COMPLIANCE AND CREATIVITY IN GRID COMPUTING

Anthony E. Solomonides
CEMS Faculty, University of the West of England, Bristol, BS16 1QY, UK
tony.solomonides@uwe.ac.uk

Abstract

Grid computing ("the grid") is a promising new technology to enhance the services already offered by the internet. This new paradigm offers rapid computation, large scale data storage and flexible collaboration by harnessing together the power of a large number of commodity computers or clusters of other basic machines. The grid was devised for use in scientific fields, such as particle physics and bioinformatics, in which large volumes of data, or very rapid processing, or both, are necessary. Unsurprisingly, the grid has also been used in a number of ambitious medical and healthcare applications. While these initial exemplars have been restricted to the research domain, there is a great deal of interest in real world applications. However, there is some tension between the spirit of the grid paradigm and the requirements of medical or healthcare applications. The grid maximises its flexibility and minimises its overheads by requesting computations to be carried out at the most appropriate node in the network; it stores data at the most convenient node according to performance criteria. On the other hand, a hospital or other healthcare institution is required to maintain control of its confidential patient data and to remain accountable for its use at all times. Despite this apparent conflict in requirements, we suggest that certain characteristics of the grid provide the means to resolve the problem: in the spirit of this paradigm in which "virtual organisations" arise ad hoc, "grid services" may negotiate ethical, legal and regulatory compliance according to agreed policy.

Introduction: the computing context

I will introduce some of the issues that concern us through examples from several recent projects in the field of ‘healthgrid’. I will first motivate the concept of grid computing. ‘Distributed computer systems’ predate even the internet and the World Wide Web (‘the web’). By means of a network of interconnections, computers are able to share a workload that would ordinarily be beyond the capacity of any one of them; they may also distribute data to different locations according to need or frequency of use. On the other hand, since the explosion of the web in every conceivable statistic – users, nodes, volume of information – we are familiar with its ability to serve information and misinformation in equal measure. The grid combines the technical features of distributed systems and the web, but efforts are also being made to ensure that it is not beset by the same problems of abuse, misuse and contamination as the web has been.

The ideal grid, envisaged as a servant of a new paradigm of scientific research called ‘e-science’, would provide transparent processing power, storage capacity and communication channels for scientists who may from time to time join the grid, do some work and then leave, so that the alliances they form in their scientific endeavours might be described as ‘virtual organizations’ or VOs for short. Different sciences have different needs, and the grid concept has become differentiated: particle physics generates enormous amounts of data which must be kept, but not necessarily instantly processed; on the other hand, data in bioinformatics is not large by comparison – it is, of course, in plain terms, large – but requires intensive processing. In extending the application of grid computing to e-health, another feature becomes pre-eminently necessary: that of collaboration.

An important consequence of the fluidity of collaboration in grid computing has been in the choice of ‘architecture’ for grid systems. ‘Architecture’ is used loosely in computer systems to describe the manner in which hardware and software have been assembled together to achieve a desired goal. Favoured also in the commercial application of the web, the so-called ‘Service-Oriented Architecture’ has been widely adopted in grid applications. In effect, it means that needed services – software applications – once constructed, are provided with a description in an agreed language and made available to be ‘discovered’ by other services that need them. A ‘service economy’ is thus created in which both *ad hoc* and systematic collaborations can take place.

Compared with data from physics or astronomy, medical data is less voluminous, but requires much more careful handling. Among the services it therefore calls for are ‘fine grained’ access control – e.g. through authorization and authentication of users – and privacy protection through anonymization or pseudonymization of individual data or ‘outlier’ detection and disguise in statistical data. There are, of course, many more specialist medical services, as our examples below reveal. It is a current requirement in the United States, for example, that if head images are communicated outside the team immediately caring for a patient, all facial features which might identify the patient must be removed.

Breast Cancer and MammoGrid

Breast cancer is arguably the most pressing threat to women’s health. For example, in the UK, more than one in four female cancers occur in the breast and these account for 18% of deaths from cancer in women. Coupled with the statistic that about one in four deaths in general are due to cancer, this suggests that nearly 5% of female deaths are due to breast cancer. While risk of breast cancer to age 50 is 1 in 50, risk to age 70 increases to 1 in 15 and lifetime risk has been calculated as 1 in 9. The problem of breast cancer is best illustrated through comparison with lung cancer which also accounted for 18% of female cancer deaths in 1999. In recent years, almost three times as many women have been diagnosed with breast cancer as with lung cancer. However, the five year survival rate from breast cancer stands at 73%, while the lung cancer figure is 5%. This is testament to the effectiveness of modern treatments, provided breast cancer is diagnosed sufficiently early. These statistics are echoed in other countries. The lifetime risk of breast cancer in the USA has been estimated as 1 in 8. Here also incidence has increased but mortality decreased in the past twenty years. Twenty years ago breast cancer was almost unknown in Japan but its incidence now approaches Western levels. (For a world-wide picture, see [1].)

The statistics of breast cancer diagnosis and survival appear to be a powerful argument in favour of a universal screening programme. However, a number of issues of efficacy and cost effectiveness limit the scope of most screening programmes. The method of choice in breast cancer screening is mammography (breast X-ray); for precise location of lesions and ‘staging’ (establishing how advanced the disease is) ultrasound and MRI may be used. A significant difficulty lies in the typical composition of the female breast, which changes dramatically over the lifetime of a woman, with the most drastic change taking place around the menopause. In younger women, the breast consists of around 80% glandular tissue which is dense and largely X-ray opaque. The remaining 20% is mainly fat. In the years leading up to the menopause, this ratio is typically reversed. Thus in women under 50, signs of malignancy are far more difficult to discern in mammograms than they are in post-menopausal women. Consequently, most screening programmes, including the UK's, only apply to women over 50.

The increasing use of electronic formats for radiological images, including mammography, together with the fast, secure transmission of images and patient data, potentially enables many hospitals and imaging centres throughout Europe to be linked together to form a single grid-based "virtual organization". It is not yet precisely understood what advantages might accrue to radiologists working in such virtual organizations, as the technological possibilities are co-evolving with an appreciation of potential uses; but one that is generally agreed is the creation of huge "federated" databases of mammograms, which appear to the user to be a single database but are in fact retained and curated in the centres that generated them. Each image in such a database would have linked to it a large set of relevant information, known as metadata, about the woman whose mammogram it is. Levels of access to the images and metadata in the database would vary among authorized users according to their "certificated rights": healthcare professionals might have access to essentially all of it, whereas, e.g., administrators, epidemiologists and researchers would have limited access, protecting patient privacy and in accordance with European legislation.

The Fifth Framework EU-funded MammoGrid project (2002-05) [3] aimed to apply the grid concept to mammography, including services for the standardization of mammograms, computer-aided detection (CADe) of salient features, especially masses and 'microcalcifications', quality control of imaging, and epidemiological research including broader aspects of patient data. In doing so, it attempted to create a paradigm for practical, grid-based healthcare-oriented projects, particularly those which rely on imaging, where there are large volumes of data with complex structures. Clinicians rarely analyse single images in isolation but rather in a series or in the context of metadata. Metadata that may be required are clinically relevant factors such as patient age, exogenous hormone exposure, family and clinical history; for the population, natural anatomical and physiological variations; and for the technology, image acquisition parameters, including breast compression and exposure data.

As a research project, MammoGrid encompassed three selected clinical problems:

i Quality control: the effect on clinical mammography of image variability due to differences in acquisition parameters and processing algorithms;

ii Epidemiological studies: the effects of population variability, regional differences such as diet or body habitus and the relationship to mammographic density (a potential biomarker of breast cancer) which may be affected by such factors;

iii Support for radiologists, in the form of tele-collaboration, second opinion, training and quality control of images.

The MammoGrid proof-of-concept prototype enables clinicians to store digitized mammograms along with appropriately anonymized patient metadata; the prototype provides controlled access to mammograms both locally and remotely stored. A typical database comprising several thousand mammograms has been created for user tests of clinicians' queries. The prototype comprises (a) a high-quality clinician visualization workstation (used for data acquisition and inspection); (b) an interface to a set of medical services (annotation, security, image analysis, data storage and queries) accessed through a so-called *GridBox*; and (c) secure access to a network of other *GridBoxes* connected through grid middleware. The *GridBoxes* may therefore be seen as gateways to the grid.

The prototype provides a medical information infrastructure delivered in a service-based grid framework. It encompasses geographical regions with different clinical protocols and diagnostic procedures, as well as lifestyles and dietary patterns. The system allows, among other things, mammogram data mining for knowledge discovery, diverse and complex epidemiological studies, statistical analyses and CADe; it also permits the deployment of different versions of the image standardization software and other services, for quality control and comparative study.

It was always the intention of MammoGrid to get rapid feedback from a real clinical community about the use of such a simple grid platform to inform the next generation of grid projects in healthcare. In fact, a Spanish company has already entered into negotiations to commercialize the project and to deliver a real, MammoGrid-based radiology service in the region of Extremadura. Thus, many ideas which came up as questions, issues or obstacles in research, must be solved in a real-life system within the next two or three years.

We may now imaginatively consider what may happen in the course of a consultation and diagnosis using the MammoGrid system. A patient is seen and mammograms are taken. The radiologist is sufficiently concerned about the appearance of one of these that she wishes to investigate further. In the absence of any other method, she may refer the patient for a biopsy, an invasive procedure; however, she also knows that in the majority of cases, the initial diagnosis turns out to have been a false positive, so the patient has been put through a lot of anxiety and physical trauma unnecessarily. Given the degree of uncertainty, a cautious radiologist may seek a second opinion: how can the MammoGrid system support her? She may invoke a CADe service; the best among these can identify features which are not visible to the naked eye. Another possibility is to seek out similar images from the grid database of mammograms and examine the history to see what has happened in those other cases. However, since each mammogram is taken under different conditions, according to the judgement of a radiographer ('radiologic technician') it is not possible to compare them as they are. Fortunately, a service exists which standardizes and summarizes the images, provided certain parameters are available – the type of X-ray machine and its settings when the mammograms were taken. Perhaps at this particular moment the radiologist's workstation is already working at full capacity because of other imaging tasks, so it is necessary for the image to be transmitted to a different node for processing. Since our grid is distributed across Europe, it now matters whether the node which will perform the standardization is in the same country or not. Let us suppose that it is a different country. A conservative outcome is to ensure that, provided the regulatory conditions in the country of origin and in the country where the processing will take place are mutually compatible (i.e. logically consistent, capable of simultaneous satisfaction) that they are both complied with. If one set requires encryption, say, but the other does not, the data must be encrypted. If both sets of regulations allow the image to be transmitted unencrypted but one country requires all associated data transmitted with the image to be pseudonymized, this must be done. These are human decisions, but it is clear that they can be automated. Where will responsibility lie if something goes wrong in this process? In any case, the story has further ramifications: the whole idea of MammoGrid is to build up a rich enough database of images and case histories to provide a sound basis both for diagnostic comparison and for epidemiology. Once standardized and returned, is the image now to be stored and made available to others for comparative use, or is it to remain outside the system. This is now a question of informed consent. Will a service, in the sense we have already used the term, be trusted to determine whether such informed consent as the patient has given covers this question?

We now consider the comparison the radiologist wanted to make – the reason for standardizing the image to begin with. The intention is to find images which are sufficiently similar and whose associated history gives an indication of the associated risk. For example, if from among the ten most similar instances, seven turn out to be malignant, there would be good reason to proceed to the more invasive stage of investigation. But how is the database to be queried so as to suggest valid comparisons? Clearly, this goes beyond image similarity. The risks for a childless woman of 65 are very different from a 50-year old mother of three. Image similarity would not be sufficient to warrant a comparison. Thus we must transmit, as part of the database query, data that potentially identify the patient; and the result of the query may provide data which potentially identify patients. On a need-to-know basis, the radiologist has to know details of the cases, but not necessarily the names of the patients, although it would not be difficult to imagine a case where the name reveals something about ethnic background and this turns out to be significant. In a fully deployed system, there may be relevant cases and images from several countries; the system must be capable of 'policy bridging', as described above, to ensure that all regulatory conditions are met. Indeed, if the impact of including a case from one particular country would be to render the comparison less useful overall, perhaps the system should be able to reject that particular case – in other words, to apply a criterion which maximizes the information obtained subject to satisfaction of applicable laws and regulations – where the 'applicable set' is itself a variable.

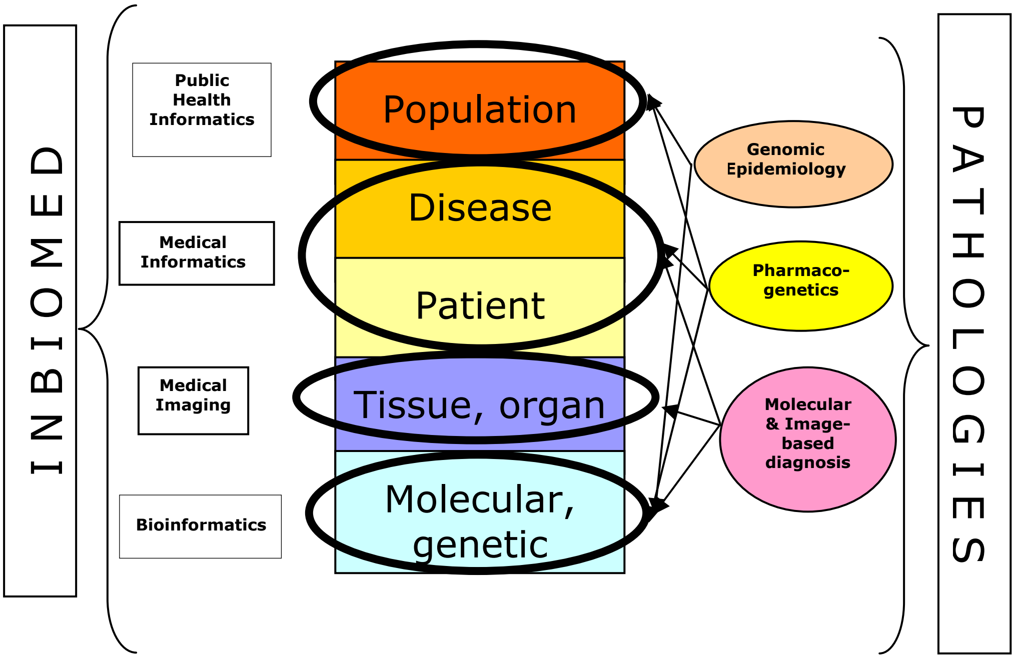
Evidence-Based and Individualized Medicine

Hitherto, I have given a 'naïve' account of one system and its approach to diagnosis. How is such a system to fit into the modern conception of evidence-based medicine, i.e. medicine that is based on scientific results, rather than on the doctor's intuition, personal knowledge and craft skill? Evidence-based practice rests on three pillars: medical knowledge, as much as possible based on 'gold standard' (double-blind, controlled) clinical trials whose results have been peer reviewed and then published; knowledge of the patient, as complete as the record allows; and knowledge of the resources, procedures and protocols available in the setting where the encounter with the patient is taking place.

There is a very extensive literature on knowledge management and the difficulties and opportunities it presents. Some work currently undertaken in the healthgrid context, such as on ontologies and on knowledge representation, is relevant here. A development which is bringing economics into conflict with the traditional approach to the establishment and dissemination of knowledge is online publication of research results. While in medicine at present this is restricted to electronic publication of papers that have already been peer reviewed and are in the pipeline for printing in a journal, in other fields of science, notably physics, immediate online publication of un-peer reviewed results so that they can be viewed and critically assessed is now common. In another field, the journal *Nature* recently conducted a comparative study of errors in *Wikipedia* and in the *Encyclopaedia Britannica*; the results were equivocal, leading some to argue that an online, user-managed encyclopaedia is less error prone, although there have been many hacking attacks on *Wikipedia*. In the case of medicine, not only malicious postings, but poor research may have serious results. The American Medical Informatics Association is currently promoting the concept of a world bank of clinical trials. Here it may be said that the traditional approach to knowledge has failed; negative results are often not published and, as certain legal cases have brought to light, even results suggestive of risks are kept under wraps. Another practice that would benefit from being documented is the effective prescription of certain drugs beyond their designed purpose or licence, where nevertheless anecdotal clinical evidence has led practitioners to believe they are effective.

However, the MammoGrid application we have described above (and other similar projects) takes us a step further in the direction of 'dynamic' construction of knowledge. If images and histories are to be used as part of the diagnostic knowledge in new cases, it is imperative that they are collected with as much care and rigour as the cases in a controlled trial. Therefore, it is essential to know the 'provenance' of the data with precise details of how it has been handled (e.g. if standardized and subjected to CADe, which algorithms were used, set to what parameters, by whom, and if capture and interpretation were subject to appropriate practice standards). I have labelled this set of issues "the question of practice-based evidence for evidence-based practice". If this were to be accepted as an appropriate source of diagnostic information, the underlying grid services which maintain it would have to make quality judgements without human intervention.

A major breakthrough in healthcare is anticipated from the association of genetic data with medical knowledge. In the healthgrid research community we have a map that has become almost an article of faith:



Disciplines, levels of being and pathology diagnostics (acknowledgement: F. Martin-Sánchez)

This view of the 'life' is in fact shared by many different disciplines, system biology being the most obvious among them. Drug development is increasingly driven by a molecular view of the world, using a variety of models to understand both how drugs act and how their action may be enhanced, inhibited or frustrated. This usually means understanding what proteins are present and, therefore, which genes code for those proteins. In the foreseeable future, we may anticipate certain drugs to be available in subtypes to account for the specific genetic endowment of the patient.

This would suggest that genetic information would have to be accessed routinely in the course of healthcare. Viewing this as part of the information held on a patient raises a number of difficult problems. Among these are the predictive value and the shared nature of genetic information. Knowing a person's genome could mean knowing what diseases they may or may not be susceptible to. Knowing one person's genetic map also reveals that of his or her siblings' in large measure. This introduces a range of questions, from confidentiality to 'duty of care' issues. If physicians will be held liable both for what they do and what they do not do, is it necessary for the underlying knowledge technology to 'be aware' and to inform them of the possibilities?

The grid could provide the infrastructure for a complete 'electronic health record' with opportunities to link both traditional patient data and genetic information to bring us closer to the ideal of genomic medicine. Among many questions being investigated in current projects is a set concerning development and illness in childhood, especially conditions in which genetic predisposition is at least suspected and in the diagnosis of which imaging is also essential. Physicians want to know how certain genes impact the development of diseases and radiologists want to know what the earliest imaging signs are that are indicative of a disease. For example, the Health-e-Child project [5] is investigating paediatric rheumatology, cardiac dysmorphology and childhood brain tumours using this approach. Consider its aims:

i To gain a comprehensive view of a child's health by vertically integrating biomedical data, information, and knowledge, that spans the entire spectrum from genetic to clinical to epidemiological;

ii To develop a biomedical information platform, supported by sophisticated and robust search, optimization, and matching techniques for heterogeneous information, empowered by the Grid;

iii To build enabling tools and services on top of the Health-e-Child platform, that will lead to innovative and better healthcare solutions in Europe:

- Integrated disease models exploiting all available information levels;
- Database-guided biomedical decision support systems provisioning novel clinical practices and personalized healthcare for children;
- Large-scale, cross-modality, and longitudinal information fusion and data mining for biomedical knowledge discovery.

With major companies looking to translate research results into products, successful outcomes from this and other projects would bring the scenario described above closer to reality.

Next Steps

The SHARE project, a so-called 'specific support action' within the European Information Societies Technology programme, will over the two years 2006-2007 be seeking to define a research road map that will allow not only the technology to be developed but the social issues also to be addressed, with the goal of establishing a healthgrid as the infrastructure of choice for European biomedical activity in the next ten years. The SHARE collaboration includes both computer scientists, experts on social requirements and medical law specialists. The project begins with the fundamental assumption that technical and social requirements must be addressed concurrently. It has identified these challenges to the modernization of health systems [7]:

- creating and populating, connecting and understanding patient records across organization boundaries and, in due course, across different national health systems;

- increasing the openness and accessibility of systems - e.g. providing patients with ownership of their healthcare record - while
- ensuring privacy, confidentiality and ethical compliance in the socio-legal plane, and
- maintaining data integrity, security and authenticity (e.g. provenance and semantics) in the technical plane;
- providing appropriate levels of authorization and authentication of users across all the services and the citizen;
- discovering, grading and certificating trustworthy sources of knowledge and case information to guide future action; finally,
- winning the trust and commitment of the medical professions at a time of immense change and economic pressure.

At present it seems unlikely that technology will be allowed to determine answers to questions of a legal nature, much less so of an ethical nature. Yet the extent to which we trust financial affairs to the internet and the extent to which we have allowed privacy to be invaded by online transactions, 'cookies' and preference tracking (to say nothing of store loyalty schemes) [8] suggests that we may be more flexible in our attitudes that our legal attitudes may imply. Indeed, as far as personal data are concerned, the financial analogy has been made before in the concept of a personal data bank. Would patients be less trusting of a 'bank' with their health record than they are with their money?

I have argued that 'healthgrid', the augmented application of grid computing to health, presents an opportunity to review not only information technology for health – a major enough task – but also our approach to the complex issues of ethical, legal and regulatory compliance as mediated by the technology. The case in favour of the technology, in terms of improved information and knowledge for clinicians, patients, public health officials, administrators and governments, is not difficult to make. The need for ethical and legal safeguards cannot be circumvented, but in itself this may prove an insuperable obstacle for the deployment of the new technology. One way forward is to analyse precisely these 'social' requirements and enhance the technology with the means to apply them automatically with minimal human intervention.

Acknowledgements

I am deeply indebted to my colleagues, in chronological order, on the MammoGrid, SHARE and Health-e-Child projects; also to the HealthGrid organization and the co-authors of the 'White Paper'; finally to the HealinG consortium. In relation to legal, ethical, security and trust issues, I must particularly thank Brecht Claerhout, Jean Herveg and James Lawford Davies — though I alone am responsible for my mistakes!

References

[1] *Frequency of Cancers Around the World*, The Scientist, Vol. 17, Cancer Supplement, 22 09 2003, at: http://www.the-scientist.com/article/display/14131/

[2] The Information Societies Technology project: *MammoGrid – A European federated mammogram database implemented on a Grid infrastructure*, EU Contract IST-2001-37614.

[3] Warren R *et al.*, *A comparison of some anthropometric parameters between an Italian and a UK population: 'proof of principle' of a European project using MammoGrid* To appear, 2006.

[4] Warren R *et al.*, *A Prototype Distributed Mammographic Database for Europe* To appear, 2006.

[5] The Information Societies Technology Integrated Project: *Health-e-Child – An integrated platform for European paediatrics based on a grid-enabled network of leading clinical centres*, EU Contract Number IST-2005-027749.

[6] The Information Societies Technology Specific Support Action: *SHARE*, EU Contract Number IST-2005-027694.

[7] SHARE Consortium, *The Healthgrid Framework*, not yet published.

[8] David H Freedman, *Why Privacy Won't Matter*, Newsweek (International Edition), 3rd April 2006; at http://www.msnbc.msn.com/id/12017579/site/newsweek/.