics), which aims to develop a secure computer infrastructure to support medical applications. The project consists of a middleware development team as well as three application teams, with the focus of the middleware being the facilitation of secure and ethical aggregation of distributed data from remote sources. The middleware must be interoperable in the sense that it is able to interface with the wide diversity of existing (and future) health information systems

in place around the world, and to provide both secure access and secure transfer of data. The middleware is based on the principles of [1], and its implementation is termed *sif* (service-oriented interoperability framework) [2].

The application teams are contributing to the validation of *sif* through the input of requirements and evaluation of prototypes. An iterative development cycle is helping to ensure that the framework meets the needs of a wide variety of applications. Within the project, the applications are: the use of the technology in studying patients suffering from long-term conditions such as diabetes and asthma; mammography training and auditing; and image analysis for cancer research. This paper describes the latter application, and particular attention is given to the analysis of mammographic images. Although this paper focuses on a particular application, we stress that the *sif* architecture is generic in that it can expose data from a wide variety of data sources.

Medical research, in particular studies of an epidemiological nature, or those in which an algorithm needs to learn from a sufficiently comprehensive training set, require large datasets in order that sufficient statistical power be attained, that is it should include a comprehensive sampling of the total, clinically significant variation present within the population. Since it is extremely rare that a single institution can provide such a comprehensive sample, such studies typically necessitate multiple institutions, spread sparsely geographically. (A particular example of this is a phase 3 clinical trial of a drug, or PMA testing of a new device or software system.) In such circumstances, an efficient and secure data transfer infrastructure is essential if data is to be shared effectively in real time (which, for example, it is typically not in the case of a phase 3 clinical trial). In this paper we describe such a study we are preparing to conduct for the evaluation of a technique we have been developing over the past five years for measuring radiodensity in mammographic images. The image processing algorithm is first to be assessed through synthetic phantom experimentation: the imaging of test objects for which the radiodensity is known precisely, and thus the accuracy of the technique in describing reality may be established. When the experimental results deem the performance of the normalisation software to be acceptable a study will commence with our clinical partner on a 1.25 terabyte patient dataset [3]. The dataset consists of all the digital mammographic images acquired over a two year period at their clinic (using an IMS Giotto full field digital mammography unit), and it is hoped that these will all be processed and subsequently analysed. Analysis will aim to identify any relationship between the observed normalised radiodensity measures and the underlying pathologies of both tumour and healthy tissue, as well as the assessment of risk as currently considered in breast density. Currently, radiologists identify possibly malignant lesions through “contrast features”, for example dense lines emanating from a density (termed spicules) suggesting invasion of the surrounding tissue by a malignancy, or small bright round densities, which are possible microcalcifications. The use of radiodensity adds yet another weapon into the armoury to verify that the feature identified is in fact that which is suspected. For example, in the case of a microcalcification, a measurement of the radiodensity in the small region of interest, relative to its surroundings, can be checked to ensure the feature is exhibiting the likely attenuation properties of calcium. This paper describes the use of *sif* to support the continued development of this algorithm, and, in particular, its clinical assessment and validation.

1 Measuring Radiodensity in Mammographic Images

For several decades the correlation between radiological features of the breast and the likelihood of the breast containing, or subsequently developing, a malignant lesion, has been studied in the field of research termed breast density. Work in this area was pioneered by Wolfe in 1969 [4] who proposed a four category classification for assessing mammographic parenchymal patterns: in particular the prominence

of ductal patterns and the occurrence of connective tissue hyperplasia. Wolfe presented findings showing that each of the four groups, from lowest to highest density, had breast cancer incidence rates of 0.1, 0.4, 1.7 and 2.2 [5]. Since Wolfe's pioneering studies, much work has been contributed to the field by many different authors. However, the most sustained and significant of these contributions is that made by Norman Boyd and his colleagues. In 1982 Boyd et al [6] defined a six category classification (SCC) which focuses on mammographic hyperplasia, and through the use of this assessment technique reported good results in discriminating between images on the basis of their likelihood of acquiring breast cancer. However, these measures suffer a shared limitation, namely reader subjectivity: one highly experienced radiologist may place an image in one category, whilst another may argue that it falls into a different one. This led Byng et al [7] to develop an interactive thresholding technique to segment, and thereby quantify, mammographic hyperplasia. Such measures are termed "area measurements" since they treat the mammogram as a 2D image, ignoring the third dimension (depth through the breast), and treat the projected image as entirely representative. In fact, mammograms are integrated x-ray attenuation maps of the breast, which is tightly compressed to spread dense tissue patches that may be indicative of cancer. A mammogram results from projecting the real 3D compressed breast to a 2D image, thereby losing all depth information between the plates.

To take account of the three-dimensional breast, "volumetric measurements" of breast density have been developed. Such measures quantify the tissue present in the cone between a detector pixel, and the x-ray focal spot, using the likely x-ray attenuation coefficients of the constituent tissues. In 1996 Highnam and Brady proposed [8] the h_{int} representation which measures volumetric density. They developed a model of image formation considering the path of x-ray photons from point of emission in the x-ray tube, to absorption at the detector. Several alternative techniques of measurement have been subsequently proposed, for example Kaufhold et al [9], whom approximate a transfer function describing imaging formation gleaned from tissue equivalent phantom images.

Inspired by Highnam and Brady's work [10], we have developed a second generation of their model [3, 11, 12]. The extra power made available by modern computers, and advances in medical physics, has enabled the removal of many of the simplifying assumptions in their model, leading to a more comprehensive image formation model. The developed algorithm consists of an image formation model considering the production of x-ray photons in the tube, their interactions through absorption and scattering phenomena within the breast, the selective absorption process occurring dependant on angle of incidence as they pass an anti-scatter grid, and their subsequent detection at the image receptor pixel. This model is used to calculate the equivalent thickness of reference material required to create the observed intensity at each pixel within the image, independent of scattered radiation. This thickness, divided by the ray traversal distance, yields the normalised radiodensity measure, which is thus independent of the incident photon flux characteristics, the detector response, and the breast thickness. To overcome the limitations in anatomical representation we highlighted in [12] an alternative normalised measure has been adopted in which the attenuation of the breast per unit traversal distance is compared to that of a reference material. This is analogous to the universally used Hounsfield unit (HU) in Computed Tomography (CT). Figure 1 shows one of the tissue equivalent test objects used for validation in a partially assembled state, and the clinically installed IMS Giotto full field digital mammography unit.

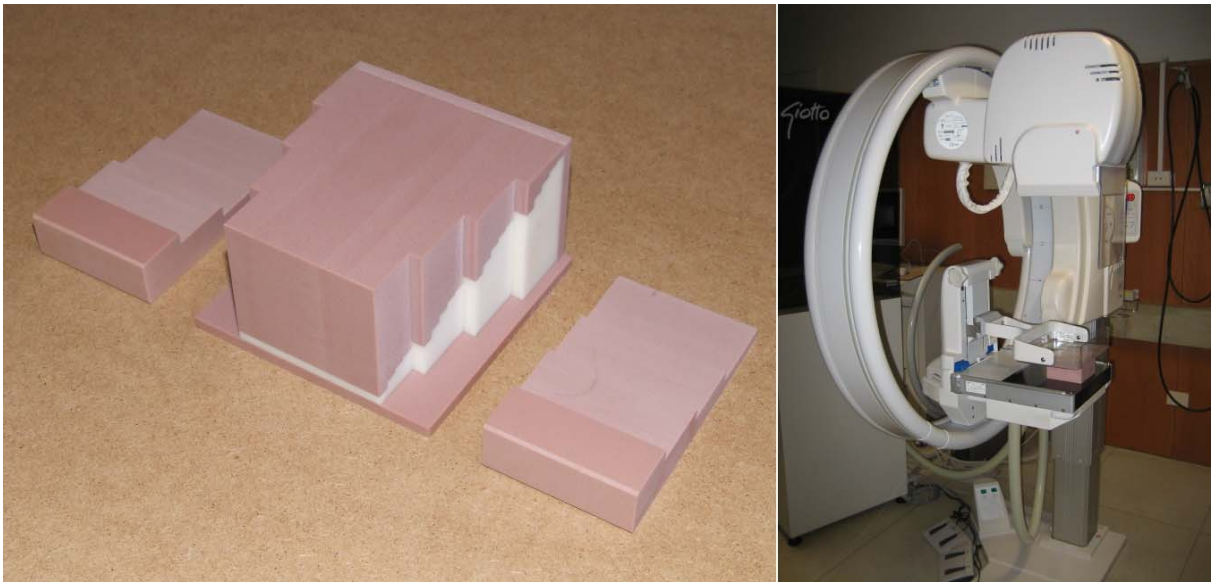

Figure 1 The mammographic test object manufactured from tissue-equivalent resins used for algorithm validation, and the IMS Giotto MD.

2 The *sif* Architecture

The *sif* architecture was developed to fulfil the need to share distributed data in a secure, federated and data-agnostic fashion. Its design was influenced by the eDiaMoND [13] and MammoGrid [14] projects, both of which the first two authors were involved in.

Following [1], *sif* has a clear requirement to facilitate fine-grained access control to data sources, with policies being determined locally by data owners; there is also a need to guarantee secure transfer between endpoints. A key driver has been the joining of data sources to provide "bigger and better" research and healthcare: fundamental to this is the presentation of a federation of multiple data sources as a single entity as described in [15]. In order to develop a generic solution we have taken a data-agnostic approach, by which we mean that *sif* is capable of "plugging in" to a variety of data sources regardless of underlying technology or data format. (See [2] for an overview of the motivation for sif.)
It is, perhaps, worth reflecting upon the philosophy behind, and the current status of, the middleware.

Some of the requirements that have informed the design and development are described below.

- *Facilitating buy-in*: In the 'real world', complexity and cost are factors that significantly influence procurement decisions. Yet, they are often not mentioned in the health grid context. It has been important to us that researchers should be able to utilise data sources---without requiring them to be ported to new operating systems or database management systems, or requiring the data to be transformed into a new structure to facilitate interoperability. This is contrary to the approach adopted in the eDiamond and Mammogrid projects, and the "Grid image workflow paradigm" followed by Globus MEDICUS [16], where fixed data schemas have been designed and implemented within which the DICOM data is stored upon the grid nodes themselves. It has also been important to us that application developers should be able to 'pick up and play with' sif: to this end, data is accessed through a simple API.

- *Abstraction*: Through experience, we have come to the conclusion that it is only through a loosely-coupled approach that the delivery of more genuinely generic solutions are possible. Thus, one of our drivers is the facilitation of technology-agnosticism for application developers via the data-agnosticism philosophy of *sif*.

- *Interoperability*: If one were to take a simplified view, one might characterise the issue of data interoperability as facilitating both *database interoperability* (between Dr Smith's breast cancer research database in San Francisco and Dr Thomas' colorectal cancer research database in New York) and *database management system interoperability* (between the IBM DB2 database utilised by Dr Smith and the Oracle database utilised by Dr Thomas). Our concern is the latter; the issue of what we might term semantic interoperability is left to application developers. This leads to a 'bottom-up'---as opposed to a 'top-down'---construction of virtual organisations, which we shall explore in the next section.

- *Security*: In a health grid context, one can think about security in terms of storage, access and transfer. With respect to a *sif* deployment, the responsibility for secure storage resides with the data owner; as such this is not a concern here. Secure access and transfer, are, however, essential concerns. With respect to the former, access policies are constructed by data owners; with respect to the latter, secure channels are established between external nodes. The requirements for secure access and transfer have been influence by relevant UK and European legislation.

At the heart of *sif* is a plugin mechanism, through which the goal of interoperability across diverse systems is realised. Each plugin is capable of connecting to a particular resource, presenting a standardised interface to the middleware, which in turn provides a federated view to applications. All communications between the resource and the client pass through the middleware using the standard interfaces provided. The middleware thus provides a decoupling between client and resource: that is, the exact details of the resource are abstracted away from the client. The middleware, acting as an intermediary, provides encrypted client connections to ensure secure data transfer across public and private networks, as well as client user identification and authorisation. The identification mechanism utilises secure signed certificates, and once authenticated, a user is subject to the data access restrictions set by the data owner.

There are three types of plugin:

- *Data plugins* are concerned with interfacing queryable data sources to the middleware. Examples of such data sources include relational databases and XML databases.

- *File plugins* are concerned with interfacing file systems, be they local file systems, or network file systems (such as NFS or CIFS).

- *Algorithm plugins* are concerned with the facilitation of the (possible remote) execution of algorithms on data. Examples of algorithms include basic image processing or reconstruction routines.

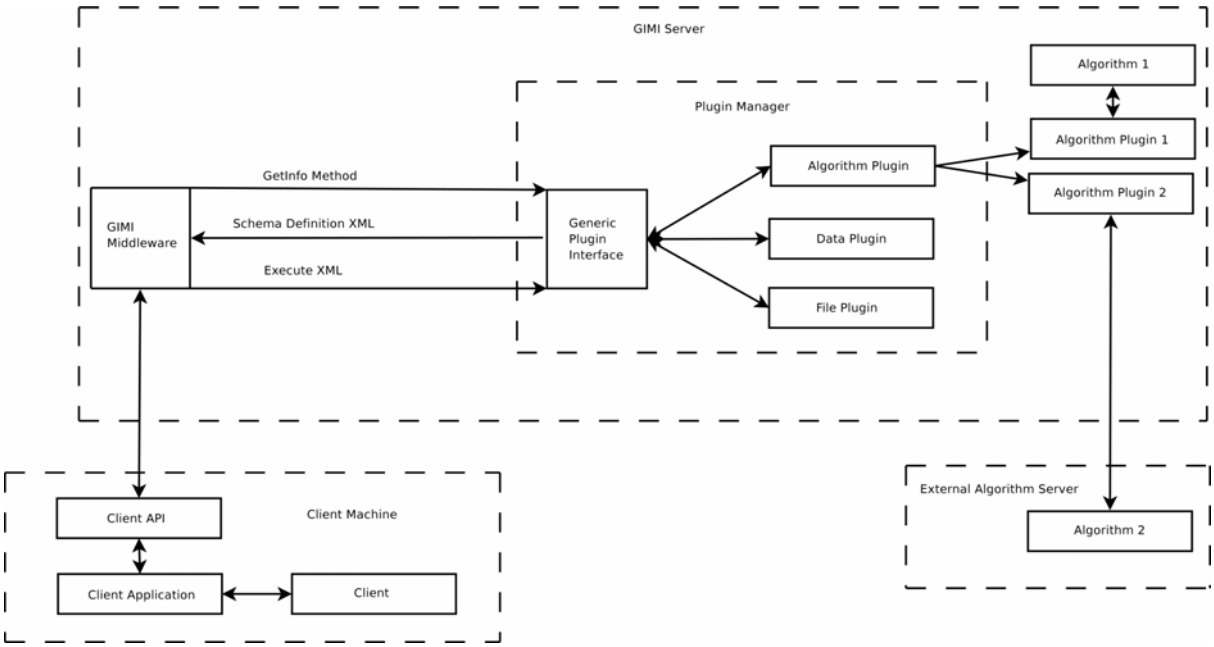

Figure 2 An overview of the plugin architecture.

Through the adoption of a standard interface for each of the three types of plugin it becomes possible to add heterogeneous resources into the infrastructure. These can be managed by the application developer who has the ability to install, uninstall, and update plugins. Importantly, there is no need for the resource being advertised through the plugin system to directly represent the physical resource. What is advertised as a single data source may come from any number of physical resources, or even another distributed system.

Figure 2 shows the overall architecture of the plugin system, with particular focus on two algorithm plugins. A client is interacting with an application running on their local workstation. The application communicates with the server via the sif client API, which is standard for all applications. The application uses the *GetInfo* method to obtain the definition of the algorithm input and outputs which are represented as XML schemas. The client provides appropriate inputs which are then passed through the *sif* API to the server as an XML document. The requested algorithm plugin is then executed with the given inputs contained within the XML document. In the example diagram, Algorithm 1 is executed locally and Algorithm 2 is executed on a remote machine. In the case of Algorithm 2 the job of the plugin is to communicate with the external algorithm server. The results of the execution (which must be consistent with the output schema) are returned to the client application which then handles them as appropriate.

In a distributed context, *sif* acts as a federation layer: if a user runs a query across several data nodes, then the middleware will distribute that query to the nodes and aggregate the results. However, the reason that sif can expose a wide variety of data sources is that it makes no assumptions about structure or semantics: while sif facilitates distributed queries, it is up to the end-user (or application) to ensure that the queries (and results) are meaningful. This, of course, makes the task of federation much easier. Other issues, such as guaranteed consistency and replication are also irrelevant in this context: the former because access is read-only; the latter because that is the responsibility of the data owner---each site is considered autonomous.

Figure 3 depicts the operation of the federation functionality. The client first formulates a federated query, which is made up of sub-queries and details of how their results should be combined. Each sub-query

contains details of the server which hosts the plugin, the plugin identifier and the query to be executed. In step 1 the client application utilising the API sends a query to the client's local node. The local node then decomposes the overall query into its individual query elements. These individual queries are then forwarded to the relevant remote nodes (step 2), or processed locally, if appropriate. Each node then processes each individual query it has received before returning (step 3) the results along with additional information about the success of the processing to the originating local node. Once a result has been received by the local node from all the nodes processing queries, the resulting data is then combined as requested by the original query submitted by the client. In addition, all the information relating to the success (or otherwise) of each of the sub queries is aggregated. This combined data and report on the processing is then returned to the client application (step 4).

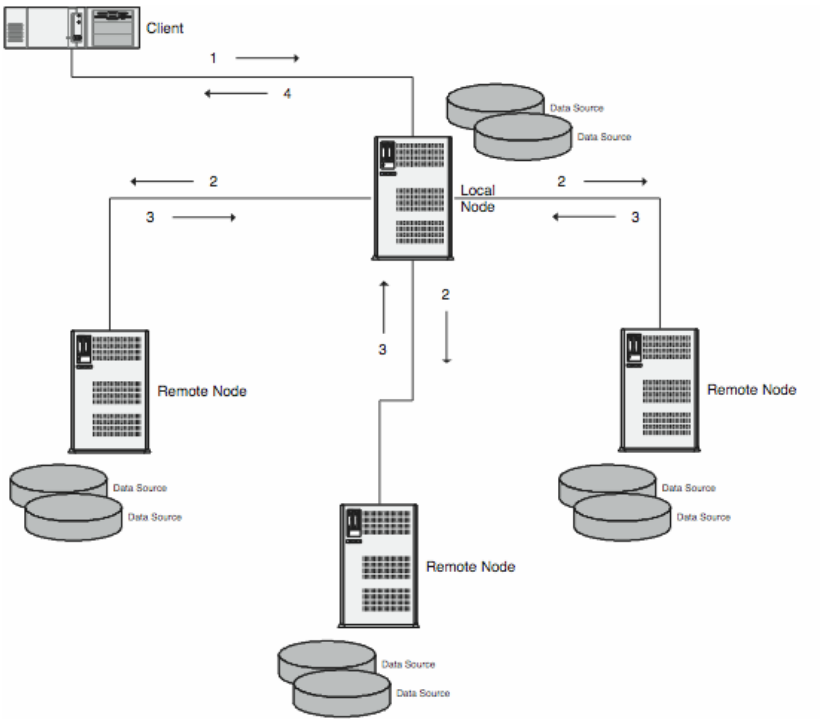

Figure 3 A federated query

3 Connectivity within the Clinical Trial

As alluded to earlier, there are, within the GIMI project, three application areas that are helping to validate the development of *sif*. In this section we describe one of these applications, which pertains to the analysis of mammographic images. In particular, the application is concerned with the validation and continued development of the algorithm described in Section 1.

The Giotto Image MD, like most digital mammography systems, incorporates DICOM 3.0 communication standards [17] allowing it to fit into any RIS/PACS environment. Our clinical partner has the unit connected to a full PACS system to which all acquired images are sent. As well as the image, the PACS system holds the associated metadata, some of which is patient sensitive, names, dates of birth, identification numbers and such.

The short-term goal driving the application of *sif* that is described in this paper is to facilitate direct access to image data from the PACS server of our clinical partner, by overlaying it with both a file (for transfer of image data) and data plugin (for querying the metadata to establish which studies to request via the file plugin). So, for example, in the event of a phantom acquisition being required, this could be acquired by a local radiographer, sent to the PACS system, and then be available to a researcher immediately—in contrast to the current arrangement of writing recordable media, such as a DVD, and sending it via the public post (which raises issues of integrity and reliability). In the longer term, we would hope to use a similar approach to gain access to anonymised patient images and associated metadata.

Of course, consideration must be given to protecting patient privacy and confidentiality: simply adding an AET (application entity title) for the researcher is not appropriate, since this only allows very coarse restrictions on the basis of IP address and TCP port. The approach to access control provided by *sif*—whereby fine-grained authorisation policies can be constructed locally by data owners in accordance with their needs (and, in some cases, obligations)—has the potential to provide assurance to these data owners that any access granted will be restricted appropriately. This, then, has the potential to offer functionality over and above that afforded by the all-or-nothing approach of the addition of an AET.

Currently, we are utilising file, data and algorithm plugins to refine the existing algorithm on legacy data sets. Two DICOM operations are of primary interest in this respect:

- **C-MOVE.** This operation is used to move image data. The C-MOVE SCP (Service Class Provider) is requested to act as a C-STORE SCU (Service Class User) and to copy composite instances to a requested AET, which may or may not be the original C-MOVE SCU (although it normally is). Interfacing to this operation has been successfully implemented as a file plugin.

- **C-FIND.** This operation is akin to an SQL query, whereby a data set is passed from the SCU to the SCP containing two sorts of attribute:

    - those which need to be matched (equivalent to the WHERE clause of an SQL query); and

    - those to be returned to the SCU (equivalent to the SELECT clause of an SQL query).

    Interfacing to this operation has been successfully implemented as a data plugin.

The algorithm plugin facility allows us to share our image processing and normalisation algorithms with our clinical partners. Figure 4 shows the output of the segmentation algorithm used as an input to the normalisation routines, which has been shared in this way. Efficient bidirectional communication is thus possible which provides an iterative development cycle in which new versions of an algorithm may be made available instantly to the clinical users, who are then able to apply it to the available images, review the output, and subsequently report back the success or limitations of the refinements. Complications of keeping many different potentially remote computers up-to-date are thus reduced. In addition, it facilitates access to an algorithm without the need to distribute the code or a binary executable (which might be reverse-engineered): this, then, has the potential to enable collaboration without risking the loss or compromise of intellectual property.

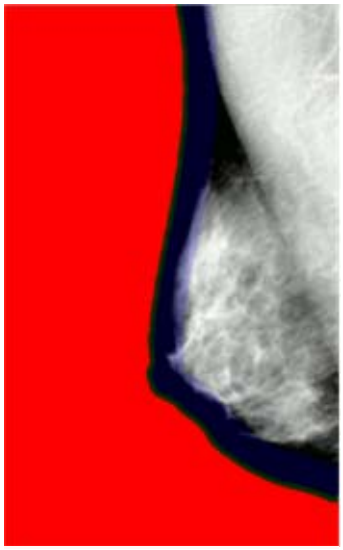

Figure 4 The output of the algorithm to segment the breast-air boundary and inner periphery at which the breast begins to occupy the full separation between the compression paddles.

4 Discussion

The vast upsurge in the popularity of digital mammography, and in digital imaging in general, has led to a need for secure and scalable computing infrastructures that can provide secure and ethical access to remote and distributed data to those who have a legitimate usage requirement. The middleware described in this paper, though it continues to be under development, shows considerable promise in fulfilling this need. The development of *sif* within the GIMI project is being undertaken in collaboration with a number of application teams, drawn from a variety of domains. This close user involvement is ensuring that the framework developed meets the needs of a diverse community. In this paper, we have concentrated on one such application, pertaining to image analysis of mammograms. Currently, this application of *sif* is helping to validate and refine a normalisation algorithm; in the longer term the framework has the potential to facilitate genuine, real-time “joined-up” healthcare.

Reference

[1] A. Simpson, D. Power, M. Slaymaker, and E. A. Politou, "GIMI: generic infrastructure for medical informatics," presented at 18th IEEE Symposium on Computer-Based Medical Systems Proceedings, 2005., 2005.

[2] A. C. Simpson, D. J. Power, D. Russell, M. A. Slaymaker, G. K. Mostefaoui, G. Wilson, and X. Ma, "The development, testing, and deployment of a web services infrastructure for distributed healthcare delivery, research, and training," in *Managing Web Services Quality: Measuring Outcomes and Effectiveness (to appear)*, 2008.

[3] C. Tromans, M. Brady, D. Van de Sompel, M. Lorenzon, M. Bazzocchi, and C. Zuiani, "Progress Toward a Quantitative Scale for Describing Radiodensity in Mammographic Images," presented at International Workshop on Digital Mammography, 2008.

[4] J. N. Wolfe, "The prominent duct pattern as an indicator of cancer risk," *Oncology*, vol. 23, pp. 149-58, 1969.

[5] J. N. Wolfe, "Risk for breast cancer development determined by mammographic parenchymal pattern," *Cancer*, vol. 37, pp. 2486-92, 1976.

[6] N. F. Boyd, B. O'Sullivan, J. E. Campbell, E. Fishell, I. Simor, G. Cooke, and T. Germanson, "Mammographic signs as risk factors for breast cancer," *Br J Cancer*, vol. 45, pp. 185-93, 1982.

[7] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, "The quantitative analysis of mammographic densities," *Phys Med Biol*, vol. 39, pp. 1629-38, 1994.

[8] R. Highnam, M. Brady, and B. Shepstone, "A representation for mammographic image processing," *Med Image Anal*, vol. 1, pp. 1-18, 1996.

[9] J. Kaufhold, J. A. Thomas, J. W. Eberhard, C. E. Galbo, and D. E. Trotter, "A calibration approach to glandular tissue composition estimation in digital mammography," *Med Phys*, vol. 29, pp. 1867-80, 2002.

[10] R. Highnam and M. Brady, *Mammographic image analysis*. Dordrecht; London: Kluwer Academic, 1999.

[11] C. Tromans, "DPhil Thesis: Measuring Breast Density from X-Ray Mammograms," in *Engineering Science*: Oxford University, October 2006.

[12] C. Tromans and M. Brady, "An Alternative Approach to Measuring Volumetric Mammographic Breast Density," presented at International Workshop on Digital Mammography, 2006.

[13] J. M. Brady, D. J. Gavaghan, A. C. Simpson, M. Mulet-Parada, and R. P. Highnam, *eDiaMoND: A Grid-enabled federated database of annotated mammograms*: Wiley, 2003.

[14] S. R. Amendolia, F. Estrella, W. Hassan, T. Hauer, D. Manset, R. McClatchey, D. Rogulin, and T. Solomonides, "MammoGrid: A Service Oriented Architecture Based Medical Grid Application," *Lecture Notes in Computer Science*, vol. 3251, pp. 939-942, 2004.

[15] M. A. Slaymaker, D. J. Power, D. Russell, G. Wilson, and A. C. Simpson, "Accessing and aggregating legacy data sources for healthcare research, delivery and training," presented at SAC, 2008.

[16] S. G. Erberich, J. C. Silverstein, A. Chervenak, R. Schuler, M. D. Nelson, and C. Kesselman, "Globus MEDICUS - Federation of DICOM Medical Imaging Devices into Healthcare Grids," presented at HealthGrid, 2007.

[17] NEMA, *Digital Imaging and Communications in Medicine (DICOM) Standard, version 3*, 2000.