PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2010) **368**, 3891–3905
doi:10.1098/rsta.2010.0153

GIMI: the past, the present and the future

By Andrew Simpson[1,*], David Power[1], Douglas Russell[1], Mark Slaymaker[1], Vernon Bailey[2], Chris Tromans[3], Michael Brady[3] and Lionel Tarassenko[2]

[1]*Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK*
[2]*Institute of Biomedical Engineering, Old Road Campus Research Building, University of Oxford, Old Road Campus, Headington, Oxford OX3 7DQ, UK*
[3]*Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK*

In keeping with the theme of this year's e-Science All Hands Meeting—past, present and future—we consider the motivation for, the current status of, and the future directions for, the technologies developed within the GIMI (Generic Infrastructure for Medical Informatics) project. This analysis provides insights into how some key problems in data federation may be addressed. GIMI was funded by the UK's Technology Strategy Board with the intention of developing a service-oriented framework to facilitate the secure sharing and aggregation of heterogeneous data from disparate sources to support a range of healthcare applications. The project, which was led by the University of Oxford, involved collaboration from the National Cancer Research Institute Informatics Initiative, Loughborough University, University College London, t+ Medical, Siemens Molecular Imaging and IBM UK.

Keywords: e-Research; security; data; healthcare

1. Introduction

As the volumes of data collected about citizens, patients and consumers increase, there are an increasing number of concerns with respect to the treatment of those data. Within the UK, these concerns have been further heightened in response to a variety of data-related incidents, such as those associated with HM Revenue and Customs, when the entire child benefit database was sent (unregistered and unencrypted) to the National Audit Office—only for the disks to fail to arrive. In the healthcare context, such issues have, of course, been discussed over a long period of time. The very nature of the data means that concerns about privacy and appropriate use are typically at the forefront of the minds of those responsible for the design, implementation and deployment of information systems. For example, Anderson (2008) summarizes the key issues pertaining to the UK's 'NHS database', while Blobel (2007) identifies the formal modelling of policies, and performing policy bridging, as the main challenges to be

*Author for correspondence (andrew.simpson@comlab.ox.ac.uk).

One contribution of 16 to a Theme Issue 'e-Science: past, present and future I'.

met in establishing trustworthy distributed e-Health solutions. Work of interest includes that of Kalra *et al.* (2005), which describes the approach to security and confidentiality adopted within the CLEF (Clinical e-Science Framework) project, and Ainsworth *et al.* (2006), which addresses security issues in the context of data collection for epidemiology.

Simpson *et al.* (2008) describe an approach to the secure sharing and aggregation of medical-related data. The approach, manifested in terms of SIF (service-oriented interoperability framework), was developed within GIMI (Generic Infrastructure for Medical Informatics) (Simpson *et al.* 2005), a multidisciplinary, collaborative project led by the University of Oxford, and which involved collaboration between the National Cancer Research Institute Informatics Initiative, Loughborough University, University College London, t+ Medical, Siemens Molecular Imaging and IBM UK. The SIF middleware forms one part of the underlying technology of GIMI; the other is an approach to *evolving* access control—a means of updating, dynamically, access control policies on the basis of system observations or changes in the environment. It is the former that is the focus of this paper. The development of these technologies was driven and validated by three key application areas, associated with healthcare delivery, research and training, and which were chosen for their diverse set of requirements. The first, pertaining to support for the self-management of long-term conditions, was motivated by the desire to use existing datasets to support decision-making: two applications, one for asthma sufferers and one for diabetes patients, were developed. The second application concerned image analysis for cancer care, with the middleware developed within GIMI providing a means of accessing—securely—hundreds of high-quality digital images held remotely to train and validate novel algorithms. The third involved the development of an intelligent training application—whereby cases could be selected on the basis of individual's 'weaknesses'—for breast radiologists.

SIF takes a *data-agnostic* approach to facilitating the sharing and aggregation of data, and its security mechanism has been designed to be sufficiently flexible that the needs of the data owner in terms of *access*—no matter how exotic—should be met. In this respect, the security drivers of SIF—certainly in terms of granularity—have been sympathetic to those of Zhang *et al.* (2007). SIF acts primarily as a secure gateway and data integration framework; issues such as semantic interoperability and support for workflow are not of concern.

Power *et al.* (2005) describe a number of security use cases that, subsequently, were used to underpin the design of SIF, with those use cases being influenced directly by experiences gained from the e-DiaMoND (Brady *et al.* 2003) and NeuroGrid (Geddes *et al.* 2005) projects. In this paper, we reflect upon and categorize the *actual* ways in which SIF has been used to support the secure sharing and aggregation of data. Interestingly, none of these modes of use correspond to the classic 'data grid' pattern that originally drove much of our work. We conclude that, perhaps, the 'federated' virtual organization pattern represented by our previous contributions was a simplification—both of desired patterns of usage and, more generally, of patterns of actual collaboration. The paper has two overarching themes: a discussion of the development of SIF, and the description of four patterns of data aggregation that have come to light as a result of our research.

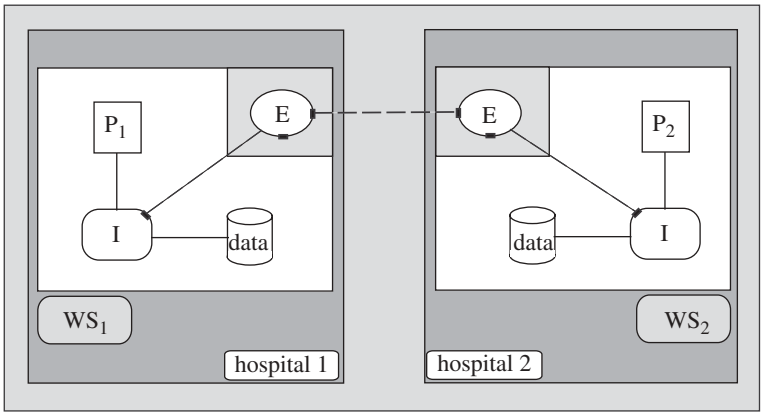

Figure 1. The SIF view of a virtual organization.

2. Motivation and background

The origins of GIMI can be traced to the e-DiaMoND project. There, a prototype system for the sharing of digitized mammograms and related patient data to support a variety of applications was developed using a combination of IBM enterprise solutions, emerging Grid and e-Science technologies—in the form of Globus TOOLKIT 3 and OGSA-DAI (Open Grid Services Architecture Data Access and Integration)—and bespoke code. The lessons learnt from e-DiaMoND—in particular, those associated with security, interoperability and abstraction—were subsequently considered in Power *et al.* (2005). (The high-level architecture of Power *et al.* (2005), whereby external services (E) mediate access according to local policies (P_1 and P_2) is illustrated in figure 1.) While early versions of what was to become SIF underpinned the MRC-funded NeuroGrid project (Geddes *et al.* 2005) and a prototype demonstrator for the National Cancer Research Institute Informatics Initiative (Pitt-Francis *et al.* 2006), it is through GIMI that the fundamental ideas described of Power *et al.* (2005) have been realized.

The fundamental motivation behind GIMI is the development of a means of sharing, federating and using data securely—and in a lightweight, portable, and generic fashion. Going further, we wish to develop a means of supporting 'big ideas'—bigger and better research, personalized healthcare, and joined up e-Government—but in a way that does not require organizations to throw away existing systems, change practices or invest heavily in new technology. GIMI's drivers can, therefore, be characterized in terms of: (i) interoperability, heterogeneity and portability—any kind of data stored on any kind of database or file system should be capable of being accessed and shared via a standard interface; (ii) secure data sharing—data access and transfer should be in accordance with the data owners' wishes, no matter how prescriptive; (iii) low costs of entry—in terms of installation and deployment, system footprint and effort required on behalf of application developers; and (iv) abstraction—via a simple applications programming interface (API), developers can construct applications to aggregate and use data without concerning themselves about issues such as secure data transport.

3. The current system

In this section, we provide a brief overview of our middleware framework, SIF. More detailed descriptions are available elsewhere: for example, SIF's support for federation is described in Slaymaker *et al.* (2008*a*); support for fine-grained access control is described in Slaymaker *et al.* (2008*b*); and the 'plug-in' mechanism—which gives rise to SIF's data agnosticism—is described by Russell *et al.* (2009).

SIF is a Web services framework that is based upon freely available (and, where possible, open-source) technologies that can run on multiple platforms, with bespoke code written in Java. We reiterate two of the drivers of SIF—interoperability and security—as described by Simpson *et al.* (2008) below.

— *Interoperability.* Taking a simplified view (and leaving aside higher-level concerns such as semantics), one might characterize 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 used by Dr Smith and the Oracle database used by Dr Thomas). Our concern is the latter; issues of semantic interoperability are left to application developers.

— *Security.* In an e-Health 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, of concern. With respect to the former, access policies are constructed by data owners (and represented in terms of XACML (eXtensible Access Control Markup Language)—see www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml). With respect to the latter, secure channels are established between external nodes. As discussed in Power *et al.* (2005), the requirements for access and transfer have been influenced by UK-centric concerns, in the form of: National Health Service (NHS) guidelines (such as, for example, the principles of the Caldicott Guardian—see www.addenbrookes.org.uk/advice/medethlaw/confidential1.html); UK legislation (such as, for example, the Data Protection Act—see http://www.hmso.gov.uk/acts/acts1998/19980029.htm); and wider European legislation (such as, for example, the European Convention on Human Rights—see http://www.opsi.gov.uk/acts/acts1998/ukpga_19980042_en_1).

The users of SIF are (at one interface) data owners who wish to share their data, and (at the other interface) application developers who wish to make use of those data in some way. A loosely coupled approach sympathetic to our aforementioned data and technology-agnostic philosophy is taken. Importantly, the fundamental assumption is that data owners determine access control policies: our assumption is that there is a decentralized security model—although support for a centralized model is possible, should it be necessary. Sinnott *et al.* (2008) discuss the pros and cons of centralized and decentralized security models.

Communities of collaborators are able to come together in a 'bottom-up' fashion: for example, application developers only need to worry about interoperability between relevant data sources—rather than worrying about interoperability across the whole virtual organization. Suppose, say, that data source S1 might contain data and files pertaining to both breast and colorectal cancer, data source S2 might contain data and files pertaining to breast cancer, and data source S3 might contain data and files pertaining to colorectal cancer. S1 and S2 might form one virtual organization that is concerned with breast cancer; S1 and S3 might form a second virtual organization that is concerned with colorectal cancer. Each virtual organization, then, would be concerned with facilitating semantic interoperability to share relevant data: breast cancer in the case of the first virtual organization, and colorectal cancer in the case of the second virtual organization. If, at a later date, the two virtual organizations were to merge to form a single community of interest, then, at that point, issues of interoperability between the breast and colorectal cancer datasets would have to be considered.

Via the SIF API, an application developer can determine how much of the underlying data should be exposed to end-users: depending on the context, a full SQL (Structured Query Language) interface may be appropriate; alternatively pre-formulated queries that abstract unnecessary details may be 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. The reason that SIF can expose any relational database 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.

SIF offers support for three types of 'plug-in': data plug-ins, file plug-ins and algorithm plug-ins. By using a standard plug-in interface, it becomes possible to add heterogeneous resources into a virtual organization. Importantly, there is no need for the resource being advertised through the plug-in 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.

Currently, data plug-ins treat all data sources as relational databases, with the plug-in being responsible for all translations between the native data format and SQL. The plug-in user or application developer can retrieve schemas for known resources; as such, the user of the plug-in is able to perform a join or a union query on data from fundamentally different data sources. The interface is modelled on SQL, taking a standard SQL query string as input and returning Java WebRowSets. Each data plug-in exposes a defined view of some subset of the data that is available, with these views being advertised upon request via the API.

Algorithm plug-ins are designed to execute arbitrary algorithms written in Java, or, alternatively, incorporate other, existing algorithms written in different languages via a Java plug-in—with such algorithms being run on a remote machine if necessary. The definition of an algorithm plug-in includes required input and output files, as well as an XML (eXtensible Markup Language) schema describing the input and output arguments. When executed, the plug-in manager

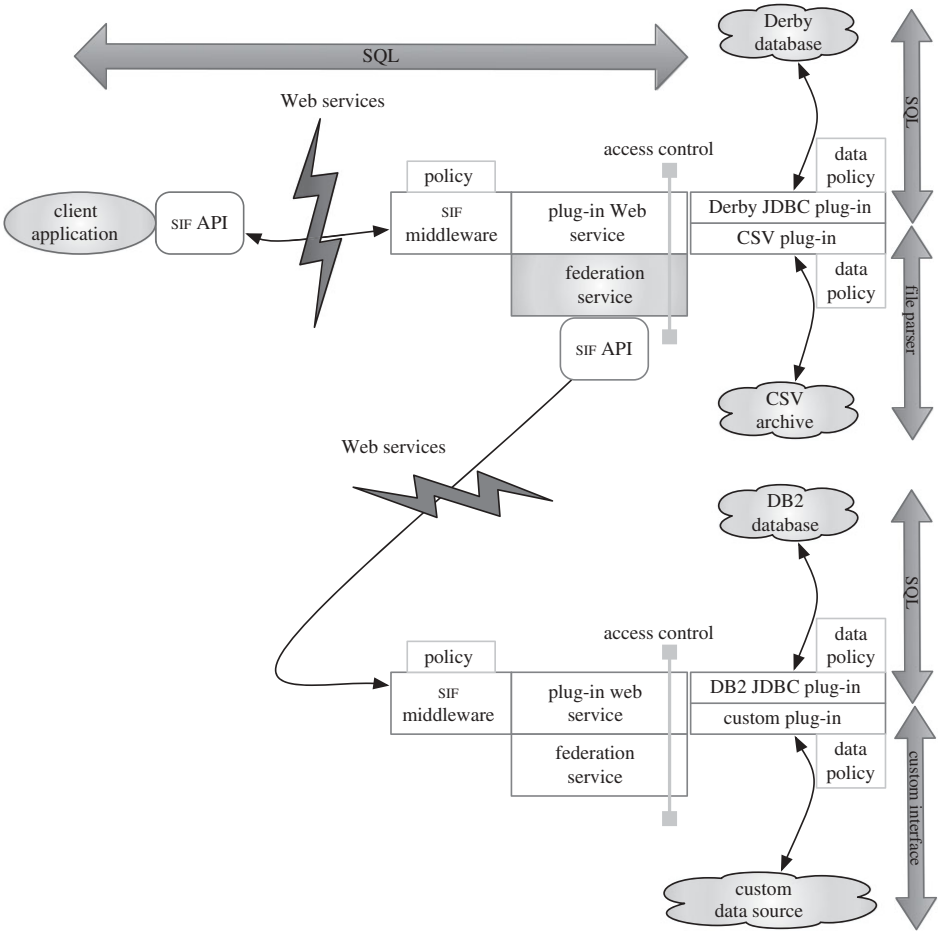

Figure 2. SIF: architecture.

will ensure that the inputs and outputs to the plug-in conform not only to the general parameters of the plug-in type in general, but also to the requirements of the plug-in being executed.

File plug-ins expose real or virtual file systems through a very tightly defined interface. Currently, the file plug-in interface functions similarly to an FTP interface in that the interface supports *put*, *get* and *list* functionality. A pair of plug-ins has been written for PACS (Picture Archiving and Communications System) systems—the predominant way of storing and retrieving medical images and related data—which typically hold DICOM (Digital Imaging and Communications in Medicine) files. The DICOM format stores the images and the associated data together in the same file; a PACS server usually allows querying of some of the fields from the DICOM header to select the files of interest. The query is a job for a data plug-in; the file retrieval is a job for a file plug-in.

The architecture of a SIF deployment is given in figure 2. SIF can be thought of as being composed of three parts: the core middleware, the plug-ins and the client-side API. The core middleware manages the installed plug-ins, giving them

a standard interface to be written against. It also provides a federation service to facilitate the construction of queries against multiple data sources, not necessarily in the same SIF instantiation or machine. As all data are represented in SIF as if they were a standard SQL database, these queries take the form of SQL queries across distinct data sources each exposed via a separate plug-in. The access control framework enforces policies created by the owners of the data and the owners of the machine on which SIF is being hosted, allowing data owners to restrict the data they expose to users and server owners to control who the permitted users of services are.

The middleware has capabilities for transferring files and data: installing, removing and updating plug-ins; advertising and defining resources exposed by plug-ins; and providing system status information. The core middleware exposes this functionality through a number of Web services, all of which use strong cryptography to ensure privacy. The client-side API is a wrapper around Web service calls to create the simplest possible interface for a new application developer to implement against; it also provides a number of helper functions to assist in common tasks.

4. Patterns of use

In this section we consider the four broad categories of use that are currently representative of the applications supported by SIF. In each case, SIF's support for flexible and expressive access controls has been necessary to engender the appropriate level of trust—both between partners in the relevant virtual organization and between categories of developers (application, middleware and plug-in).

(*a*) *'Secure pipelines'*

The simplest use case involves no federation whatsoever. In many contexts, researchers wish to gain access to data stored on a remote device: rather than the time-honoured CD or DVD as a mechanism for transmission, one might wish to access images directly from a PACS machine. In the context of clinical data, of course, such access and transfer needs to be undertaken with significant guarantees with respect to security.

We might consider there being two sides to the 'trust equation': the first is concerned with limiting who can access data and under what circumstances; the second is concerned with limiting what authorized users can do with data.

With respect to the former, SIF's approach to access control—giving rise to expressibility via XACML and giving rise to dynamism via 'evolving access control'—means that data owners can prescribe very fine-grained policies, and also, perhaps, policies that take into account conditions on the system (or, perhaps, users' previous actions). This gives rise to the potential for policies such as 'Dr X can access up to 20 GB worth of images per day' or 'Researcher Y can, for a given patient, access either field A, or field B—but not both'.

While there is a need to support requirements such as 'you may view, but you may not copy', this is something that cannot be supported (at least currently) by SIF. The experiences of those concerned with developing trusted infrastructures, such as, for example, Cooper & Martin (2006) and Huh & Martin (2009), are being followed closely in this respect.

(*b*) *Lightweight aggregation*

The most common use case pertains to the aggregation of data from a variety of data sources. Earlier experiences from projects such as the aforementioned e-DiaMoND and NeuroGrid gave rise to the 'bottom-up' philosophy of §2: issues of syntactic and semantic interoperability are left to domain experts—in the form of plug-in writers and application developers—to resolve. In this context, SIF provides a means of abstraction: issues such as secure access and transport and data aggregation are abstracted from the application developer as much as reasonably possible. One key difference between emerging applications and those envisaged in earlier projects is that there is now a need for more lightweight means of collaboration—whereby legacy systems remain unchanged as much as possible.

(*c*) *'Windows' on research data*

There is, in effect, a three-tiered limitation on access to data sources via SIF: the plug-in writer exposes that part of the data source that may be accessible (there may be several plug-ins per data source); SIF's access control mechanism, positioned outside of the data source, restricts access according to credentials, as well as other properties (there will, typically, be many users per application); and, finally, the application developer may restrict access further (there may be several applications associated with each plug-in). Quite separately from the middleware, the data source may very well have its own restrictions on behaviour.

Projects such as Integrating Biology Virtual Research Environment (IBVRE; Lloyd *et al.* 2007) have demonstrated that there are drivers for aggregating research data from disparate sources, which are then accessed via a single, central portal—with that portal controlling access to the variety of data sources. Typically, teams that wish to share data in this way will want to do it in a restricted fashion—they will only want to share a particular subset of their data. Equally, those responsible for developing the front-end solution will want to provide appropriate assurance to the data owners that only appropriately authorized users can access the supplied data.

(*d*) *Integrating central systems with 'outliers'*

The 'classic' data grid use cases are based on the assumption of the bringing together of disparate and disjoint (and possibly heterogeneous) sources into a single logical whole. Increasingly, however, many organizations are moving to a hybrid scenario whereby a single, centralized data source captures much essential data pertaining to the organization, while departments maintain their own local systems. There is, though, still a need to link the single, central system with the 'outliers': either to ensure that the centre gains a 'big picture' view of the organization or to enable those on the periphery to ensure that they can use the central data effectively for their specific needs. Of course, one might envisage this

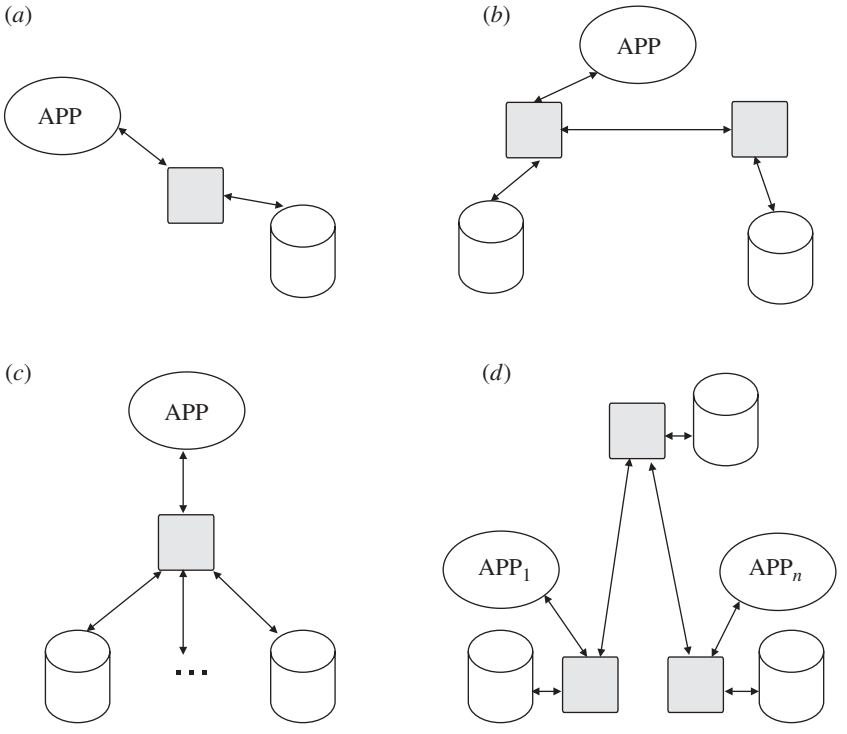

Figure 3. Use cases: (*a*) simple pipeline; (*b*) lightweight federation; (*c*) windows; and (*d*) outlier integration.

scenario as a more generalized form of the 'secure pipeline' use case; one might also consider the 'single, central' data source to, in actuality, be the aggregated data source of §4*c*.

(*e*) *Summary*

These use cases are sketched in figure 3: (*a*) illustrates the 'simple pipeline' case; (*b*) illustrates the lightweight federation case; (*c*) illustrates the 'windows' case; and (*d*) illustrates the outlier integration case. In each case, the shaded box indicates the presence of a SIF node, while 'APP' represents an application.

Evidently, one can map these patterns on to the wider enterprise computing context. For example, many of the challenges facing those in e-Governance will be familiar to those with a background in e-Science or e-Research: social issues surrounding community building; semantic interoperability issues; technological challenges; etc. Similarly, there are clear parallels between these e-Health patterns and secure data sharing requirements with e-Governance.

Within the UK government context, for example, a series of data security issues, as alluded to in §1, has increased awareness of the issues surrounding the appropriate treatment of personal data. Pattern (*a*) in figure 1 does away with the need for the manual transfer of CDs and DVDs in an e-Health context and could also do so within an e-Government context, reducing the risk of 'lost disks in the post'; pattern (*b*) links two legacy systems to enable a temporary

virtual organization (through the construction of two or more plug-ins and an application) to perform a particular task, much as two government agencies might come together temporarily; etc. Of course, this is a 10 000 feet view that does not take into account many technical and social complexities; it is, though, illustrative of how experiences from the e-Research arena may start to be leveraged within other contexts.

5. Applications

Currently, SIF supports eight healthcare-related applications: the four applications of §1; the aforementioned NeuroGrid and National Cancer Research Institute prototype demonstrator; a report generator for the OPTIMA (Oxford Project to Investigate Memory and Ageing) project (see http://www.medsci.ox.ac.uk/optima); and a portal for breast cancer research, developed with colleagues at Swansea University. Having considered the four patterns of use in the abstract, we now consider the manifestation of the first three of these classes in terms of current applications of SIF; we consider the fourth in §6.

(a) *'Secure pipelines': from PACS to desktop*

The GIMI project had, at its core, three application teams that drove and validated the development of SIF: one team was concerned with healthcare research, one with healthcare training, and one with healthcare delivery. The healthcare research application pertained to support for image analysis for cancer care.

Highnam *et al.* (1995) developed the h_{int} representation and an algorithm to generate it for quantitative analysis of mammograms (X-rayed images of breasts). Tromans & Brady (2006) developed the 'next generation' of that model, in which (owing to increased computing power, among other factors) many of the assumptions of the original model have been removed. Of course, to train and validate such an algorithm, vast quantities of high-quality digital mammograms are required. Rather than transfer such data via traditional means (CD or DVD), SIF has been used to transfer data from servers based in remote hospitals to the researcher's desktop in Oxford. The file plug-in mechanism is used to access files, and the algorithm plug-in mechanism is used as a means of executing existing algorithms. The use of SIF in support of this application is described by Tromans *et al.* (2008).

(b) *Lightweight aggregation: supporting the self-management of long-term conditions*

The focus of the healthcare delivery team within GIMI was the development of applications to help support the self-management of long-term conditions. The prototype asthma application involved the linking of data held at t+ Medical's servers with data from the UK's Met Office (the UK's national weather service) to help determine and predict the effect of changes in weather conditions on asthma patients. The prototype diabetes application was concerned with the transfer of patient data between mobile phone devices, medical practitioners' databases and servers to facilitate the real-time monitoring and prompting,

where appropriate, of diabetes patients. In both cases, the driver was for a lightweight, secure means of integrating data from disparate sources to enable the novel use of data.

(*c*) *'Windows' on research data: bringing life to legacy data*

The OPTIMA project aims to improve the understanding of the changes that occur as the brain ages, via a longitudinal study involving both patients with memory problems and some control subjects. Data have been collected over a period of more than 20 years. Previously, researchers' access to the existing database was available only through the data manager: researchers and investigators would construct their questions of interest; a query would be formulated by the data manager; and a spreadsheet would be returned to the researcher or investigator. A data plug-in was created, and an application has been developed that allows the construction of (simple) queries by non-expert users. The application interacts with the data plug-in to retrieve data based on a full range of criteria as applicable to the data, visualizing those data for the user and allowing the option to save them out as a CSV file for processing.

Access is tightly controlled on a user-by-user basis, and all interaction with the database is fully audited. Further, it should now be possible to share this resource (in an appropriate fashion) with the wider community. An immediate consequence of this work is that discussions are now under way to use the federation aspects of SIF to allow other research groups working on brain-related topics within Oxford to share their resources. By starting from the OPTIMA data, and growing outwards, this will be a proving ground for the bottom-up philosophy, with a view to supporting the organic growth of communities.

6. Moving forward

As with any technology, innovations are being driven by the needs of users. Moving forward, further applications are being developed via funding from the Oxford Biomedical Research Centre (BRC) (see www.oxfordbrc.org), with an expansion of the aforementioned OPTIMA report generator to incorporate other research groups' data being an initial focus. Initial steps are being undertaken to measure SIF's viability to support applications in other domains; funding from the Oxford Centre for Integrative Systems Biology (see www.sysbio.ox.ac.uk) to underpin a variety of data sharing efforts has been secured in this respect. Importantly, however, despite the fact that SIF was developed initially to meet the needs of those working with healthcare-related data, the resulting system makes no assumptions with respect to the type of data being accessed and shared, nor, indeed, its use. As such, it has the potential to be applied within many other contexts—such as the manifestation of our fourth pattern.

(*a*) *Integrating central systems with 'outliers': the Oxford experience*

The University of Oxford has an intrinisically distributed structure: while 'the university' certainly exists and comprises four divisions, which, in turn, comprise numerous departments, there are also over 30 colleges—each (undergraduate and postgraduate) student being a member of a college. Over the past five years or

so, the university has developed a series of centralized IT systems (for financial management, student management, personnel management, etc.), yet, at the same time, each department and each college maintains its own systems. The most recent application has involved developing a plug-in for one of these systems (the student management system), as well as for several external systems, to enable the linking of central data with outlying data to enable the integration of central and local views of the student lifecycle. The construction of a bespoke application and associated access control policies took a matter of weeks—allowing administrators to link student data from disparate systems in a straightforward fashion.

7. Discussion and conclusions

We have summarized the development of SIF: a Java and Web services framework for the secure aggregation of heterogeneous data. We have also discussed the four broad classes of use case that have arisen in our work. We do not claim that this survey is exhaustive—we have yet to encounter a 'cloud'-like scenario, for example—but these are the patterns of use that have cropped up time and again. Interestingly, the classic 'data grid' scenario as evidenced by, for example, e-DiaMoND and NeuroGrid, which were fundamental to the origins of SIF, is notable by its absence. This may be for a variety of reasons. First, it may be that, where such collaborations do exist, the researchers have their own established methods of doing things, or, alternatively, the benefits that may be afforded by a lightweight solution pale into insignificance when compared with the non-technical issues associated with establishing trust, ensuring semantic interoperability, etc.—meaning that more 'heavyweight' solutions are a sensible choice. An alternative reason may be that such pilot projects were not representative of true needs: they did not actually represent the kinds of collaboration that researchers wished to undertake, or, perhaps, were a simplification of the four patterns that we have experienced.

The experiences of De Roure & Goble (2009) will echo with many who have been responsible for delivering solutions for e-Scientists. There (among other valuable conclusions), the following six 'principles of user engagement' are given: 'keep your friends close'; 'embed'; 'keep sight of the bigger picture'; 'favours will be in your favour'; 'know your users'; and 'expect and anticipate change'. Fundamentally, even though technologists are developing technological solutions, it is essential that there is appropriate 'pull' from application scientists: to abuse a well-known phrase, if you build it they may not necessarily come. Any technological solution *has* to be driven and validated by its end-users.

We have met a variety of challenges along the way: some technical, some social. Perhaps the three most pertinent lessons that we have learnt are the following.

— *One man's database is another man's spreadsheet.* When computer scientists and software engineers use the term 'database', there is a generally understood meaning of the term. Often this is the same in other disciplines; sometimes, however, it is not. Clarifying how and where data are stored is now, typically, the first item on the agenda for any new engagement that we have.

— *Trust is predominantly a social concept.* Why should anyone trust us to ensure that our middleware will protect their data? More importantly, even if a collaborator is convinced, how might we convince their data manager or systems administrator? The technical matter of opening up ports is trivial; the social matter of engendering trust certainly is not.
— *Being on the 'bleeding edge' can be ... interesting.* SIF is based on Java and Web services. While, in the abstract, working with open standards is a good thing, in practice, things are not always so rosy: for example, attempting to ensure that an interoperability framework works with incompatible closed-source commercial extensions to open standards is very painful. As a further example, SIF was developed at a time when Web service standards and their implementations were evolving rapidly, and, as such, problems of interoperability were an all-too-common occurrence. While Java 6 now incorporates much of the core Web service functionality required, SIF is still dependent on libraries that were created during the period of flux and were never updated to be compatible.

There are comparisons to be made with other technologies emerging from the e-Science domain. OGSA-DAI (Antonioletti *et al.* 2005), for example, 'aims to provide the e-Science community with a middleware solution to provide access to and integration of data for applications working across administrative domains' (Antonioletti *et al.* 2007). As a data integration framework, OGSA-DAI has enormous benefits; SIF's focus on security, however, offers a degree of flexibility and expressibility over that afforded by OGSA-DAI. PERMIS (PrivilEge and Role Management Infrastructure Standards; Chadwick *et al.* 2008) provides a role-based authorization infrastructure, which goes further than our security concerns, in that it is concerned not only with authorization, but also with supporting distributed credentials management. The work of the VOTES (Virtual Organisations for Trials and Epidemiological studies) consortium (Sinnott *et al.* 2007) is also of relevance: there, security and usability are highlighted as key concerns in ensuring the acceptance of service-oriented technologies in providing seamless access to aggregated data sources. Whereas the VOTES consortium uses OGSA-DAI in its delivery of services, within SIF, we have attempted to combine security and aggregation in a single framework.

There are three areas of work that we are addressing in the short term. First, as we address new domains, new requirements inevitably emerge. For example, we are in the process of developing a variety of applications for the Oxford Centre for Integrative Systems Biology (OCISB). The focus here is not on security—that is of little concern—but on data transfer: the amount of data to be transferred leaps from tens of megabytes to several gigabytes as we move from the healthcare context to the biological sciences. Second, there are several rewrites that are about to be undertaken, the most notable of which are a port to Java 1.6 and a switch to a modular approach to authorization, which will permit the use of mechanisms that are more accessible or familiar to policy writers—but perhaps less expressive—than XACML. Ultimately, we wish SIF to become as independent of authorization mechanisms as it is of back-end technologies and data models. Finally, we are taking initial steps in terms of moving SIF from a closed-source product to an open-source one.

The work described in this paper was funded by the Technology Strategy Board's Collaborative Research and Development Programme. Ghita Kouadri Mostefaoui, Xiaoqi Ma and Graeme Wilson contributed to the development of SIF. We would like to acknowledge the efforts of our collaborators within GIMI and the BRC: Grzegorz Agacinski, Simon Berry, Alastair Gale, Alan Hogg, Paul Lewis, Oliver Lyttleton, David Smith, Paul Taylor, Igor Toujilov, Gordon Wilcock and Moi Hoon Yap.

References

Ainsworth, J., Harper, R., Juma, I. & Buchan, I. 2006 Design and implementation of security in a data collection system for epidemiology. *Stud. Health Technol. Informatics* **120**, 348–357.

Anderson, R. 2008 Patient confidentiality and central databases. *Br. J. General Practice* **58**, 75–76. (doi:10.3399/bjgp08X263992)

Antonioletti, M., *et al.* 2005 The design and implementation of Grid database services in OGSA-DAI. *Concurr. Comput.: Pract. Exp.* **17**, 357–376. (doi:10.1002/cpe.939)

Antonioletti, M., *et al.* 2007 OGSA-DAI 3.0—the whats and whys. In *Proc. UK e-Science All Hands Meeting 2007, Nottingham, UK, 10–13 September*, pp. 158–165. See http://www.allhands.org.uk/2007/proceedings/.

Blobel, B. 2007 Comparing approaches for advanced e-health security infrastructures. *Int. J. Med. Informatics*, **76**, 454–459. (doi:10.1016/j.ijmedinf.2006.09.012)

Brady, J. M., Gavaghan, D. J., Simpson, A. C., Mulet-Parada, M. & Highnam, R. P. 2003 eDiaMoND: a grid-enabled federated database of annotated mammograms. In *Grid computing: making the global infrastructure a reality* (eds F. Berman, G. C. Fox & A. J. G. Hey), pp. 923–943. New York, NY: Wiley.

Chadwick, D. W., Zhao, G., Otenko, S., Laborde, R., Su, L. & Nguyen, T. A. 2008 PERMIS: a modular authorization infrastructure. *Concurr. Comput.: Pract. Exp.* **20**, 1341–1357. (doi:10.1002/cpe.1313)

Cooper, A. & Martin, A. P. 2006 Towards a secure, tamper-proof grid platform. In *Proc. Sixth IEEE Int. Symp. on Cluster Computing and the Grid* (*CCGrid 06*), *Singapore, 16–19 May*, pp. 373–380. Los Alamitos, CA: IEEE Computer Society. (doi:10.1109/CCGRID.2006.103)

De Roure, D. & Goble, C. 2009 Software design for empowering scientists. *IEEE Softw.* **26**, 88–95. (doi:10.1109/MS.2009.22)

Geddes, J. *et al.* 2005 NeuroGrid: using grid technology to advance neuroscience. In *Proc. 18th IEEE Symp. on Computer-Based Medical Systems* (*CBMS 2005*), *Dublin, Ireland, 23–24 June*, pp. 570–573. Los Alamitos, CA: IEEE Computer Society. (doi:10.1109/CBMS.2005.76)

Highnam, R. P., Brady, J. M. & Shepstone, B. 1995 A representation for mammographic image processing. In *Proc. First Int. Conf. on Computer Vision, Virtual Reality and Robotics in Medicine*. Lecture Notes in Computer Science, no. 905, pp. 365–371. Berlin, Germany: Springer. (doi:10.1007/BFb0034972)

Huh, J. H. & Martin, A. P. 2009 Towards a trustable virtual organisation. In *Proc. 2009 IEEE Int. Symp. on Parallel and Distributed Processing with Applications* (*ISPA 2009*), *Chengdu, China, 9–12 August*, pp. 425–431. Los Alamitos, CA: IEEE Computer Society. (doi:10.1109/ISPA.2009.72)

Kalra, D., Singleton, P., Milan, J., MacKay, J., Detmer, D., Rector, R. & Ingram, D. 2005 Security and confidentiality approach for the Clinical e-Science Framework (CLEF). *Methods Information Med.* **44**, 193–197.

Lloyd, S. *et al.* 2007 Integrative biology: the challenges of developing a collaborative research environment for heart and cancer modelling. *Future Gener. Comput. Syst.* **23**, 457–465. (doi:10.1016/j.future.2006.07.002)

Pitt-Francis, J., Chen, D., Slaymaker, M. A., Simpson, A. C., Brady, J. M., van Leeuwen, I., Reddington, F., Quirke, P. & Gavaghan, D. J. 2006 Multimodal imaging techniques for the extraction of detailed geometrical and physiological information for use in multi-scale models of colorectal cancer and treatment of individual patients. *Comput. Math. Methods Med.* **7**, 177–188.

Power, D. J., Politou, E. A., Slaymaker, M. A. & Simpson, A. C. 2005 Towards secure Grid-enabled healthcare. *Softw.: Pract. Exp.* **35**, 857–871. (doi:10.1002/spe.692)

Russell, D., Power, D. J., Slaymaker, M. A., Kouadri-Mostefaoui, G., Ma, X. & Simpson, A. C. 2009 On the secure sharing of legacy data. In *Proc. Sixth Int. Conf. on Information Technology: New Generations* (*ITNG 2009*), *Las Vegas, NV, 27–29 April*, pp. 1676–1679. Los Alamitos, CA: IEEE Computer Society. (doi:10.1109/ITNG.2009.21)

Simpson, A. C., Power, D. J., Slaymaker, M. A. & Politou, E. A. 2005 GIMI: Generic Infrastructure for Medical Informatics. In *Proc. 18th IEEE Symp. on Computer-Based Medical Systems* (*CBMS 2005*), *Dublin, Ireland, 23–24 June*, pp. 564–566. Los Alamitos, CA: IEEE Computer Society. (doi:10.1109/CBMS.2005.59)

Simpson, A. C., Power, D. J., Russell, D., Slaymaker, M. A., Kouadri-Mostefaoui, G., Ma, X. & Wilson, G. 2008 A healthcare-driven framework for facilitating the secure sharing of data across organisational boundaries. *Stud. Health Technol. Informatics* **138**, 3–12.

Sinnott, R. O., Ajayi, O., Jiang, J., Stell, A. & Watt, J. 2007 User-oriented security supporting interdisciplinary life science research across the Grid. *New Gener. Comput.* **25**, 339–354. (doi:10.1007/s00354-007-0022-8)

Sinnott, R. O., Chadwick, D. W., Doherty, T., Martin, D., Stell, A., Stewart, G., Su, L. & Watt, J. 2008 Advanced security for virtual organizations: the pros and cons of centralized vs decentralized security models. In *Proc. 8th IEEE Int. Symp. on Cluster Computing and the Grid* (*CCGrid 08*), *Lyon, France, 19–22 May*, pp. 106–113. Los Alamitos, CA: IEEE Computer Society. (doi:10.1109/CCGRID.2008.67)

Slaymaker, M. A., Power, D. J., Russell, D., Wilson, G. & Simpson, A. C., 2008*a* Accessing and aggregating legacy data sources for healthcare research, delivery and training. In *Proc. 23rd Annu. ACM Symp. on Applied Computing* (*SAC 2008*), *Fortaleza, Ceará, Brazil, 16–20 March*, pp. 1317–1324. New York, NY: ACM Press. (doi:10.1145/1363686.1363994)

Slaymaker, M. A., Power, D. J. & Simpson, A. C. 2008*b* On the facilitation of fine-grained access to distributed healthcare data. In *Proc. Secure Data Management 2008*. Lecture Notes in Computer Science, no. 5159, pp. 169–184. Berlin, Germany: Springer.

Tromans, C. E. & Brady, J. M. 2006 A scatter model for use in measuring volumetric mammographic breast density. In *Digital Mammography, Proc. 8th Int. Workshop* (*IWDM 2006*), *Manchester, 18–21 June*. Lecture Notes in Computer Science, no. 4046, pp. 251–258. Berlin, Germany: Springer. (doi:10.1007/11783237_35)

Tromans, C., Brady, J. M., Power, D. J., Slaymaker, M. A., Russell, D. & Simpson, A. C. 2008 The application of a service-oriented infrastructure to support medical research in mammography. In *Medical Imaging on Grids: Achievements and Perspectives* (*MICCAI-Grid Workshop*), *New York, NY, 6 September*, pp. 43–52. See http://www.i3s.unice.fr/~johan/MICCAI-Grid08/pdf/tromansMICCAIG.pdf.

Zhang, N., Yao, L., Nenadic, A., Chin, J., Goble, C., Rector, A., Chadwick, D. W., Otenko, S. & Shi, Q. 2007 Achieving fine-grained access control in virtual organisations. *Concurr. Comput.: Pract. Exp.* **19**, 1333–1352. (doi:10.1002/cpe.1099)