Obesity e-Lab: Connecting Social Science via Research Objects

Iain Buchan[1], Shoaib Sufi[2], Sarah Thew[1], Ian Dunlop[2], Urara Hiroeh[1], Dexter Canoy[1], Georgina Moulton[1,] John Ainsworth[1], Angela Dale[3], Sean Bechhofer[2], Carole Goble[2]

[1]School of Community Based Medicine, University of Manchester, UK

[2]School of Computer Science, University of Manchester, UK

[3]School of Social Sciences, University of Manchester, UK

buchan@manchester.ac.uk

Abstract. Despite a progressive approach to open access datasets, Social Science does not routinely capture and re-use its research processes. This is a barrier to inter-disciplinary research. The public health problem of obesity, with its interwoven social, behavioural and biomedical factors, illustrates the need for more sharable research processes facilitating insights across disciplines. Within this broad need we have identified the central requirement to support secondary research from large surveys such as the Health Surveys for England – a requirement that generalises to other social research topics. We present the e-Laboratory (e-Lab) architecture, for bringing together datasets, investigators and methods around specific questions and packaging the research process into a sharable entity – the Research Object (RO). The Obesity e-Lab project is using obesity research questions and communities to generate a variety of ROs supporting, for example, information mapping between different survey years, transformation of child body mass index measures into research-ready forms, and geo-visualisation of obesity measurements and models. Our collaborators are building e-Labs in other disciplines including biology, health sciences and chemistry. By participating in a programme of building different but interoperable e-Labs, Social Science could stimulate and sustain new research with other disciplines – exporting, importing and coproducing ROs.

Introduction

Social research takes place in a range of settings, both in and around social science. For the deepest insights and impacts social science needs to connect with research processes in other disciplines and settings, for example medical research, economic policy making or local healthcare provider decision making. This connection is impeded by the lack of sharing of reproducible packages of research, incorporating both data and processes for data transformation and analysis. Social Science has, however, led the way among disciplines in the curation of datasets for broad (if not quite open) access to researchers from different disciplines and organizations.

The obesity epidemic provides an example of one such problem which requires more realistically complex research that combines social, behavioural, biomedical and environmental perspectives in ways that might otherwise take place in silos. Clinically,

obesity is a condition of excess fat tissue which can lead to the development of major chronic diseases, such as diabetes and cancer (Kopelman & J, 1998). The rapid rise in obesity prevalence therefore poses an important public health challenge. Although clinical interventions aimed at reducing excess fat in obese individuals are known to be effective, there are not yet any known effective interventions to reduce the obesity burden at a population level (Canoy & Buchan, 2007). Factors such as diet and physical activity are considered 'proximal' determinants of obesity but these factors are generally embedded in a social context within which certain 'obesogenic' lifestyles may perpetuate (McPherson, Marsh, & Brown, 2007). However, research in obesity rarely crosses academic disciplines. Some of the important research questions tend to require expertise in biology, medicine and social science yet the opportunities for collaborative research can be limited. Researchers may be unfamiliar with the research community of other disciplines, lack awareness of relevant data sources or have less understanding of the theoretical concepts underpinning the data collected by other disciplines. Furthermore the infrastructure required to support collaborative working with other disciplines may be lacking.

The e-Laboratories initiative is building a generic architecture to improve access to research data, as well as supporting the sharing of methods and expertise. This architecture has the potential to support a wide range of social research topics. The Obesity e-Lab (ObE) is specialising this approach for the communities, techniques and data sources used in obesity research with a focus on secondary analysis of large surveys, such as the *Health Survey for England* (HSE). The obesity research community is diverse, including social scientists and epidemiologists, as well as government and National Health Service (NHS) analysts. Our aim is to support better use of resources both by reducing some of the difficulties associated with working with large surveys and by promoting collaboration and sharing of practices to help provide answers to obesity-related questions.

In this paper, we review the issues associated with the secondary use of existing data, in particular survey data, and consider previous attempts to support researchers in the sharing of data, methods and expertise. We then describe the concepts and architecture of e-Laboratories (e-Labs) and Research Objects (ROs), and discuss implementation progress in Obesity e-Lab project thus far. We conclude by considering the future development and challenges of the Obesity e-Lab.

Secondary Research on Large Surveys

Secondary analysis is generally used to refer to the additional analysis of survey data originally gathered for other purposes. In the UK the Economic and Social Research Council (ESRC) requires that all data collected during the course of ESRC funded research is lodged in the United Kingdom Data Archive (UKDA) (ESRC, 2009) and made available to other researchers, and other similar projects encourage the publication and sharing of survey data by social scientists globally (King, 2009). This data can then be used for secondary analysis.

These archives of pre-existing survey data represent a massive potential resource for public health researchers and social scientists. However, use of these resources is not always straightforward; a researcher wishing to carry out secondary analysis faces a number of challenges. First, they must identify the most appropriate data source to answer their particular research question. The UKDA alone contains over 5000 datasets, and it can be difficult to get a sense of what is available within each set. Often users pick datasets for reasons of familiarity, for example because they are already in use within their department rather than because they are confident they have identified the correct dataset (Freese, Forthcoming).

Having chosen one or more datasets the researcher must select the variables within the dataset that are pertinent to their research questions. This can be no small undertaking, for example the 2004 HSE questionnaire, a widely used public health dataset, contains approximately 1600 variables (UKDA, 2009). When browsing the list of variables for a dataset, naming conventions do not always make the meaning of each variable apparent, and to properly understand what a variable represents the researcher must consult the accompanying data dictionary and the original questionnaire, often integrating information from several separate documents. Identifying appropriate variables may represent days or weeks of work.

Moving beyond data selection the researcher must consider other issues around the use of the data, for example, the construction of derived variables, harmonisation of a variable that has been measured over several years so that they may be meaningfully compared, how to apply weights to surveys to account for sampling differences [see (Dale, 2006) for a full discussion of these issues]. The expertise and experience of the individual researcher is a significant factor in these operations, and it can be difficult for novice researchers to discover how others have tackled these problems.

Replicability and Provenance

There is growing recognition within the social science and epidemiological communities that space limitations on journals, together with the increasing complexity of analyses, means that authors often have to summarise their results and can rarely include for, example, all the steps used to derive a variable or the procedures used to apply weights. This makes it difficult for other researchers to replicate or develop the published work – despite the fact that replicability may be seen as a quality criterion for quantitative research (Bryman, 2004).

Traditionally, public health researchers and social scientists have used scripts, such as Stata 'do files' as a way of recording their work, for example documenting the creation of derived variables and the steps and parameters used within their analyses. These scripts provide a useful aide memoir to the analyst in developing their thinking about their research, and become an important record of the steps taken to reach their eventual conclusion. Increasingly, researchers are being encouraged to submit their scripts and raw data to journals, both to allow verification or replication of their conclusions, and to allow other researchers to build on past work (Freese, 2007). Advances in statistical analysis and visualization technology are supporting increasingly complex analyses, spread across multiple software packages, so that analysts' traditional methods of recording their work, such as statistical analysis scripts only record one piece of the jigsaw and are less effective. There is a need to document the individual steps taken in the analysis which is not dependent on a single statistical analysis package.

As well as providing a record of their thinking for the researcher, the ability to re-run analyses facilitates the verification of results for publication and the sharing of data and analytical methods which may be extended by other researchers. For example, working out the code to produce specific table formats or graphs is time-consuming and the ability to benefit from others experience is considered very valuable. The approach also enables the researcher to share their work and work collaboratively with other researchers.

Previous Work

The growth of e-Science in the UK during the last decade has spawned the development of technologies that allow sharing of resources particularly for complex data and computational

requirements (Hey & Trefethen, 2002). Amongst the technologies developed include the Grid framework (Stevens, Robinson, & Goble, 2003) as well as the Service Oriented Architectures which typically provide a workflow tool for orchestration using Web Services (Fraser, 2005). Advances in e-Science have also led to the development of infrastructure supporting collaboration and sharing within research. There are a variety of terms used –*Virtual Research Environments* (VRE), *cyberenvironments* (James & Robert, 2007; Liu, McGrath, Myers, & Futrelle, 2007), *collaboratories* – to describe infrastructure supporting collaboration and sharing within research. A VRE provides infrastructure that helps to manage the complexities present when working in distributed collaborations. It comprises a set of online tools and other network resources interoperating to support or enhance the processes of a wide range of research practitioners within and across disciplinary and institutional boundaries. VREs should support the processes of conducting research; be based, where possible on loosely-coupled tools and services; use open standards; and be accountable through the use of appropriate logging services and provenance data (JISC, 2006).

A scientific workflow is the description of a process that specifies the co-ordinated execution of multiple tasks so that, for example, data analysis and simulations can be repeated and accurately reported. Alongside experiment plans, Standard Operating Procedures and laboratory protocols, these automated workflows are one of the most recent forms of scientific digital methods, and one that has gained popularity and adoption in a short time [workflow]. They represent the methods component of modern research and are valuable and important scholarly assets in their own right.

myExperiment (myExperiment, 2009) is a VRE for the social curation and sharing of scientific research objects such as workflows. myExperiment.org has already gathered 900+ users worldwide and caught the imagination of the scientific and the Web communities. To date, myExperiment has primarily been used for the sharing of workflows and in silico experiments, although additional content types are supported.

Web Portals provide web-based applications that integrate information from a number of different services or components, providing a unified presentation for the user and supporting features including single sign-on. Within a portal, a *portlet* provides a self-contained pluggable component. Portals provide collaborative tools for a VRE such as wikis, blogs and shared calendars (Yang, Allan, & J, 2008) but do not provide a generalised framework for handling aggregations or composite objects representing stages in a process

Our approach (the e-Lab) tackles the problem through a focus on the work objects – also known as Research Objects - that are created and manipulated in the course of scientific investigation, along with the services that are needed in order to support the creation, manipulate and publication of those objects. The long term vision of the e-Lab is to support the sharing of objects both *within* and *across* collaboratories, with the research objects encapsulating the shared content that may travel *between* VREs. The concept of Boundary Objects, as a means of cross-discipline communication, was first identified by Star and Griesemer (Star & Griesemer, 1989) two decades previously.

Research Objects represent aggregations of content along with metadata describing the content items and their relationships within the aggregation. The Open Archives Initiative Object Reuse and Exchange (OAI-ORE, 2009) defines standards for the description and exchange of aggregated resources but does not cover aspects such as lifecycle or control over the mutability of both aggregation and content items, which are key to the management of Research Objects.

Within myExperiment, *Packs* allow the aggregation of objects, and provide a partial implementation of Research Objects. Current work on exposing the myExperiment content through data publication using the Resource Description Framework, building on common metadata schema (FOAF, SIOC, Dublin Core) and standardised aggregation mechanisms (OAI-ORE) will support the reuse of this content outside of myExperiment (De Roure et al., 2009).

The e-Lab Approach

An e-Laboratory (or e-Lab) is a set of integrated components that, used together, form a distributed and collaborative space for e-Science, enabling the planning and execution of in-silico experiments - processes that combine data with computational activities to yield experimental results. The experiments are methods when they are instantiated with appropriate parameters, datasets and configurations. They automate the capture of instrument measurements, analyses and visualisations, and form the cross-linking automated pipelines of computational processes (specialist programs, workflows, scripts) that draw upon pooled materials such as datasets, models, parameter sets, publication articles and the results of analytical processes under controlled conditions. These materials are accessed through managed resources increasingly deployed as services. Workbenches, or dashboards, are typically web-browser based portals that give unified access to these electronic materials or rich applications. Logs automatically record the provenance of results arising from these automated processes, including their configuration.

The methods and materials that in-silico scientists use are the e-Lab's digital content; scientific experiment data, scientific models and algorithms, scientific pipelines and workflows, scientific publications, and metadata (annotations and provenance information related to all content and itself content). These encapsulated bundles of digital content are known in the e-Lab as Research Objects (ROs). ROs are the assets that an e-Laboratory operates over. As such they can be linked, replicated, transferred, enhanced and elaborated by multiple users, distributed widely and processed into new forms and generated afresh. Materials spawn and circulate between researchers and other laboratories and into different resources.

Rich metadata is needed in order to support search and navigation, (re)use, management, exchange, integration and execution. In addition, such metadata itself requires management.

Research Objects and Rich Publishing

ROs are the entities that an e-Laboratory: creates, stores, accesses and manages; exchanges with other e-Laboratories; publishes to external sites, deposits in external resources; and displays through work- benches. The motivation behind ROs (and the associated services that produce and consume them) is to improve the curation, accessibility and repeatability of research.

A RO might be:

- A single workflow or collection of workflows with instructions, examples and default input data;
- Laboratory data from instruments, coupled with blogged log book entries;
- A collection of all the digital items associated with one experiment;
- A reproducible research article with the workflows and data required to reproduce the results described in the article;

A RO may contain sufficient information to allow an experiment to be *repeated* including execution/invocation of any services used in the experiment. Alternatively, the RO may be *replayed*, showing the steps that were performed, but without necessarily requiring an execution environment for the services, workflows or applications originally used in the investigation. A RO may be *repurposed*, perhaps through the replacement of one service with another, in order to perform a related analysis.

Research Object Reuse and Exchange

ROs provide a standardised mechanism for the aggregation of resources, along with metadata describing the bundle of data and analyses - for example, a representation of the *research question* that an analyst is hoping to answer, or the fact that a result arises from the invocation of a particular service.

Our implementation of ROs builds on the aggregation mechanisms defined in the Open Archives Initiative Object Reuse and Exchange Specification [OAI-ORE]. ORE provides a vocabulary and abstract model for describing aggregated resources, along with a number of concrete serialisations of the model (in particular in RDF/XML). Our Research Object Upper Model extends ORE and provides additional vocabulary allowing, for example, the description of states in the RO lifecycle: *draft*, *under review*, *published* etc. These states then restrict the transformations or operations that can then be performed on an RO, i.e. preventing modifications on objects in a *published* state.

Particular RO domain schemas can be used to extend these vocabularies and add additional domain specific information to the aggregations. A number of standardised vocabularies are emerging that will also provide vocabulary capturing key aspects of the experimental process (e.g. Open Provenance, SWAN/SIOC, OBO relations ontology)

This layered approach allows services to provide some level of functionality over ROs, without necessarily having to understand the specific details of the relationships represented within the RO. The standardised model also facilitates the exchange of ROs between different e-Labs.

The Obesity e-Lab Architecture

Within the Obesity e-Lab project, we are building a system to support users in navigating, working with and collaborating around survey data and in the process, building ROs focused on their particular obesity based research questions. An example research question might be "Which areas of the North West of England have childhood obesity prevalence higher than national estimates?", the corresponding RO might include HSE data, locally gathered obesity data from schools, scripts to perform weighting and data analysis, and map based visualizations.

Figure 1 provides an overview of the Obesity e-Lab architecture which is composed of four underlying functional modules: *variable selection, analysis scripts, visualization methods* and *reference methods*, which are presented to the user via a single 'workbench' style interface.

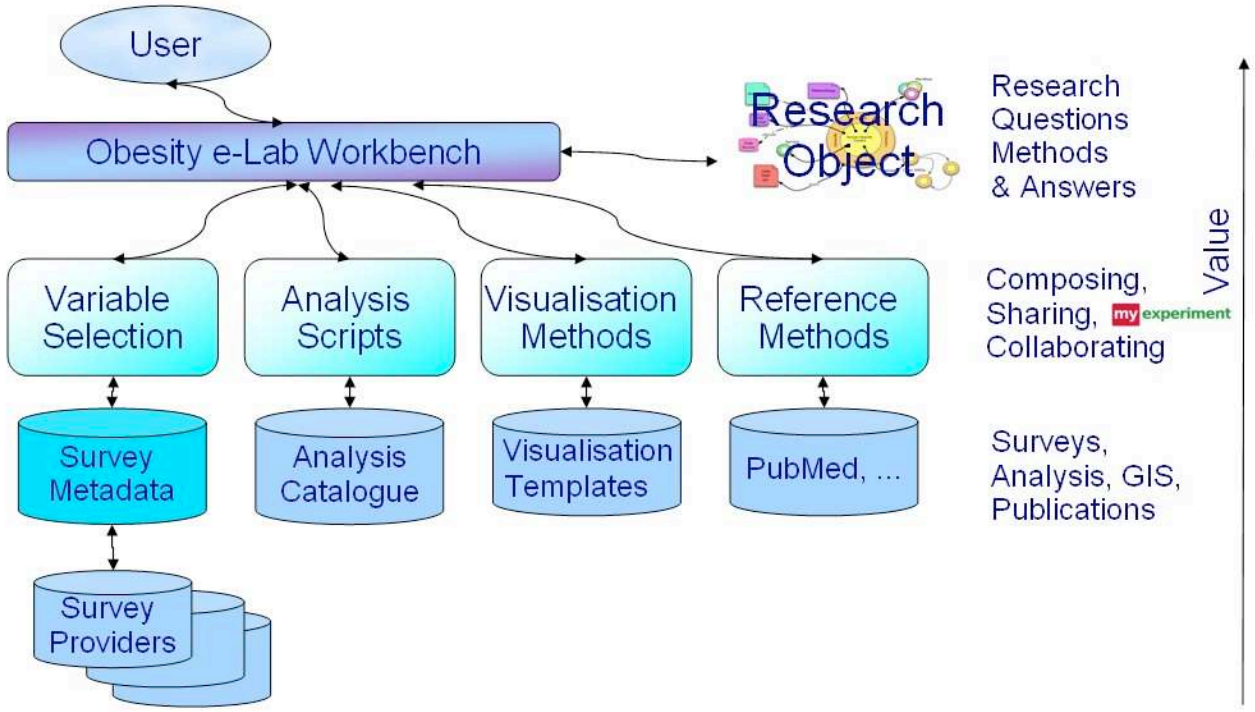

Figure 1: The Obesity E-Lab Architecture

Variable Selection

In order to select appropriate variables from a large survey, a user must navigate a list of potential candidate variables, and review metadata such as the wording of the original survey question or the method of calculation of a derived variable. Currently such information is spread between several different documents. Obesity e-Lab is using text-mining approaches to link this information, allowing the researcher to find the complete variable definition in one location.

The software will also aid users in searching more effectively for variables. Rather than browsing a long, flat list of variables, a faceted browsing mechanism supports the user in searching and filtering the data by different dimensions, effectively reducing the space in which users must search, for example, searching and filtering by year, category (e.g. ethnicity, disease, lifestyle), derived or directly measured.

Analysis Scripts

The Analysis Scripts module has two functions. Firstly it captures a detailed record of the users' research process – for example documenting which dataset and year a user has selected, and which variables they have downloaded, creating the beginning of a RO. Secondly the module supports users in linking their analysis with particular variables. This may take the form of statistical scripts, spreadsheets, pseudo-code or notes. Linking variables with scripts takes the user one step closer to an executable RO, and provides a useful record of their thinking. Moving beyond individual analysis, users can share scripts or expertise, either with the wider Obesity e-Lab community or with a trusted group of colleagues. As these annotations are associated with particular datasets or variables, other users can find and make use of this expertise.

Visualisation Methods

Researchers often want to complement traditional statistical analyses with visual investigations of their data, which can serve a number of different purposes, for example geographic visualisation, hypothesis discovery or statistical process monitoring. Obesity e-Lab will provide a library of visualisation methods, beginning with the ADVISES geo-visualisation application (Thew et al., 2009), to allow the user to incorporate graphical or map-based analysis into their RO. By recording the dataset in use and the chosen parameters these visual analyses can be shared and re-run.

Reference Methods

The system will support the user in linking publication references to their analysis. This would allow a user to link their analysis script to a paper. For example, there are numerous published definitions of 'childhood obesity', a researcher might wish to link their analysis script with the paper defining the classification they have users. Similarly, users could link analysis scripts with resulting publications, providing provenance for their work, as well as helping other researchers build on their outputs.

System development

A working prototype system has been developed which provides access to a cached subset of four years of the HSE to a set of pre-registered users. The prototype focused on supporting the browsing and selecting of data for analysis. Variables were pre-categorized by an epidemiologist, allowing the end user to search and browse for sets of related variables, and the facility to compare variables between years was provided, for example to compare the different ways ethnicity has been categorized over time. Figure 2 shows a user searching for geographic variables within the system, and extending the default search with their own custom search terms. Having identified the variables of interest the user is able to download a chosen sub-set of the data to their desktop for further analysis. This prototype was a Adobe Flex based web application, backed by a Java RESTlet based interface to a MySQL database.

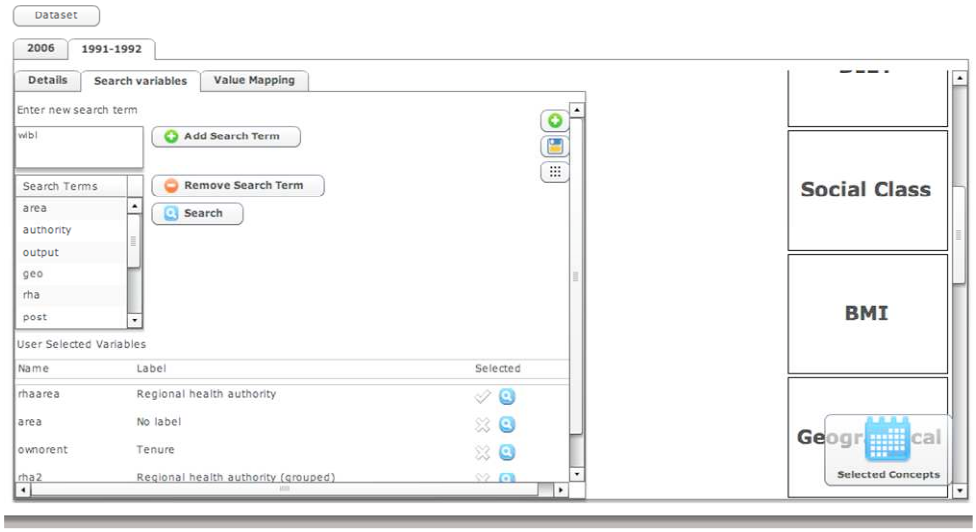

Figure 2. The Prototype – screenshot of the search function

The primary aim of this prototype was to provide experience for the technical team in working with large surveys and as a working proof-of-concept. However, the prototype has also

supported discussions with users about interface design and their requirements and will be a component of our on-going requirements work.

Conclusions and Future Work

The obesity epidemic requires a response which combines expertise from diverse disciplines, including social science, medicine and public health research. Furthermore it requires the collaboration of academic researchers and public health researchers based in the NHS. In response the Obesity e-Lab project has developed and specialised the e-Lab architecture for use in the domain of obesity research, supporting both individual researchers in accessing survey data, and promoting sharing of expertise and collaboration within the research community. A proof of concept prototype has allowed us to test the selection, manipulation and download of variables from the Health Survey for England dataset.

Having established technical proof of concept the project is now working with academic and NHS public health researchers to understand current working practices and investigate their requirements. It can be particularly difficult for users to state their requirements upfront when involved in an innovative software project which will change their working practices (Thew et al., 2009). Consequently we will work closely with public health researchers in iterative cycles of requirements gathering, building and end-user testing, developing a shared vision of an e-Lab, and allowing users to develop an understanding both of their requirements and of the potential of the technology.

Supporting collaborative thinking and the sharing of resources and expertise are important features of this project. Research Objects provide encapsulations of research work, bundling together the resources involved in an experiment or investigation. The e-Lab approach helps to promote sharing and reuse in a number of ways. Firstly it provides a common infrastructure in terms of identified services that can operate over research objects. Secondly it provides a common implementation of that infrastructure and facilitates sharing of objects across laboratories and institutions. The experiences and learning from the Obesity e-Lab project will feed into the larger programme of e-Lab development.

Acknowledgments

This research was funded by the *Economic and Social Research Council* under the *National Centre for e-Social Science* programme.

References

Bryman, A. (2004). *Social Research Methods*. Oxford: Oxford University Press.

Canoy, D., & Buchan, I. (2007). Challenges in Obesity Epidemiology. *Obesity Reviews, 8*(Supplement 1), 1 - 11.

Dale, A. (2006). Quality issues with Survey Research. *International Journal of Social Research Methodology, 9*(2), 143-158.

De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., et al. (2009). *The myExperiment Open Repository for Scientific Workflows*. Atlanta, Georgia, US.

ESRC. (2009). UK Data Archive. Retrieved 6th May 2009, 2009, from http://www.data-archive.ac.uk/

Fraser, M. (2005). Virtual Research Environments: Overview and Activity. *Ariadne*(44).

Freese, J. (2007). Replication Standards for Quantitative Social Science: Why Not Sociology? *Sociological Methods Research, 36*(2), 153-172.

Freese, J. (Forthcoming). Secondary Analysis of Large Social Surveys. In E. Hargittai (Ed.), *Research Confidential*. Ann Arbor, MI: University of Michigan Press.

Hey, A., & Trefethen, A. (2002). UK e-Science Core Programme and the Grid. *Future Generation Computer Systems*(8), 1017-1031.

James, D. M., & Robert, E. M. (2007). Cyberenvironments: adaptive middleware for scientific cyberinfrastructure, *Proceedings of the 6th international workshop on Adaptive and reflective middleware: held at the ACM/IFIP/USENIX International Middleware Conference*. Newport Beach, CA: ACM.

JISC. (2006). JISC VRE Roadmap. Retrieved 6th May 2009, 2009, from http://www.jisc.ac.uk/pub_vreroadmap.html

King, G. (2009). The Dataverse. Retrieved 6th May 2009, 2009, from http://thedata.org/

Kopelman, P. G., & J, S. M. (1998). *Clinical Obesity*. Oxford: Blackwell Science Ltd.

Liu, Y., McGrath, R. E., Myers, J. D., & Futrelle, J. (2007). *Towards A Rich-Context Participatory Cyberenvironment*. Paper presented at the International Workshop on Grid Computing Environments, Reno, NV.

McPherson, K., Marsh, T., & Brown, M. (2007). *FORESIGHT, Tackling Obesities: Future Choices - Modelling Future Trends in Obesity and the Impact on Health*.

myExperiment. (2009). myExperiment. Retrieved 6th May 2009, 2009, from http://myexperiment.org

OAI-ORE. (2009). OAI-ORE. Retrieved 2009, 6th May 2009, from http://www.openarchives.org/ore/

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39 (Vol. 19, pp. 387-420).

Stevens, R. D., Robinson, A. J., & Goble, C. A. (2003). myGrid: personalised bioinformatics on the information grid (Vol. 19, pp. i302-304).

Thew, S., Sutcliffe, A., Procter, R., Bruijn, O. d., McNaught, J., Venters, C. C., et al. (2009). Requirements Engineering for E-science: Experiences in Epidemiology. *IEEE Software, 26*(1), 80-87.

UKDA. (2009). Health Survey for England. Retrieved 6th May 2009, 2009, from http://www.esds.ac.uk/government/hse/

Yang, X., Allan, R., & J, F. (2008, 3-5th December 2008). *The NCeSS Portal - a Web 2.0 Enabled Collaborative Virtual Research Environment for Social Scientists* Paper presented at the 4th International Conference on Semantics, Knowledge and Grid Beijing, China.