# e-Labs and Work Objects: Towards Digital Health Economies

John D. Ainsworth* and Iain E. Buchan

School of Community Based Medicine,
University of Manchester,
Manchester Academic Health Science Centre,
Manchester, M13 9PL, United Kingdom
{john.ainsworth,iain.buchan}@manchester.ac.uk

**Abstract.** The optimal provision of healthcare and public health services requires the synthesis of evidence from multiple disciplines. It is necessary to understand the genetic, environmental, behavioural and social determinants of disease and health-related states; to balance the effectiveness of interventions with their costs; to ensure the maximum safety and acceptability of interventions; and to provide fair access to care services for given populations. Ever expanding databases of knowledge and local health information, and the ability to employ computationally expensive methods, promises much for decisions to be both supported by best evidence and locally relevant. This promise will, however, not be realised without providing health professionals with the tools to make sense of this information rich environment and to collaborate across disciplines. We propose, as a solution to this problem, the e-Lab and Work Objects model as a sense-making platform for digital health economies - bringing together data, methods and people for timely health intelligence.

**Keywords:** Health Intelligence, Collaboration, Work Objects, e-Lab, Digital Economy, Health Economy, Analysis Workbench.

## 1 Introduction

In the 1970s Archie Cochrane and colleagues alerted the medical profession to the need to weed out subjectivity and anecdote from clinical practice [1]. At the same time there was a move to improve the safety of medicines. Since then the evidence-based care movement has grown and is now accepted by most healthcare professionals to be best practice. However, there are serious problems with the evidence on which healthcare and public health practice is based: it is expensive to produce; it takes a long time to produce; it takes a long time to influence professional practice; it is crude, relating to the average participant and simple treatment definitions under ideal conditions – in other words, it gives a low-resolution picture of how a patient might respond to treatment or a how a subgroup of the community might respond to a public health intervention. There is

* Corresponding author.

R. Mehmood et al. (Eds.): EuropeComm 2009, LNICST 16, pp. 205–216, 2009.

also a lack of public benefit from investments in science and public services, due to fragmentation of communities, data and analytical methods. In other words, silos of research that could be more effective and efficient if the researchers had easy ways to find and share resources when they need them. The divisions are common between disciplines, for example social vs. biomedical science investigations of obesity. But they also exist within disciplines, for example between biomedical scientists investigating nutritional vs. physical activity components of obesity. Most of the health informatics literature on electronic health records and putting evidence into practice is about weaving the existing evidence-base into healthcare decision-making. The role of clinical information systems in improving the evidence-base, however, has been neglected, but they are essential to providing a timely and more flexible evidence base for future healthcare. This future could be called *high resolution healthcare*; it would enable personalised medicine, efficient and opportunistic clinical trials, complex (including genomic) epidemiology, and tactical development of local services based on local environmental factors and outcomes at the population level. High-resolution care and research requires information systems to link relevant data, methods and people in a clear and timely fashion.

The history of public health intelligence shows rapid advancement in the discipline through the application of information technology [2], [3] and [4]. Increasingly complex analysis methods requiring High Performance Computing (HPC) resources are being used. Simultaneously, there has been a rapid increase in the range of data sources available to the public health practitioner, encompassing electronic health records, research databases, geographical information systems and socio-demographic profiles. Ubiquitous connectivity and middleware enables HPC resources to be shared, and data collections to be accessed from anywhere. However the applications used to make sense of these electronic resources themselves tend to be very specific to the problem being addressed resulting in isolation of outputs and duplication of effort when the same problem is solved for each discipline [5].

## 2 Related Work

Over the course of the past decade, we have witnessed the growth of e-Science [6] and much progress in developing the middleware required for sharing resources, both computational and data. The plethora of Grid frameworks [7] and grid deployments represents the main thrust of these efforts, but it has not become the universal infrastructure envisioned by its pioneers [8]. In fact the most successful Grid deployments are actually as part of a complete vertical application such as the CERN Large Hadron Collider Grid. Service Oriented Architectures (SOA), usually realised through Web Services, offer an alternative approach to sharing, typically by providing a workflow tool for orchestration [9]. The e-Science movement has also spawned numerous Virtual Research Environments [10] drawing on the collaboratory concept [11], but no generic, reusable, electronic equivalent of the laboratory workbench or lab notebook has emerged. The Open Provenance

Model [12] provides a standard way of capturing the history of the production of digital objects, with the goal of providing repeatability of *in-silico* experiments. myExperiment [13] draws on the social networking paradigm to provide a platform for curating and sharing scientific workflows. myExperiment also contains an aggregation mechanism known as a "pack", which enables user to bind related artefacts together. This capability is further developed as Research Objects in [13]. The Open Archives Initiative (OAI) have developed a standard for aggregating web-based resources through the Object Reuse and Exchange protocol [14], which is being widely adopted within the digital repositories community. The concept of Boundary Objects, as a means of cross-discipline communication, was first identified by Star and Griesemer [15] two decades previously.

## 3 Motivating Use Cases

The use cases presented below serve to illustrate the need for an electronic laboratory for health.

### 3.1 Obesity Investigations

The obesity epidemic [16] and its potential to break financial models of healthcare has raised the urgency of understanding the epidemiology of obesity and the effectiveness of large-scale measures to tackle it. However, identifying the determinants of obesity, which are very complex, requires understanding social and behavioural as well as biomedical mechanisms [17]. Obesity-relevant information is contained in a number of large surveys, such as Health Surveys for England and the British Household Panel Survey. However, these surveys are difficult to navigate, and are under-used in obesity research. The difficulty arises from the number of variables measured in each survey, and subtle differences in measurement techniques and variable names, which can only be resolved by digging through supporting documentation. Researchers fail to learn from one another about finding, extracting and analysing relevant data. Furthermore, individual researchers may be unable to reproduce an analysis, based on a complex survey after they have forgotten the steps they took. The statistical analysis is usually encapsulated in scripts, but this is not usually chained to the data extraction. Surveys that are repeated on a regular basis, for example the annual Health Survey for England, may have differences in measurement, sampling, or simply labelling of variables, which makes analysis across surveys difficult. It is unsurprising therefore that social and health scientists asking similar questions using HSE would usually in isolation from one another. Social researchers don‘t usually know where or how to get at the full range of data relevant to obesity research, for example data collected by healthcare services or schools. And for obesity research in the public health service, there is often a lack of analytical capacity, for example to resolve spatial or temporospatial patterns of obesity from geocoded data sets.

### 3.2 Genetic Epidemiology

Understanding the genetic basis for disease, and how genetic factors interact with environments and behaviours is a grand challenge for science. Biotechnologies are providing vast amounts of genetic and gemonic data. For example, out of the three million or so genetic factors that vary between people, half a million factors can now be measured on a blood sample for around two hundred dollars. These points of variation, or Single Nucleotide Polymorphisms (SNPs), are usually studied for their relation to disease states by running statistical analyses over tens of thousands of study subjects, hundreds of thousands of genetic factors and a handful of other factors such as age. This is a computationally expensive task [18], even with the crudest types of analysis. Ideally more relaistically complex analyses, such as seeking clusters of interacting genetic factors, would be commonplace, but this is restricted by statistical and computational limits at present. The development and/or application machine learning methods may make the more compelx analyses tractable. Validation of the causal relationship between a genetic variation and a disease state, must take into account environmental exposures of individuals as these may contribute significantly. This information can be acquired through a clinical study of the cases or from medical records. The successful interpretation of genotype and phenotype data requires a specialist understanding of the disease. The ideal genomic research information system would enable collaboration between methodologists (bioinformaticians, biostatisticians and biomathematicians), domain experts (clinicians, epidemiologists and biologists) and computer scientists. The system would provide a timely thinking space for teams of experts to co-develop insights into the genetic basis of disease from a combination of perspectives.

### 3.3 Pharmacovigilance

Post-marketing surveillance of medicines (also known as Phase IV of clinical trials) is required to assess the safety, and to some extent the effectiveness, of newly licensed medicines ‘in the wild’, oustide the artificial environments of clinical trials. Phase III clinical trials do not usually include all of the types of patient, for example women of child bearing age or patients with other dieases taking other medicines, who might be eligible for treatment with the drug after it is licensed. Therefore the evidence from clinical trials does nto provide a full picture of the public health implications of the drug. Regarding saftety: a system of Adverse Event Reporting (AER) is employed, which relies on clinicians identifying, and reporting harmful affects to a central authority. It may be the case that adverse reactions are not identified as being caused by a particular medicine and so not reported. Important signals about the safety and effectiveness of newly licensed medicines could be extracted from electronic health records. For example, if patient A has the same indication for new medicine X as patient B, but patient A’s physician is not yet prescribing X, then a natural experiment takes place - the challenge is to identify appropriate natural control patients like A and make careful statistical analyses to compare X with existing treatment ‘in the wild’. However, there is no central database that can be analysed; the data is held within multiple systems that not only cover a subset of the population but it further fragmented by the type of care being provided, typically primary and secondary care. The difficulty of combining the relevant data is further compounded by the need to preserve patient privacy and to comply with the information governance requirements of each organisation that holds a part of the patient's overall the health record. An ideal system would enable analysts to extract anonymised data across a federation of electronic health record databases, effectively treating it as a single virtual population data set. Effective analysis requires a combination of statistical method expertise and clinical expertise to interpret the findings [19].

### 3.4 Modelling Healthcare for Populations

Long-term conditions, such as Coronary Heart Disease (CHD), consume the largest proportion of healthcare budgets, and are a major focus of public health initiatives. Moving interventions 'up stream' to earlier stages of disease would reduce the amount of suffering over the average lifetime and save money. Health policy makers and those planning and managing local health services are poorly served by over-simple estimates of the potential public health impacts of making changes to the pathways of care or taking preventive public health measures. These estimates are often unreliable [20], because the models do not represent the complexity of the disease, population or care over time. It is possible to construct graphical models [21] and to use Discrete Event Simulation to model a disease in a population [22]. Such a simulation would enable the user to test various different scenarios, with the ability to modify both clinical and public health interventions, and measure both the effectiveness based on clinical outcomes and costs. Larger simulations, in terms of the population size, results in better accuracy but require greater computational resources. Discrete event simulations are amenable to parallelisation, and so there is a benefit to employing HPC resources. The construction of models requires collaboration between health economists, epidemiologists, biostatisticians and typical decision-makers/leaders (public health professionals, healthcare managers, and clinicians). The execution of simulation scenarios is of interest to public health professionals, clinicians and service commissioners and the results of simulations are used to inform policy decisions. The ideal system would enable user to construct and share models around 'what if scenarios' easily; to execute individual simulations quickly; and to share simulations and their results.

### 3.5 Use Case Summary

From these domain specific use cases, we can identify a set of common requirements. The electronic laboratory must:

1. Provide a mechanism for organising work, such that it can be shared, repeated, audited, reused and reviewed.
2. Provide easy access to resources such as data sets and computational resources.

3. Provide support for the scientific method such that investigations can be planned, constructed, executed recorded and repeated.
4. Provide support for collaboration through the formation of ad-hoc communities of interest, both within and between disciplines.

Our goal is to support both the reuse of content and the reuse of software. The curation and discovery of content via Work Objects within an e-Lab can serve to act as an organisational memory, as training materials, as an accelerant to the discovery process, and as means to reduce duplication. Within health services we envisage a key benefit of the e-Lab/Work Object paradigm to be analytical capacity building among the workforce. The reuse of content between e-Labs will require a standard interexchange format for Work Objects to be developed. The e-Lab software architecture must foster the reuse of functionality and interoperability, but allow specialisation for domain specific tasks.

## 4 The e-Lab

An e-Lab is an information system for bringing together people, data and analytical methods at the point of investigation or decision-making. It provides a secure environment for managing, exploring and analysing data from anonymised, integrated health records. The functional architecture of the e-Lab is shown in Figure 1. The e-Lab provides access to three different types of workspace for each user: personal space that is private to the user; group collaboration spaces that are visible only to members of the group; and public space that is visible to all e-Lab users. Collaboration facilities – such as people search and messaging – are provided, as is the capability to organise communities of interest around Work Objects. Syndication is available for users to track the development of Work Objects. The e-Lab enables access to computational and data resources. Computational resources may range from private compute clusters, required for secure processing of medical records, genomic data and images, to national and international Grids. The e-Lab embeds anonymised clinical data and enforces information governance policies. The e-Lab provides the capability to link across data sources, to perform statistical analysis and visualise the results. Users may upload their own data sets – retaining full control over access rights – and the e-Lab will add it to the data resource catalogue so that it can be used in the same way as the embedded data resources. We distinguish between ‘expert users’ and ‘routine users’. Expert users are able to create and publish methods to support the knowledge discovery process into the e-Lab as Work Objects. These Work Objects can then be re-used by routine users to accelerate their own knowledge discovery. The e-Lab is secured through both technical and operational governance procedures. Maintaining privacy and confidentiality of individuals whose anonymised medical records are stored in the e-Lab is paramount. Privacy preserving data linkage [23] and statistical disclosure control [24] is used. Furthermore, users are only permitted to access data for which they have the approval of the governance board and full audit trails of all activity in the e-Lab are maintained. The e-Lab employs a Service Oriented Architecture (SOA), which

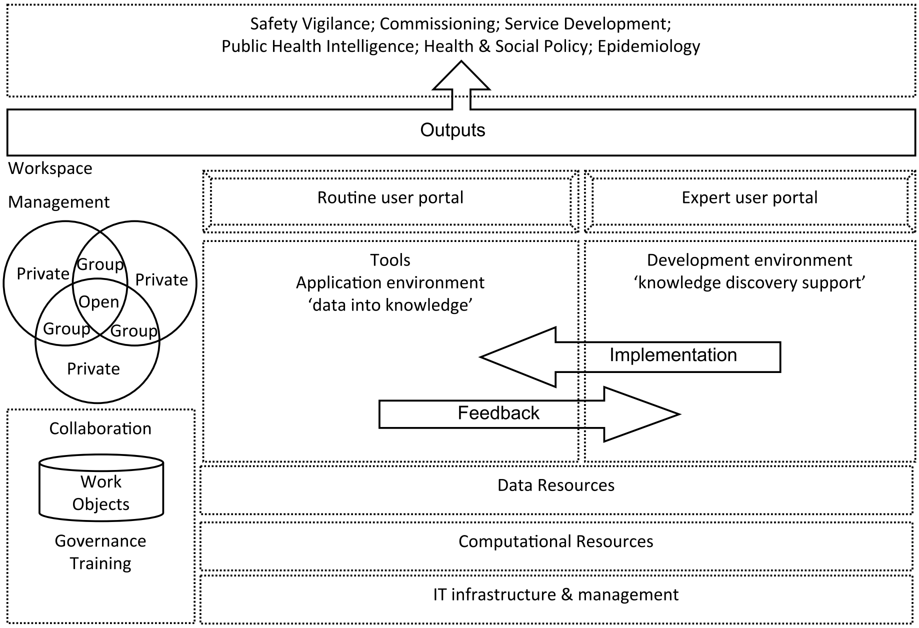

**Fig. 1.** The e-Lab functional architecture

enables both reuse of software and reuse of operational services between e-Lab deployments. We define the core set of e-Lab services to be a Work Object repository, data set repository, metadata catalogue, statistical analysis, visualisation, governance, and access control.

## 5 Work Objects

Work Objects are central to the e-Lab, providing the capability to curate and share information, which in turn builds analytical capacity and organisational memory. Work Objects are collections of digital content assembled to support a specific work task or a series of work tasks – for example to provide a persistent record of an investigation, to publish to a community of interest a statistical method for reuse, or to group together training examples for a tutorial.

***Repeatability.*** A useful analogy can be drawn between a Work Object and a scientific paper. In theory, the paper should give the community all the information necessary to reproduce the results of the research, however there is rarely sufficient information in the paper for another scientist exactly to recreate the investigation. A Work Object representing an investigation can capture all the information necessary to reproduce the results, by recording each step in the process, the data sources used, any transformations applied, the analysis methods and models used, and the commentary underpinning the interpretation of results.

***Reuse.*** Furthering the analogy with scientific papers, a Work Object must be able to reference other Work Objects, in a similar fashion to citations in papers and these references must be navigable. However the Work Object concept goes further. It is possible to embed a Work Object inside another Work Object. For example a Work Object containing a method of statistical analysis could be used inside any number of Work Objects each representing an investigation.

***Permanence.*** A Work Object must provide a persistent record of activity and the associated findings. The process of publishing a Work Object into the public domain must cause a permanent record to be made. A Work Object contains metadata that enables searches to be made over a collection of Work Objects.

***Typing.*** A Work Object must provide a mechanism that enables constraints to be place on its contents, to define application specific content types, and to describe relationships between the content items. This mechanism enables Work Objects to be typed, and consequently systems that are aware of the type of Work Object that they are producing or consuming can provide a richer user experience. The typing of a Work Object requires the specification of the allowed content items, their format and the required number of each; it requires specification of the precedence of content items, for example "data set A and method B must be populated before results C"; it requires specification of production relationship between contents items, for example "executing query I on data source J produces data set K". This specification defines a Work Object's lifecyle that compliant systems will enforce. As an example a Research Object must contain a definition of a research question; the design of the investigation; the ethical approval; the measurements; a record of the steps used to transform the data into results; the results; finished documents about the results. The typing mechanism is extensible, allowing for new types of Work Object to be created as and when required by a community of users.

***Graceful degradation of understanding.*** All systems producing and consuming Work Objects must implement the Work Object as a container; it is not necessary to understand any specific Work Object type. We term these systems Work Object Compliant. An example of this type is a Work Object Repository that is able to store Work Objects, and provides the capability to search for specific Work Objects by querying the metadata. Systems that produce/consume Work Objects and understand one or more types are application specific but are able to reuse components that are Work Object Compliant . Furthermore, Work Objects inherit from OAI ORE [14], and so any system that it Work Object Compliant is also ORE compliant as shown in Figure 2.

Content Items contained in a Work Object maybe embedded directly or indirectly referenced by URI. There are pros and cons associated with either approach. Embedded Content Items can be guaranteed to be immutable and are always accessible. There can be no such guarantees with Reference Content Items, although it may be possible to enforce this through service level agreements with the content provider. It is impractical to embed some content items because of their size, for example genomic data sets, and impossible for others

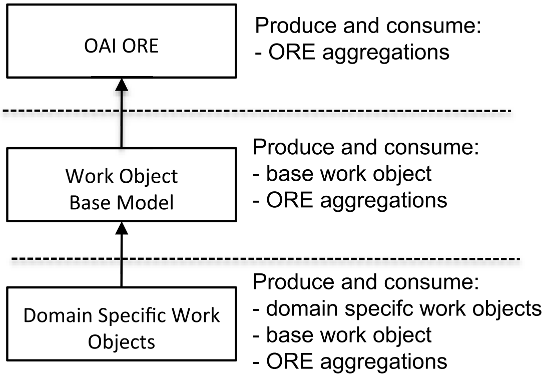

**Fig. 2.** Levels of Work Object Compliance and Understanding

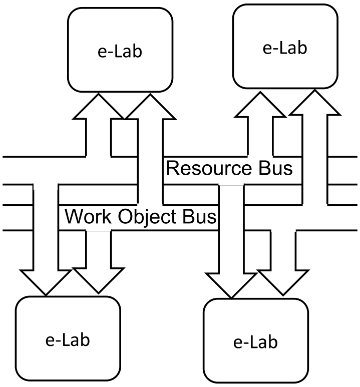

**Fig. 3.** e-Lab Federation

where they are subject to copyright. A published Work Object is considered to be in the public domain, however it is possible to restrict access to Content Items. Any Content Item may be encrypted so that it is not visible without prior arrangement with the author. This applies to both embedded content items and the URI of a referenced content item. Furthermore, although the URI of a Referenced Content Item may be visible, the content provider may apply access control.

## 6 Discussion and Future Work

The e-Lab and the Work Object work together to provide a solution to the problems of resource access, collaboration, reuse and organisation of work. We expect that it will be used in the UK NHS to build analytical capacity and accessible organisational memory. The e-Lab model will be fully developed in the North

West e-Health (NWeH) project, a collaborative effort between the University of Manchester, Salford Primary Care Trust and Salford Royal Foundation Hospital Trust. NWeH is developing the e-Lab and the Work Object software. The first operational e-Lab will be deployed in Salford in 2009, with further deployments following across the North West of England. These community e-Labs will be federated creating a virtual e-Lab for large-scale population-based research containing data on over 2 million people across the North West of England, and Work Objects contributed from NHS personnel from all members of the federation – tapping into the existing culture of sharing across the NHS. The core e-Lab software will be further developed across a range of projects in the Northwest Institute for Bio-Health Informatics (http://www.nibhi.org.uk) including the Shared Genomics Project [18], the Obesity e-Lab [25] and the Manchester Collaboration for Leadership in Applied Health Research and Care (Systems Research Theme, which is producing new methods for care pathway modelling and simulation).

We have presented e-Labs, and their enclosed Work Object repositories. This model can be extend to enable sharing of resources and sharing or Work Objects between communities centred around an e-Lab. We introduce the Resource Bus and the Work Object Bus as a means of federating e-Labs (Figure 3). If a community wishes to trust another community it can export a Work Object, which could be used by the receiving community to accelerate service provision or research and ensure that it is consistent across communities. A published Work Object is considered to be in the public domain, however it is possible to restrict access to it contents. Any content item may be encrypted so that it is not visible without prior arrangement with the author or if the content item is indirectly referenced, the content provider may apply additional access control. The Resource Bus enables the sharing of data and computational resources between Health Economies. The Resource Bus enables users to discover the resources that are available to them from other e-Labs, contingent on the trust relationships that exist between any two e-Labs. These resources can then be used as part of an investigation. For example, e-Labs can expose their embedded health data resources, derived from integrated electronic health records, onto the Resource Bus, creating a single virtual database of the entire population from the participating health economies. The virtual population database can then by accessed in the same way as any embedded e-Lab data resources. This model of distributed collaboration ensures that access control and governance arrangements of each e-Lab are maintained at a local level, which is not possible with traditional approaches that utilise a central data warehouse.

The e-Lab Technical Architecture Group at the University of Manchester was established to bring together projects from disciplines outside of health such as bioinformatics and chemistry. The goal of this group is to standardise Work Objects and define a common, reusable e-Lab infrastructure.

**Acknowledgments.** We would like to thank the members of the Obesity e-Lab (funded by the UK Environmental and Social Research Council), Shared Genomics (funded by Microsoft), the Manchester Collaboration for Leadership in

Applied Health Research and Care (funded by UK National Institute for Health Research), and North West e-Health (funded by the UK Northwest Development Agency) projects for their input and the e-Lab Technical Architecture Group at the University of Manchester for many insightful discussions.

## References

1. Cochrane, A.L.: Effectiveness and Efficiency. Random Reections on Health Services. Nuffield Provincial Hospitals Trust, London (1972)
2. Hersh, W.: Medical informatics improving health care through information. Journal of the American Medical Association 288(16), 1955–1958 (2002)
3. AbouZahr, C., Boerma, T.: Health information systems: the foundations of public health. Bulletin of the World Health Organization 83, 578–583 (2005)
4. Hersh, W.: Health care information technology progress and barriers. Journal of the American Medical Association 292(18), 2273–2274 (2004)
5. O'Carroll, P., Yasnoff, W., Ward, E., Ripp, L., Martin, E.: Public Health Informatics and Information Systems. Springer, New York (2003)
6. Hey, T., Trefethen, A.: The UK e-science core programme and the grid. Future Generation Computer Systems 18(8), 1017–1031 (2002)
7. Stockinger, H.: Dening the grid: a snapshot on the current view. The Journal of Supercomputing 42(1), 3–17 (2007)
8. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
9. Stevens, R., Robinson, A., Goble, C.: myGrid: personalised bioinformatics on the information grid. Bioinformatics 19(90001), 302–304 (2003)
10. Fraser, M.: Virtual research environments: overview and activity. Ariadne (2005)
11. Chin Jr., G., Lansing, C.: Capturing and supporting contexts for scientic data sharing via the biological sciences collaboratory. In: Proceedings of the 2004 ACM conference on Computer supported cooperative work, pp. 409–418. ACM, New York (2004)
12. Moreau, L., Freire, J., Futrelle, J., McGrath, R., Myers, J., Paulson, P.: The open provenance model. Technical Report, University of Southampton (2007)
13. De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., et al.: Towards Open Science: The myExperiment approach. Concurrency and Computation: Practice and Experience (in press, 2009)
14. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., Warner, S.: Open Archives Initative Object Reuse and Exchange (OAI-ORE). Technical report, Open Archives Initative (2007), `http://www.openarchives.org/ore/0.1/toc`
15. Star, S., Griesemer, J.: Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeleys Museum of Vertebrate Zoology, 1907-39. Social studies of science, 387–420 (1989)
16. James, P., Leach, R., Kalamara, E., Shayeghi, M.: The worldwide obesity epidemic. Obesity 9(11s), 228S–233S (2001)
17. Canoy, D., Buchan, I.: Challenges in obesity epidemiology. Obesity Reviews 8(s1), 1–11 (2007)
18. Deldereld, M., Kitching, L., Smith, G., Hoyle, D., Buchan, I.: Shared Genomics: Accessible High Performance Computing for Genomic Medical Research. In: IEEE Fourth International Conference on eScience, 2008. eScience 2008, pp. 404–405 (2008)

19. Bates, D., Gawande, A.: Improving safety with information technology. New England Journal of Medicine 348(25), 2526–2534 (2003)
20. Morabia, A. (ed.): A History of Epidemiologic Methods and Concepts. Birkhauser Verlag, Basel (2004)
21. Bishop, C.: Pattern recognition and machine learning. Springer, New York (2006)
22. Unal, B., Critchley, J., Capewell, S., Liverpool, U.: IMPACT, a validated, comprehensive coronary heart disease model. Technical Report, University of Liverpool, United Kingdom (2006), `http://www.liv.ac.uk/PublicHealth/sc/bua/IMPACT-Model-Appendices.pdf`
23. O'Keefe, C., Yung, M., Gu, L., Baxter, R.: Privacy-preserving data linkage protocols. In: Proceedings of the 2004 ACM workshop on Privacy in the electronic society, pp. 94–102. ACM, New York (2004)
24. Elliot, M., Purdam, K., Smith, D.: Patient record data: Statistical disclosure control for grid based data access. In: Proceedings of the second international conference on e-Social Science (2006)
25. Obesity e-Lab, `http://www.obesityelab.org.uk`