# The Development of Data Infrastructures for eHealth: A Socio-Technical Perspective

Jenny Ure
*University of Edinburgh*, urejenny@gmail.com

Rob Procter
*University of Manchester*, rob.procter@manchester.ac.uk

Yu-wei Lin
*University of Manchester*, yuwei.lin@manchester.ac.uk

Mark Hartswood
*University of Edinburgh*, mjh@inf.ed.ac.uk

Stuart Anderson
*University of Edinburgh*, soa@inf.ed.ac.uk

*See next page for additional authors*

Follow this and additional works at: https://aisel.aisnet.org/jais

# The Development of Data Infrastructures for eHealth: A Socio-Technical Perspective

## Authors

Jenny Ure, Rob Procter, Yu-wei Lin, Mark Hartswood, Stuart Anderson, Sharon Lloyd, Joanna Wardlaw, Horacio Gonzalez-Velez, and Kate Ho

# The Development of Data Infrastructures for eHealth: A Socio-Technical Perspective*

**Jenny Ure**
University of Edinburgh
jenny.ure@ed.ac.uk

**Mark Hartswood**
University of Edinburgh
mjh@inf.ed.ac.uk

**Joanna Wardlaw**
University of Edinburgh
joanna.wardlaw@ed.ac.uk

**Rob Procter**
University of Manchester
rob.procter@manchester.ac.uk

**Stuart Anderson**
University of Edinburgh
soa@inf.ed.ac.uk

**Horacio Gonzalez-Velez**
The Robert Gordon University
h.gonzalez-velez@rgu.ac.uk

**Yu-wei Lin**
University of Manchester
yuwei.lin@manchester.ac.uk

**Sharon Lloyd**
University of Oxford
sharon.lloyd@comlab.ox.ac.uk

**Kate Ho**
University of Edinburgh
k.l.ho@sms.ed.ac.uk

## Abstract

*We explore some recurring socio-technical problems encountered in the development of infrastructure for sharing and re-using data across sites and social scales for eHealth research. We link these problems to contradictions between underlying assumptions about data as a commodity whose reuse is not compromised when it is extracted from the context in which it has been captured, and the reality of data as entangled with, and constituted through, local practice. To illustrate these problems, we draw on the experiences of a number of HealthGrid projects developing infrastructures for data sharing and reuse, and trace the strategies that have evolved to address them. These experiences problematize the "one size fits all" model initially adopted by HealthGrids, and highlight the need for design and development strategies that are able to engage with local needs and thereby ensure that the technical infrastructure is properly aligned with the human infrastructure it is supposed to support.*

*Keywords*: e-Infrastructure, socio-technical systems, eHealth, data sharing, ICT design and development strategies, ontologies

## 1. Introduction

The HealthGrid concept is a response to the eHealth vision of accelerated progress in biomedical research and healthcare delivery.[1] Realising this eHealth vision (in common with e-Research[2] more generally), calls for radical changes in the ways in which research is conducted, including how researchers share and reuse data and collaborate (Jirotka *et al.* 2005). The role of HealthGrids is to harness innovations in digital infrastructure that can enable the seamless access to, sharing and reuse data (e.g., clinical records, genomic data, and images) irrespective of source. These innovations in digital infrastructure, known as the Grid or, increasingly commonly, as e-Infrastructure (Cyberinfrastructure in the USA), comprise networked, interoperable, service-oriented, scalable computational tools and services. Ontologies — formalised ways of describing the semantics of data (Gruber 2007a) that can be interrogated by distributed human users and computers so as to facilitate discovery, linking, and reasoning across datasets — are a key element of this e-Infrastructure.

In this paper, we examine a number of challenges for the design and development of HealthGrids. We focus on recurring problems, where assumptions about data sharing and reuse were contested or broke down at different stages and intervention points in the HealthGrid development process, and the implications they have for successful deployment and long-term sustainability. Exploring these problems provides us with a better understanding of the ways in which assumptions can affect the quality, usefulness, and currency of shared data, and of the costs, risks, and benefits for different users. Finally, we note the impact of governance issues raised when data is shared across organizational and national boundaries.

## 2. Challenges in Realising the Vision

The challenges we examine in this paper came to light at a road-mapping workshop in the UK National e-Science Centre,[3] where several HealthGrid project teams[4] came together to identify a number of common concerns at different stages in the data and information lifecycle (Ure *et al.* 2006, 2007a, 2007b). These projects were charged with (a) designing infrastructure for sharing and re-using imaging, genetic, and clinical data in overlapping or related disease domains, and (b) developing ontologies (formal encodings of the concepts within a particular knowledge domain and their relationships) to support this. The paper also draws on the outcomes of a related UK e-Social Science workshop addressing social and semantic aspects of this infrastructure development.[5]

Duguid and Brown (2000) and Bowker and Star (2000) have underlined the social and often political nature of the collection, classification, and representation of information, and how such processes can be facilitated or frustrated by the particular design choices of technical artifacts intended to support them. These sorts of issues were evident in the problems encountered by HealthGrid project teams in their efforts to develop common semantic models capable of representing and sharing data across diverse sites and scales. In addition, difficulties reflected very optimistic assumptions about the potential of standardised protocols and automated data cleaning to address problems of data quality from disparate contexts.

HealthGrids are co-evolving, socio-technical complexes (Joslyn and Rocha 2000) whose key challenge is achieving the effective alignment of coupled technical and human information

---

[1] See http://initiative.healthgrid.org/the-initiative/healthgrids-concept.html

[2] Known originally as 'e-Science', in the UK the term 'e-Research' has emerged as the more inclusive label. For a summary of e-Science aims, see Berman, F. and Hey, T. (2004), "The Scientific Imperative", Chapter 2 in: Foster, I. and Kesselman, C. (eds.), The Grid: Blueprint for a New Computing Infrastructure, 2nd edition, Morgan Kaufman Publishers.

[3] http://www.nesc.ac.uk/esi/events/709/ and https://wikis.nesc.ac.uk/mod/Main_Page

[4] Participating projects included UK: NeuroGrid (http://www.neurogrid.ac.uk/), PsyGrid (http://www.psygrid.org/), CARMEN (http://www.carmen.org.uk/), P3G (http://www.p3gconsortium.org/), Generation Scotland (www.generationscotland.org/); EU: HealthGrid Share (http://initiative.healthgrid.org/the-initiative/share-project.html), HealthAgents (http://www.healthagents.net/); US: Bio-Informatics Research Network (http://www.nbirn.net/).

[5] http://www.ncess.ac.uk/events/item/?item=169

infrastructure. The aim of those involved in the design and implementation of HealthGrids must be to manage unanticipated problems and, where possible, to gain useful synergies, whether these are technical (e.g., through enhanced interoperability of services) or human (e.g., through increased sharing and reuse of data and collaboration in research).

To understand the alignment challenge, we begin by following the data lifecycle (see Figure 1) through data collection, cleaning, and quality control (detecting and eliminating errors) to explore the implications of local working practices and knowledge for the sharing and reuse of data collected across different sites. Local variations in practices cannot be entirely eliminated by standardization of data lifecycle protocols but, more importantly, the protocols themselves will be subject to change as the technologies for data collection evolve. Other solutions must be sought if the goal is to ensure data remains re-usable in the face of such changes. In these circumstances, the capture of metadata (i.e., data about data) that documents the way in which data has been collected and prepared for use (i.e., its provenance) becomes critical to success.

The importance of providing metadata is already recognised among those research communities that have established routines for data archiving. However, with the eHealth vision (and that of e-Research more generally), which demands greater sharing and reuse of data, more sophisticated forms of provenance are seen as being necessary (Moreau *et al.* 2008). Chief among these are models known as ontologies which formally define the semantics of metadata and so enable machine-based reasoning, linking and analysis. However, looking to ontologies to manage this diversity and change merely shifts the problem of standardization rather than eliminates it.



**Figure 1. Aligning technical and human processes at different stages of the data lifecycle.**
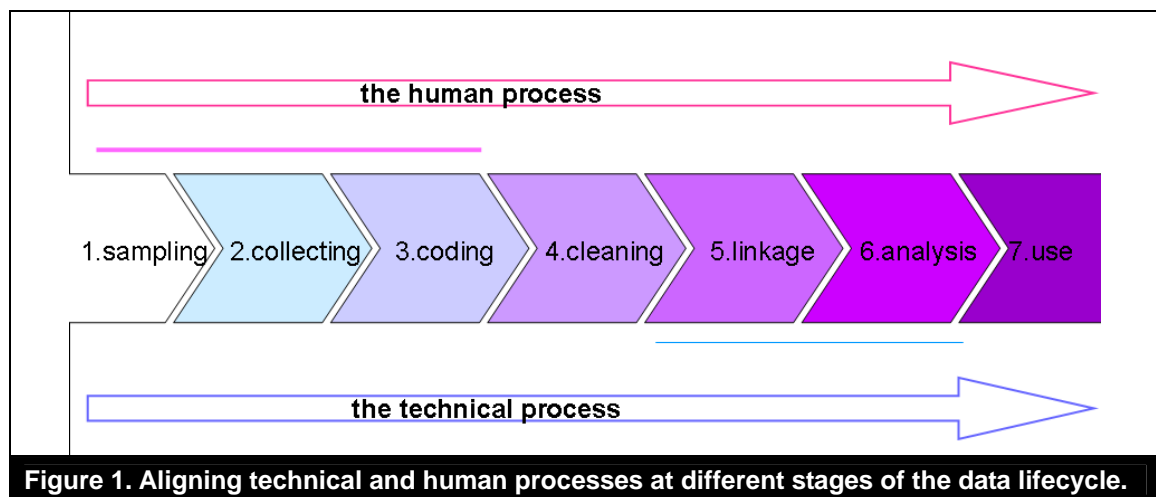
We move on to explore problems experienced by several HealthGrid projects in their attempts to define ontologies to apply within particular research domains. In particular, we contrast the view of ontology builders – that stable and common metadata standards are necessary to support large-scale data sharing (and the assumptions about the nature of distributed knowledge production, representation and governance that follow from this view) — with the need expressed by users for approaches that are compatible with more diverse local aims and practices. Such practices are sensitive to the kinds of practical problems likely to be experienced on the ground at the different sites where HealthGrid infrastructures are expected to be embedded and used (Hartswood *et al.* 2006, Randall *et al.* 2007).

## 3. The Data Lifecycle: the Reality behind the Assumptions

For all of the HealthGrid projects, significant issues were evident in the initial stages of the data lifecycle, where project managers acknowledged that high error rates during data collection as a problem. Data might be entered in the wrong boxes on forms, for example, or might be incomplete or even contradictory. While data checking and data cleaning procedures could identify particular kinds

of anomalies, it was often impossible to differentiate between variance due to disease effects, population differences, or other unspecified differences between sites and data collection teams. Such examples highlighted the (often unrecognised) role of local knowledge and communication in identifying and rectifying data quality issues, and the difficulty of replicating this in large scale, multidisciplinary and distributed collaborations, where informal opportunities for face-to-face communication are greatly reduced.[6]

The experience of the projects was that quality assurance mechanisms were able to identify unexpected data values, but input about the population, the local context, and the processes on the ground was necessary as the study was carried out. For example, differences in resting heart rate between two populations in one of the projects was initially thought to reflect higher rates of blood pressure between samples in different cities, but information gleaned from research nurses led to the discovery that, due to the fact that the lift in one hospital had broken, one group of patients had been tested for resting heart rate after climbing several flights of stairs. Again, this was an accidental discovery arising from an anomaly being discussed while one of the nurses was present and able to draw on this knowledge of the local context. Data sharing and reuse on the scale presumed by the eHealth vision has often been predicated on assumptions about the potential to harmonize protocols and tools consistently across sites. However, in practice, local changes could not be anticipated, and would not be kept current after the end of the project.

Local knowledge and agency was seen to often be central to the ongoing maintenance of data quality, particularly in the face of unanticipated and unpredictable changes in local context and practice. The NeuroGrid project provided a good example of this in the federation of brain scans between multiple sites. A variety of technical procedures are involved in image acquisition, transfer, and storage, and it is often difficult for true disease-related effects to be separated from artifacts of the technical process (Geddes *et al.* 2006). One technique used here is a harmonisation algorithm that accounts for variation due to the different makes and calibrations of scanners. It became clear from a chance exchange among hospital staff on one site that scanners were being serviced on a monthly basis and recalibrated, rendering the algorithm useless. Before this had become evident to the wider team, differences between sites due to calibration would be read as representing disease effects, or possibly population differences.

This highlights the limitations of assumptions about automated data quality mechanisms in large scale, distributed processes, and the importance of local knowledge and agency in ensuring data quality. More worryingly, it also highlights the extent to which unknown bias and errors can reduce the reliability of federated data on this scale. While such issues are not new to the Computer-Supported Cooperative Work (CSCW) research community, their cumulative impact on large multi-site projects of the sort studied here raises particularly acute questions around the interpretation and quality of data, and poses a fundamental challenge to the premise of seamless data sharing, linking, and reuse assumed by the HealthGrid projects.

The projects discovered that data linkage required the coordination and alignment of knowledge and agency at multiple local sites. The mechanisms for achieving this are problematic, however. For example, the EU HealthAgents project[7] (Gonzalez-Velez *et al.* 2009) focuses on the diagnosis and prognosis of brain tumors using magnetic resonance imaging (MRI) and spectroscopy (MRS) data, with MRI/MRS records located in different hospital and clinical centres in Europe and Asia. Here, the challenge was not only to manage data from different makes and versions of MRI scanners, but also to manage the substantially different regulatory infrastructures governing the use of that data, since partners were obliged to comply with the different regional and national governance frameworks relevant to the confidentiality of patient records in the source and target countries.

As a response to this problem, members of one of the genomic HealthGrid projects described how

---

[6] Less frequently discussed is the extent to which data can meaningfully be considered as a transferable asset *per se* (see, for example, Fortier 2006).
[7] http://www.healthagents.net/

they have evolved alternative scenarios for designing infrastructure to manage secure data linkage of confidential patient data. Traditional "role-based" access to data "pulled" from different sites was replaced by locally controlled data "push" led by local managers of data quality and security (McGilchrist *et al*. 2007).

## 4.    Problems in Agreeing on Semantics

Bringing together in the workshop a number of projects in the same disease domain (schizophrenia) allowed for consideration of those datasets required across projects. Core symptom datasets for psychosis were seen as a useful starting point for facilitating data sharing across these projects. These same projects were also finding, however, that reaching agreement on shared data models was exceptionally difficult, even within individual projects. Expectations these projects might have had, that users already engaged in multiple other studies and using familiar conventions for data definition and measurement might adopt new ones without significant incentives or pressure to do so, proved problematic.

The projects also were struggling to align the (often competing) interests of researchers, clinicians and ontologists. For ontologists designing semantic applications for logical inference, the need for well-defined definitions of classes of data and relationships in the disease domain was paramount if they were to create a logically consistent application for integrating or analysing data from disparate sources. For clinicians, the diagnosis of disease states from symptomatology was fuzzier, and more processual than is often imagined. The underlying mechanisms are not always clear, and practitioners' conceptualisations can change significantly as new knowledge comes into play. In the more tangible context of describing physiological characteristics of organs or systems in the body, the variance is more a function of historical preference for particular ways of defining parts and wholes. In imaging studies, for example, the same organ may be broken up into zones that are arbitrary or that follow historical conventions, and data is interpreted and represented in that context.

More worryingly, participants in the projects had also discovered that apparently straightforward semantic classifications often turned out to have different interpretations and implications in different contexts, and for different purposes. As evidenced at the HealthGrid workshop on schizophrenia,[8] there is some dispute among clinicians as to whether this can even be considered a single disease classification, or is simply an umbrella term for a variety of conditions as Turetsky *et al*. (2002) suggest. Many common diseases are multi-factorial and open to a wide range of shifting interpretations and reinterpretations in the light of emerging findings. If such classifications are difficult in hard sciences, they are considerably more so in bio-medical domains, where disease concepts are often fluid, ambiguous, and evolve within and between professional and national communities of practice (Dupre 2006).

### Standardising Domain Models

Data at the molecular level on synaptic proteins involved in human mental illness, such as schizophrenia, are more valuable when integrated with scanning data, genetic data, and data on treatment, as illustrated in the BioInformatics Research Network (BIRN) test-bed. Achieving this kind of integration requires convergence on a common model for mapping data from different sites, and at different scales, as illustrated in Figure 2.

For some of the HealthGrid project design teams, there was an implicit assumption that the domain was "out there" and that the challenge was to facilitate the documentation and the clarification of this at a relevant level of detail and in an acceptable format. While there was some acceptance of the fact that concepts of disease would vary, and that compromise was required, there was an underestimation of the extent to which tangible physical structures were differently conceptualized, bounded, or defined in relation to other structures for specific purposes. On closer examination, these were not merely semantic differences. Often they reflected the different practical purposes for which

---

[8] Re-use or Re-invention - a Roadmap for Data Integration (Schizophrenia as a Test Case)
 http://www.nesc.ac.uk/esi/events/709/

this information was required, the context of use, and the power of particular groups to shape adoption in practice within communities (Bodenreider *et al.* 2004, Bergman and Lyytinen 2002). Decisions about the adoption of particular standards or coding formats, for example, have implications for costs and benefits to different groups.
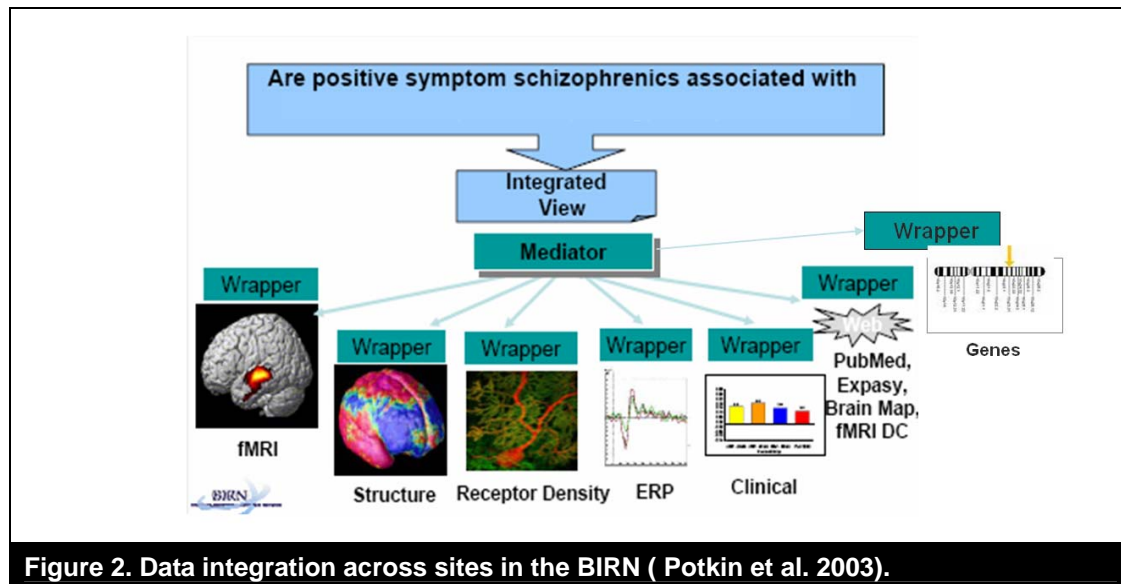


**Figure 2. Data integration across sites in the BIRN ( Potkin et al. 2003).**

To some extent, these existing taxonomies for coding medical data played a role in shaping or constraining usable models of the domain. Within NeuroGrid, for example, the Stroke research node benefited from using SNOMED[9] as a common frame of reference, with the health service community dealing daily with stroke patients on a large scale. The psychosis research node, however, had a particular interest in cross referencing with classification systems related to brain morphology, given the research focus on the size and location of lesions in psychosis. The Foundational Model of Anatomy (FMA) ontology,[10] cross-referenced to the Unified Medical Language System (UMLS)[11] frameworks used by the BIRN and other brain-mapping communities, was the preferred reference framework for sharing and integrating disparate data sets. The difficulty in reaching agreement on codes, classifications, and models was often dependent on historical and professional dependencies that are not easily reconfigured, even if agreement can be reached in principle.

Ontologies have been presented as semantic infrastructures through which data from many sites can be brought together in a framework that is meaningful to both technical and human agents. The belief in such ontological solutions to practical diversity is widespread; as Goguen (2005: p. 1) has observed, "many ontologists seem to believe in the possibility of a single unified ontology that attracts consensus because it 'reflects the real underlying reality' of a domain." Accepting conceptual diversity as a starting point, argues Goguen, suggests a different view, in which knowledge engineering "should seek ways to support it, rather than ways to overcome, suppress, or subvert it" (*ibid*: p. 1).

Problems reported by ontologists in a number of projects suggest that biological and bio-medical concepts often have socially constructed attributes that do not lend themselves to shared ontological representation using the conventions of formal logic used to specify classes and relations in ontologies (Martone *et al.* 2004, Martone 2006). Rector and Rogers (2001) and Dupre (2006) highlight the inconsistencies in many of our current concepts of biological entities and the difficulty, therefore, of achieving logically consistent shared models of spatio-anatomical elements, borders and relationships to support automated data integration and analysis. Bodenreider *et al.* (2004) described

---

[9] http://www.ihtsdo.org/
[10] http://sig.biostr.washington.edu/projects/fm/
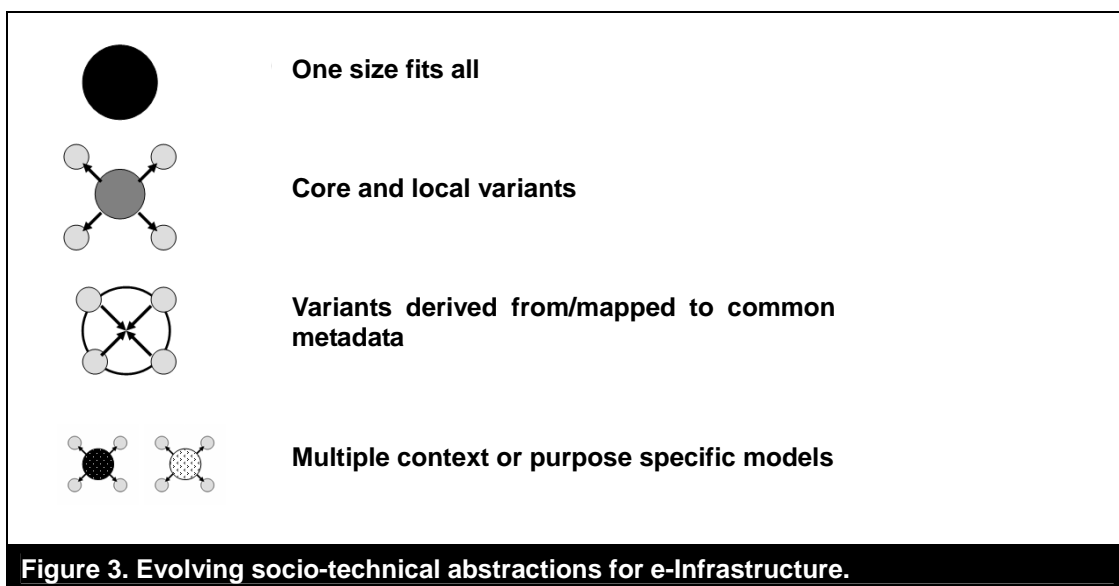[11] http://www.nlm.nih.gov/research/umls/

the problem as "the intrusion of the epistemological in the ontological." It is perhaps less than surprising, then, that the initial vision has evolved in response to the challenges thrown up in real world contexts.

# 5.    Aligning Distributed Technical and Human Infrastructure

The workshop provided an opportunity to expose for discussion and reflection the tensions between the premise of a stable, interoperable infrastructure for data sharing and re-use, and the reality that community-based information infrastructure is dynamically reconfigured, with multiple task-specific variants. Examples from other domains allowed HealthGrid project teams to consider different alignments of technical and distributed human information systems that have evolved to square this circle, and indicate how different arrangements impact the quality and usability of e-Infrastructure in real world contexts.

It is, however, unusual for different scenarios to be presented clearly to the wider project community in an effort to transparently evaluate the costs, risks and benefits to different groups. (Researchers were anxious to optimise access, for example, while gate-keepers and patient representatives were understandably anxious to minimise data linkage that could identify patients and sensitive data.) The process and its different stages here are consonant with the picture of requirements engineering described by Bergman *et al.* (2002), who argue that large-scale system requirements are constructed through a political decision process in complex socio-technical spaces, at different stages, where technical, social, economic, and institutional factors are brought together in a current solution space that provides the baseline for construction by stake-holding parties.

Within the different sites collaborating in HealthGrid projects, community interaction was actively sought as a resource for validating, updating, adapting and enhancing the quality of data and processes, and in dealing with the confidentiality and security issues associated with sensitive data. However, HealthGrid project leaders saw attempts to support communication in distributed communities of users as hard to sustain due to the time commitments required and also because of the barriers to access. For instance, users were required to obtain and install Grid access certificates and remember passwords. Representative users and, in particular, clinical users were often unable or unwilling to devote time to work that seemed distant from or irrelevant to the immediate local concerns of patients. This was seen as a significant weakness in the design of a genuinely usable data infrastructure, geared to real concerns and problems on the ground.



**One size fits all**

**Core and local variants**

**Variants derived from/mapped to common metadata**

**Multiple context or purpose specific models**

**Figure 3. Evolving socio-technical abstractions for e-Infrastructure.**

We identified four basic design strategies used in the different HealthGrid projects as means to resolve the tensions between interoperability and local usability (see Figure 3). These range from generic models that are scalable (but not locally usable) to local models that are easily understood and used in practice, but may not be interoperable with those of other communities. The early "one size fits all" approach was based on a top down classification that was hard to implement in practice. This has evolved to allow for local variants, and in the most recent models, onotologies are created bottom up from local models, for specific purposes.

## One Size Fits All

The ontology community's initial ambitions to create a "one world" view where all knowledge of a bio-medical domain might be encoded within a single, unified ontology is increasingly under challenge. Attempts to develop an ontologically consistent, machine-readable model of human physiology drawing on the Foundational Model of Anatomy (Rosse *et al*. 1998) make assumptions about the nature of data, which, in practice, have been harder to realize than perhaps originally anticipated. For example, many of the features of physiology that must be represented are not overtly evident on inspection, but are rather constructed. In the context of neuroscience, different groups segment and label the hippocampus in different ways that relate to constructions within the community over time rather than to observable characteristics of the hippocampus. Goguen (2005) challenges the one world view and suggests, instead, a need to support multiple, evolving ontologies for single domains and to provide tools to help construct partial mappings or so-called "faceted" ontologies (Motta 2007).

An interesting variation on the one world view is the use of the collaboratory (Olson *et al*. 1998, Kling *et al*. 2003) to shape/construct common understandings where possible and to manage diversity through collaboration, critical mass, and open access to tools and resources. In common with open source software strategies (Metcalfe 2007, Feller *et al*. 2007), open access can align the work of distributed groups to common ends, adding value for the network as a whole. The collaboratory approach allows collective and collaborative knowledge acquisition, annotation, and integration, and provides incentives for alignment with shared standards. The Neuroscience Information Framework (http://nif.nih.gov/) is a good example of this, building on the earlier work of the BIRN.[12] One outcome of the HealthGrid workshop was the recommendation that the participating projects collaborate in this initiative.

## Core and Local Variants

A central issue in defining shared ontologies is the difficulty of balancing the benefits of a stable semantic infrastructure against the need to accommodate the diverse preferences of user groups and the speed of change within the knowledge domain. Such trade offs are rarely evident to ontology builders until the initial prototype is demonstrated to clinicians at different sites, but can require costly redesign or compromise if the ontology is to be deployed in ways that allow users to work effectively with it. Building ontologies is a high cost investment, and ensuring their usability and sustainability is, to a large extent, dependent on squaring this circle.

One common strategy is the separation of a fairly stable core from those elements that are likely to differ among domains and user communities, or are likely to evolve and change over time. It is clear, for example, that there will be a need to incorporate new tests and treatments. The ontology task force on the BIRN project found it helpful to separate the more stable structural aspects of classification in brain imaging from the more functional ones with contextual and purpose-specific dependencies (Martone 2006).

A variation on the core and local strategy is the separation of a higher order conceptualisation of the domain (so-called "upper" ontologies[13]) to which more domain or purpose-specific ontologies can then be mapped. In effect, the upper ontology serves as "semantic glue" between different application domains. For this reason, three of the participating HealthGrid projects that were working on different

---

[12] http://wiki.esi.ac.uk/Re-use_or_Re-invention_-_a_Roadmap_for_Data_Integration
[13] http://suo.ieee.org/

aspects of stroke imaging were exploring the use of the DOLCE upper ontology (see Figure 4).
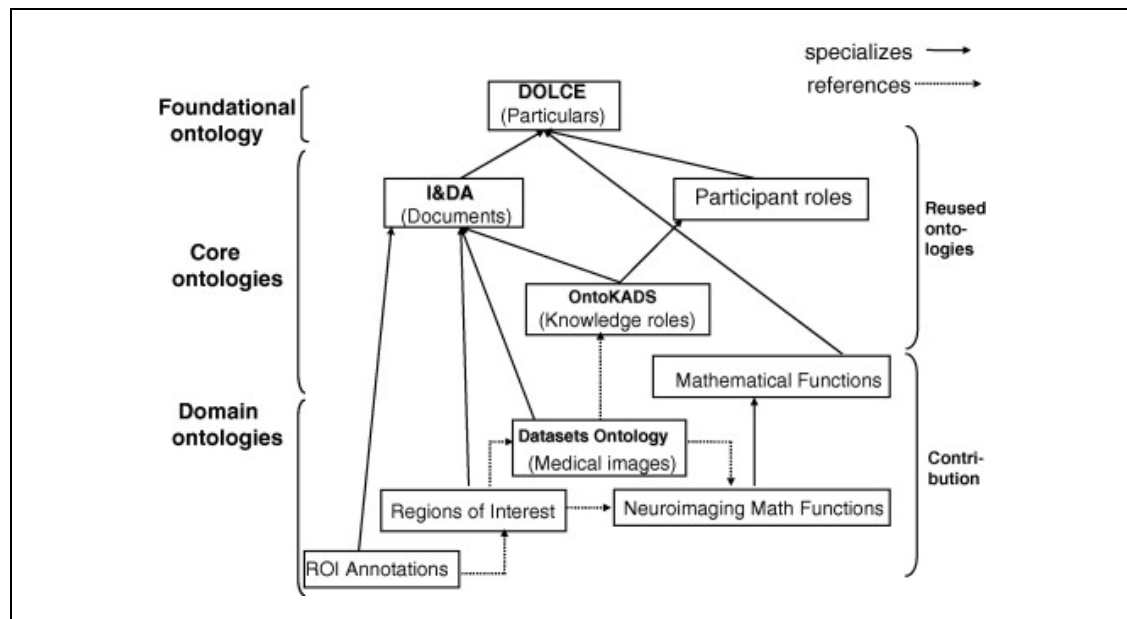


**Figure 4. DOLCE as a unifying top-level structure in OntoNeuroBase (Temal et al. 2008).**

An emerging approach for managing change over time is to capture core (i.e., stable) elements using a formal ontology language and to allow a range of less formal representations to evolve "at the edge" around different views and different (and often transient) purposes. A workshop on building ontologies[14] hosted by the UK National Centre for e-Social Science provided examples of this approach applied in other domains and of how it could be taken further in a user-centered direction. PolicyGrid (Edwards *et al.* 2009), for example, uses a core ontology in combination with folksonomies (i.e., user-generated metadata or tags). The former provides an element of formal structure to meet the need for machine-based reasoning, while the latter enables the leveraging of users' knowledge to meet evolving community requirements (see Figure 5). Gruber (2005, 2007a, 2007b) has also highlighted the potential benefits of integrating ontologies and user-generated tags in this way.

This view of semantic infrastructure not only supports local use but provides for subsequent leverage of distributed expertise and local agency in generating, managing, and sustaining content. Such approaches are particularly relevant in domains where the dynamic, evolving, and socially constructed nature of concepts is of the essence.[15] It draws also on a long tradition of work (e.g., Resnick 2002) on the leverage of social capital in system design.

## Purpose- and User-Specific Strategies

Motta (2007) predicts that in the next generation of semantic technologies, there will be a move from the goal of a centrally designed, monolithic ontology, through core and local variants, toward more purpose-specific solutions. These would be created through the automatic integration of ontology fragments, sourced bottom up from users to meet transient needs at different times and for different purposes. According to Motta (2007), this will enable data infrastructures to "aggregate data in a much more dynamic fashion, automatically identifying the semantic resources relevant to the current need, doing away with the single ontology assumption and performing both ontology mapping and co-reference resolution on the fly."

---

[14] http://www.ncess.ac.uk/events/item/?item=169&active=&subactive=

[15] Tagging data aligns formal reasoning (logical inference, database query etc) with context-dependent community based semantics (labels, groupings, online identities).
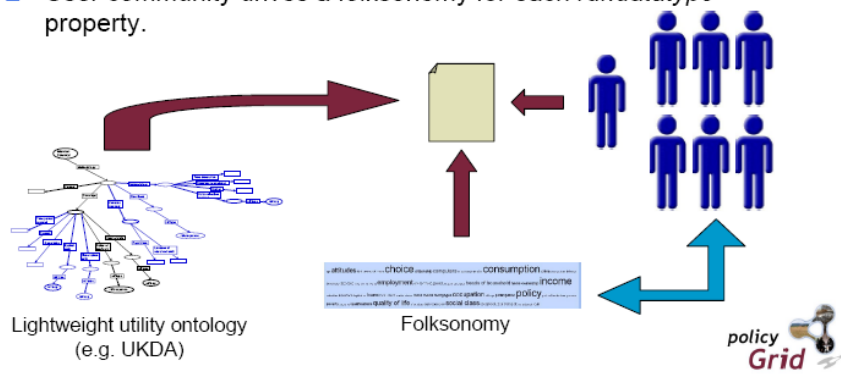
**Figure 5. Combining Ontologies and Folksonomies in PolicyGrid (Edwards 2007).**

This "automated" solution seems somewhat in tension with some of the points made earlier in relation to local practice and social intermediation, however, it highlights the move to infrastructure that might support a "many worlds" view, where the design challenge is to provide infrastructure that is compatible with and shaped by diverse constructions of the world, rather than seeking to constrain these to fit a pre-existing one. The diversity that has until recently been regarded as a challenge for designers of scalable standardized infrastructure has, paradoxically, been seen as an opportunity in eBusiness. Sites such as Amazon, for example, leverage user and community knowledge and agency as means of individualizing content and services.

## 6.    Conclusions

The experiences of various HealthGrid projects show clearly that there is a tension between the technical ideal of a stable, interoperable infrastructure for data sharing and reuse, and the reality of knowledge as evolving, socially and locally constructed, and often disputed. This was most evident in projects where applications were very diverse, or where there was rapid change.

Diversity and change was seen as an unwelcome challenge by HealthGrid project teams. In other sectors, however, there is evidence that this challenge can be met through the adoption of strategies that set out to leverage distributed human resources more effectively (Comfort 2002, Tapscott and Williams 2006). We argue that similar approaches could be used to drive the evolution of e-Infrastructures in eHealth. Recent developments such as Google Health provide infrastructure for patients to access, edit, and link their health records and already provide a sandpit for exploring a radically different paradigm (Vascellaro 2008, Mandl and Kohane 2008) where infrastructure is shaped by and for users.

Some of the key issues faced by the HealthGrid projects relate to the reconfiguring of disparate local, regional, and national governance structures for the use of data. The ethical, legal, professional, and political implications of wider access and linkage of patient data were generally not fully anticipated by the project teams, leading to delays in acquiring ethical approval and in addressing the risks of legal challenge where patient confidentiality might be breached. Jirotka *et al.* (2005) examine this issue in

the context of the UK eDiamond project,[16] and this was also widely reported at other EU eHealth[17] road-mapping workshops. These risks can be severe as, for example, when legal challenges forced the shelving of a national bio-banking project in Iceland (Abbott 2004). This example illustrates the extent to which new technical infrastructures can outstrip the ability of existing ethical, legal, and governance infrastructures to police them (Spinardi and Williams 2005).

When critical design decisions are made early and have unanticipated implications for practice, they will often be hard to change within the constraints of project time scales or budgets. McGilchrist *et al.* (2007) outline a range of design scenarios that redistribute the risks and benefits of different architectures for managing record linkage and the security of patient data in very different ways and would allow for such issues to be explored before crucial design decisions are made. Similar issues were also raised in relation to harmonisation among European HealthGrids (Breton *et al.* 2005, Wilson and Lessens 2006), underlining a need for a transnational consensus process if HealthGrids are to share data across national boundaries.

The collective experiences of the HealthGrid projects' infrastructure development reveal recurrent challenges faced at different stages, from data collection through to the creation of mechanisms for governance, and highlights the need for changes in design and development strategies. Some of these challenges are specific to eHealth, but many are arguably also evident in other sectors (Ure *et al.* 2002, 2005). This suggests a need, in particular, for:

1. better understanding, documentation, dissemination, and reuse of recurring problem: solution scenarios (Williams 2006);
2. opportunities for knowledge transfer from other sectors addressing similar challenges (Sawhney and Parikh 2001, Jha *et al.* 2007, Tapscott and Williams 2006) and from studies of other examples of infrastructural innovations (Edwards *et al.* 2007);
3. enhanced support from funding and support organisations such as e-Research centres to extend the opportunities for providing shared spaces for bringing stakeholders together in this way, and the provision of incentives for doing so.

We have seen how workshops at the UK National Centre for e-Science and UK National Centre for e-Social Science have helped exploration of the challenges of developing technical and human infrastructure for eHealth. We argue that the priority now must be for national organizations for e-Research to play a more significant role in bringing stakeholders together across domains to share recurring problem: solution scenarios, and to provide a forum for collaborative governance – by design rather than by default.

## Acknowledgements

## References

Abbott, A. (2004). "Icelandic database shelved as court judges privacy in peril", Nature 429: 118.

Bergman, M., King, J.L. and Lyytinen, K. (2002). "Large-Scale Requirements Analysis Revisited: The need for Understanding the Political Ecology of Requirements Engineering", Requirements Eng, 7:152–171, Springer Verlag.

Bodenreider, O., Smith, B. and Burgun, A. (2004). "The Ontology-Epistemology Divide: A Case Study

---

16 eDiamond Project (http://www.ediamond.ox.ac.uk)
17 EU HealthGrid Share Roadmap Project (http://www.eu-share.org)
    EU IRG e-Infrastructure Project (http://e-irg.eu/roadmap/eIRG-roadmap.pdf)

in Medical Terminology", In Proceedings of the International Conference on Formal Ontology and Information Systems, Torino, November.

Bowker, G.C. and Star S.L. (2000). "Sorting Things Out: Classification and Its Consequences." MIT Press (Paperback).

Breton, V., Dean, K. and Solmonides, T. (2005). "The HealthGrid White Paper", In Solomonides, T., McClatchy R., Breton V., Legre, Y. and Norager, S. (eds.) From Grid to HealthGrid. IOS Press ISSN 0926-9630.

Comfort, L. (2002). "Anticipating Fire: A Socio-technical Approach to Mitigation", Technology. Vol. 7, No. 53.

Davies, R.H., Twining, C.H., Allen, P.D., Cootes, D.F. and Taylor, C.J. (2003). "Shape Discrimination in the Hippocampus Using an MDL Model", In C.J. Taylor and J.A. Noble (Eds.): IPMI 2003, LNCS 2732, pp. 38–50.

Duguid, P. and Brown, J.S. (2000). "The Social Life of Information", Boston, MA: Harvard Business School Press.

Dupré, J. (2006). "Scientific Classification", Special Issue of Theory, Culture and Society of Problematizing Global Knowlege, 23, pp. 30-32.

Edwards, P. (2007). "Why Ontologies Are Only Part of the Answer for Humanities & Social Science", National Centre for e-Social Science Agenda Setting Workshop, University of Manchester. Available at: http://www.ncess.ac.uk/events/ASW/ontologies/presentations/pete_edwards_ontologies_200 70323.pdf

Edwards, P., Farrington, J., Mellish, C., Philip, L., Chorley, A.H., Hielkema, F., Pignotti, E., Reid, R., Polhill, J.G., Gotts, N.M. (2009). e-Social science and evidence-based policy assessment: challenges and solutions. Social Science Computer Review, Vol. 27, No. 4 (in press).

Edwards, P.N., Jackson, S.J., Bowker, G.C. and Knobel, C. (2007). "Understanding Infrastructure: Dynamics, Tensions and Design", National Science Foundation Workshop Report, Workshop on History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures, Univ. of Michigan. Available at: http://deepblue.lib.umich.edu/handle/2027.42/49353

Feller, J. Fitzgerald, B., Hissam, S.A. and Lakhani, K.R. (2007). "Perspectives on free and open source software". Cambridge, MA: MIT Press.

Geddes, J, Lloyd, S., Procter, R., Ure, J., Simpson, A., Rossor, M., Fox, N., Hill, D., Hajnal, J., Lawrie, S., McIntosh, A., Johnstone, E., Wardlaw, J., Bath, P. and Watson, G. (2006). "Designing for eHealth: Recurring Scenarios in Developing Grid-based Medical Imaging Systems, HealthGrid 2006, Valencia.

Goguen J. (2005). "Support for Ontological Diversity and Evolution" A short essay for the SEEK (Science Environment for Ecological Knowledge) project meeting, October 27. Available at: http://www.cs.ucsd.edu/users/goguen/papers/onto-intgn.html

Gonzalez-Velez, H., Mier, M., Julia-Sape, M., Arvanitis, T.N., Garcia-Gomez, J.M., Robles M., Lewis, P.H., Dasmahapatra, S., Dupplaw, D., Peet, A., Arus, C., Celda, B., Van Huffel, S. and Lluch-Ariet, M. (2009). "HealthAgents: Distributed Multi-Agent Brain Tumor Diagnosis and Prognosis". Applied Intelligence, June;30(3):191-202.

Gruber, T. (2005). "TagOntology - a way to agree on the semantics of tagging data". Presentation to Tag Camp, Palo Alto, CA, October 29. Available at: http://tomgruber.org/writing/tagontology.htm

Gruber, T. (2007a). "What is an Ontology?" Available at: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

Gruber, T. (2007b). "Ontology of Folksonomy: A Mash-up of Apples and Oranges". Available at: http://tomgruber.org/writing/ontology-of-folksonomy.htm

Hartswood, M., Ho, K., Procter, R., Slack, R. and Voss, A. (2006). "Grammars of collaboration: designing for e-Science." In Designing for e-Science: interrogating new scientific practice for usability in the lab and beyond. Edinburgh: National e-Science Centre, January. Available at: http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=613

Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., Hinds, C. and Voss, A. (2005). "Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study", Computer-Supported Cooperative Work (CSCW) Journal, Vol.14, No. 4.

Joslyn, C. and Rocha, L.M. (2000). "Towards Semiotic Agent-Based Models of Socio-Technical

Organizations". In H.S Sarjoughian et al. (Eds.) Proc. AI, Simulation and Planning in High Autonomy Systems (AIS 2000) Conference, Tucson, Arizona, USA, pp. 70-79.

Kling, R., McKim, G. and King, A. (2003). "A Bit More To IT: Scholarly Communication Forums as Socio-Technical Interaction Networks". Journal of the American Society for Information Science and Technology 54(1), 47-67.

Jha, S. (2007). "Distributed Programming Abstractions: What are the Challenges for Distributed High Performance Applications". National e-Science Centre Theme and Workshop. Available at: http://www.nesc.ac.uk/talks/789/jha_esi_pub_final.pdf

Mandl, K.D. and Kohane, I.S. (2008). "Tectonic shifts in the health information economy". New England Journal of Medicine, April 17;358(16):1732-1737.

Martone, M., Gupta, A. and Ellisman, M.H. (2004), "e-Neuroscience: challenges and triumphs in integrating distributed data from molecules to brains", Nature Neuroscience Vol. 7, No.5, May.

Martone, M. (2006). "Building Multiscale Models of the Nervous System through the BIRN", In Reuse or Reinvention: a roadmap for data integration, National e-Science Centre, 27 November.

McGilchrist, M., Sullivan, F. and Kalra, D. (2007). Assuring the confidentiality of shared electronic health records: Sharing data between multiple institutions offers better regulation and public protection, BMJ;335:1223-4.

Metcalfe, R. (ed) (2007). "JISC's Sustainability Study: A case study review of open source sustainability models". Available at: http://www.jisc.ac.uk/media/documents/programmes/distributed_elearning/sustainabilitystudy-1.0.pdf

Moreau, L. and Ludäscher, B. (Eds.) (2008). "The First Provenance Challenge". Concurrency and Computation: Practice and Experience, Special Issue: Vol 20, Issue 5 pp 409-418.

Motta, E. (2007). "Ontology-Based Applications in the Age of the Semantic Web", 10th Intl. Protégé Conference, July 15-18, Budapest, Hungary.

Nonaka, I. and Nishiguchi, T. (eds.) (2001). "Knowledge Emergence: Social technical and Evolutionary Dimensions of Knowledge Creation", Oxford University Press.

Olson, G.M., Atkins, D.E., Clauer, R., Finholt, T.A., Jahanian, F., Killeen, T.L., Prakash, A. and Weymouth, T. (1998). "The Upper Atmospheric Research Collaboratory (UARC)". Interactions, 5 (3) pp. 48-55.

Potkin, S.G., Turner, J.A., Brown, G.G., Glover, G.H., Heckers, S., Keator, D.B. and Grethe J.S. (2003). "Biomedical Informatics Research Network: Functional Imaging Research in Schizophrenia Test Bed". In: Soc. for Neuroscience. New Orleans, LA.

Randall, D., Sharrock, W., Rooksby, J., Lin, Y-W. and Procter, R. (2007). Ontology Building as Practical Work: Lessons from CSCW. Proceedings of the 3rd e-Social Science conference, Ann Arbor, MI, USA, 7-9 October.

Rector, A.L. and Rogers, J.E. (2000). "Ontological issues in using a description logic to represent medical concepts: Experience from GALEN". Methods of Information in Medicine , Springer Verlag, January.

Resnick, P. (2002). "Beyond Bowling Together: SocioTechnical Capital", In Carroll (Ed) HCI in the New Millennium, J.P. Addison-Wesley, pp. 247-272.

Rosse, C., Shapiro, L.G. and Brinkley, JF. (1998), "The Digital Anatomist Foundational Model: Principles for Defining and Structuring its Concept Domain". In Chute EG (ed): A paradigm shift in health care information systems: clinical infrastructures for the 21st century. JAMIA Symposium Supplement 1998: pp 820-824.

Sawhney, M. and Parikh, D. (2001). "Where Value Lives in a Networked World", Harvard Business Review, January, pp. 175-198.

Spinardi, G. and Williams, R. (2005). "The governance challenge of breakthrough science and technology". In: Lyall, Catherine and Tait, Joyce (Eds.) 'New modes of governance : developing an integrated policy approach to science, technology, risk and the environment'. Aldershot, Ashgate, pp.45-66.

Tapscott, D. and Williams, A. (2006). "Wikinomics: How Mass Collaboration Changes Everything", Atlantic Books.

Temal L., Dojat M., Kassel G. and Gibaud B. (2008). "Towards an ontology for sharing medical images and regions of interest in neuroimaging", Journal of Biomedical Informatics, Vol. 41, Issue 5, October 2008, pp. 766-778

Turetsky, B.I., Moberg, P.J., Mozley, L.H., Moelter, S.T., Agrin, R.N., Gur, R.C. and Gur, R.E. (2002). Memory-Delineated Subtypes of Schizophrenia: Relationship to Clinical, Neuroanatomical, and Neurophysiological Measures, Neuropsychology, Vol. 16, No. 4.

Ure, J., Lloyd, A. and Malins, J. (2002). "Beyond Constructivism: Generative Networked Environments", Alt-C 2001, Edinburgh.

Ure, J. and Jaegersberg, G. (2005). "Invisible Architecture: The benefits of aligning people, processes and technology." British Computer Society.

Ure, J. (2006). "Reuse or Re-invention - a Roadmap for Data Integration (Schizophrenia as a Test Case)". Available at: http://www.nesc.ac.uk/esi/events/709/ and https://wikis.nesc.ac.uk/mod/Main_Page

Ure, J., Procter, R., Martone, M., Porteous, D., Lloyd, S., Lawrie, S., Job, D., Baldock, R., Philp, A., Liewald, D., Rakebrandt, F., Blaikie, A., McKay, C., Anderson, S., Ainsworth, J., Van Hemert, J., Blanque,R I., Sinnott, R., Barillot, C., Gibaud, B., Williams, A., Hartswood, M., Watson, P., Smith, L., Burger, A., Kennedy, J., Gonzalez-Velez, H., Stevens, R., Corcho, O., Morton, R., Linksted, P., Deschenes, M., McGilchrist, M., Johnson, P., Voss, A., Gertz, R. and Wardlaw, J. (2007a). "Designing for eHealth", In Proc. HealthGrid 2007, Geneva. IOS Press.

Ure, J., Procter, R. and Lin, Y. (2007b). "A Socio-technical Perspective on Ontology Development in HealthGrids", In Proc UK e-Science All Hands Meeting, 10-13 September, Nottingham.

Vascellaro, J.E. (2008). "Google Helps Organize Medical Records", Wall Street Journal, May 20.

Williams, R. (2006). "Infrastructures and Architectures". UK e-Science Centre Workshop, 27 September. Available at: http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=700

Wilson, P. and Lessens, V. (2006). "Rising to the Challenge of e-Health across Europe's Regions". In Proc. eHealth 2006, Malaysia, May.

## About the Authors

**Jenny Ure** is a Research Fellow in the School of Informatics with a focus on the alignment of people, processes and technology in socio-technical systems in e-Health and e-Business.

**Rob Procter** is Professor and Research Director at the UK National Centre for e-Social Science at the University of Manchester. His research interests include socio-technical issues in the design, implementation, evaluation, adoption and use of interactive computer systems, with emphasis on ethnographic studies of work practices, computer-supported cooperative work and participatory design.

**Yuwei Lin** is a Research Associate at the UK National Centre for e-Social Science at the University of Manchester. Her research interests include science and technology studies (STS), free/libre open source software (FLOSS) studies, and social and organisational issues involved in the development and implementation of information and communication technologies.

**Mark Hartswood** is a Research Associate in the School of Informatics, Edinburgh University. His research interests include Medical Informatics, Computer Supported Cooperative Work and Participatory Design.

**Stuart Anderson** is a senior lecturer in the School of Informatics at Edinburgh University. His main research interest is the dependability of socio-technical systems, in particular the analysis of the role of risk and trust in such systems.

**Sharon Lloyd** is a project manager for health related IT projects at the University of Oxford and has worked on many academic research projects involving technology development. She is also a research facilitator for the Computing Laboratory and researches project management methodologies for collaborative academic projects.

**Joanna Wardlaw** is Professor of Applied Neuroimaging and Honorary Consultant Neuroradiologist at the University of Edinburgh. She is the imaging PI for a large multicentre trial of thrombolytic treatment in stroke (IST-3), and has a longstanding interest in imaging in acute stroke having been involved in several previous multicentre stroke treatment trials. She is director of the Scottish Imaging Network, A Platform for Scientific Excellence (SINAPSE, www.sinapse.ac.uk), a pooling initiative established by the Scottish Funding Council who part fund her work.

**Kate Ho** is a final year PhD student in the School of Informatics at Edinburgh University, researching requirements engineering for Grid infrastructure in eHealth and eScience.

**Horacio Gonzalez-Velez** is a lecturer with the School of Computing at the Robert Gordon University, researching computational science, with particular emphasis on algorithmic modelling for the biomedical sciences.