Widening Uptake of e-Infrastructure Services

Widening Uptake of e-Infrastructure Services

Alex Voss[1], Marzieh Asgari-Targhi[1], Rob Procter[1], Peter Halfpenny[1], Stuart Dunn[2], Elpiniki Fragkouli[2], Sheila Anderson[2], Lorna Hughes[2], Mike Mineter[3], David Fergusson[3], Malcolm Atkinson[3]

[1] ESRC National Centre for e-Social Science, University of Manchester

[2] Arts & Humanities e-Science Support Centre, King's College London

[3] National e-Science Centre, University of Edinburgh

Email address of corresponding author: alex.voss@ncess.ac.uk

Abstract. This paper presents work conducted in the e-Uptake project which aims to widen the uptake of e-Infrastructure Services for research. We focus specifically on the identification of barriers and enablers of uptake and the taxonomy developed to structure our findings. Based on this, we describe our model of uptake and the development of a number of interventions we are developing. These include training and outreach events, workshops and the deployment of a UK 'one-stop-shop' for support and event information as well as training material. Finally, we will describe how the project relates to other ongoing community engagement efforts in the UK and worldwide.

Introduction

Existing investments in e-Science and Grid computing technologies have helped to develop the capacity to build e-Infrastructures for research: distributed, networked, interoperable computing and data resources that are available to underpin a wide range of research activities in all research disciplines. In the UK, the Research Councils and the JISC are funding programmes to support the development of essential components of such infrastructures such as National Grid Service (www.ngs.ac.uk) or the UK Access Management Federation (www.ukfederation.org.uk) as well as discipline-specific efforts to build consistent and accessible instantiations of e-Infrastructures within specific discipline areas, for example, the e-Infrastructure for the Social Sciences (Daw *et al*. 2007). As the technical components and architectural styles are maturing, questions about uptake and embedding of e-Infrastructures in day-to-day working practices of researchers come to the fore. Indeed, one may argue that if these issues are not addressed, the e-Science community will not realise its full potential and to achieve sustainability.

Consequently, funders are complementing their investments in e-Infrastructures by active programmes of community engagement (Voss *et al*. 2007). As part of the community engagement strand of its e-Infrastructure programme, JISC has funded the e-Uptake project, a collaboration between the ESRC National Centre for e-Social Science at the University of Manchester, the Arts & Humanities e-Science Support Centre at King's College London and the National e-Science Centre at the University of Edinburgh. In this paper we present the project's activities to date to address issues around the *widening of uptake* of e-Infrastructure

services by eliciting information about the barriers to and enablers of uptake, developing adequate interventions such as training and outreach events, running workshops and the deploying a UK 'one-stop-shop' for support and event information as well as training material.

Research Approach

To achieve the aims of the project we need to look beyond isolated, contingent or random problems that people have encountered in employing e-Infrastructure services. Rather, we seek to identify recurring, widespread barriers that can be overcome by a set of targeted interventions that the project will make or that suggest strategies which might be followed up by e-Infrastructure stakeholders. Furthermore, the study must reflect the diversity of the target population (research active members of the UK academic community), their different interests and possible uses of the services (from the Access Grid Support Centre to the National Grid Service) and the number of potential factors influencing uptake (from individual practices to organisational factors and wider research policy). It is important that we sample not just the views of early adopters but also those of people who have not yet engaged with e-Infrastructure services so we can understand the factors underlying this. In addition, the information gathered from academic end users needs to be contrasted with the views held by service providers and technology providers as well as intermediaries such as application developers, e-Science centres and academic hosting institutions.

The first step in our research has been to review the existing literature on uptake of e-Infrastructure services in research and to analyse existing data collected as part of previous and ongoing activities of the project partners. This work has resulted in a list of barriers identified in the literature as well as an initial list of enablers and candidate interventions to address these. These findings are organised through a typology covering and further detailing the various dimensions identified above.

Identifying Respondents

As it is practically impossible to define a definitive set of respondents *a priori*, we are using an iterative approach starting with an initial set of interviews (complemented by a short questionnaire to cover basic information) of researchers with some experience with the use of e-Infrastructures in their research. Candidate respondents were identified using a combination of web searches, use of existing databases such as the UK research councils' 'grants on the web' databases and web mining. We found that it was relatively easy to compile long lists of candidates but that filtering them using our selection criteria (active research and using at least one of the JISC-funded services) involved a large amount of manual work to compile the required information from publicly available data.

Over the last three months we have conducted 49 interviews with UK-based researchers across a wide range of discipline areas, yielding about 25 hours of recorded audio. Before the interview, respondents were asked to fill in a short questionnaire so that the interviewers would have some baseline information to guide the interview process. Our sampling was stratified using the division of funding in the UK into research councils[1]. An overview of the discipline areas covered so far (based on respondents' self-classification in the questionnaire) can be found in Figure 1.

[1] The classification before the interview was based on reviews of information easily obtainable from websites and may not always coincide with respondents' self-classification in the questionnaire.

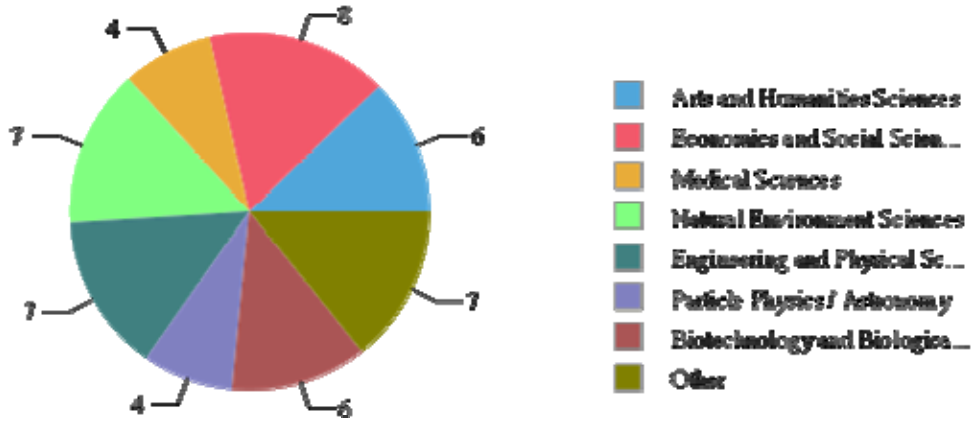

Figure 1: Respondents by Discipline

In addition to the primary stratification by research disciplines, we are also aware of the fact that other dimensions will be relevant as they influence the kinds of barriers that researchers face and the way they react to them. For example, researchers at different stages of their careers may have different interests, attitudes towards technological innovation, skills as well as investments in standard methods and tools. Consequently, we are aiming to ensure that our sample includes respondents of different levels of seniority.

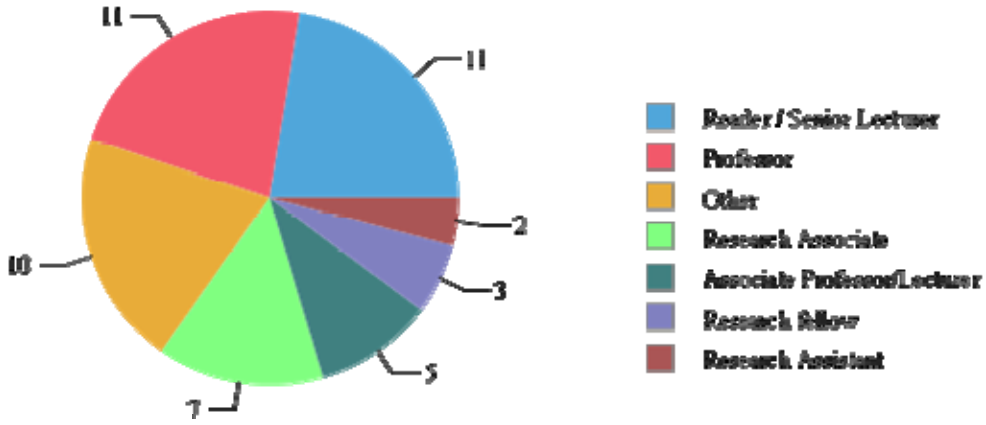

Figure 2: Respondents by position

The resulting preliminary analysis of the data will feed into the process of identifying further respondents for further waves of interviews using a theoretical sampling approach (*cf.* Strauss 1987) and a combination of web mining based on existing data gathered and snowball sampling where interviewees are able to suggest colleagues who we might approach. As Figure 2 shows, research assistants and fellows are currently relatively underrepresented in our sample (although some appear in the 'other' category) compared to more senior respondents. Respondents were, on average, aware of 68% of the services included in the questionnaire and were, on average, users of 21% of them. 75% were users of at least one service. These figures reflect the fact that the first wave of interviews has focused on early adopters and members of the e-Research community. We are going to address this issue in our subsequent waves of interviews to obtain an overall balanced sample. In addition, we will use theoretical sampling in order to verify or falsify emerging explanations of adoption processes, barriers and enablers.

Developing the Analytical Approach

The data from the first wave of interviews is currently being transcribed in preparation for analysis. There are a number of possible approaches for analysing the interview data and the choice must take into account various factors, including characteristics of the data, tools and time constraints.

One of the analytical approaches we are investigating involves marking up the transcripts using a coding scheme developed on the basis of our earlier literature review. A coding and presentation framework has been developed that allows us to manage the wealth of information produced and collated by the project and make it accessible for the project itself, for potential consumers such as service providers wishing to develop their own service provision and community engagement activities, and for researchers in this field. We have used a grounded approach (*cf*. Strauss 1987, Charmaz 2007) that develops the typology from the list of findings rather than imposing a set of pre-conceived categories onto them. As we add to our list of findings and build our analysis, we will continuously review the coding scheme and refactor it where necessary. While we do not currently envisage that this process will result in a formalised ontology, the approach taken is based on a set of XML applications that do allow us to express in an extensible way relationships of interest, such as those between barriers and enablers (E addresses B), between enablers (E1 is a pre-condition for E2) or between enablers and other entities (E is provided by organisation X).

Figure 3: Excerpt from the Coding Scheme

The coding is hierarchical to allow coding at a number of different levels of granularity and to make the coding scheme more manageable (there are 166 different codes at the moment). At the top level of the hierarchy, we distinguish between: stakeholders, existing services,

stages in the research lifecycle, research domains, social issues, technical issues and issues related to data. Further codes allow us to flag descriptions of research practices, usage of e-Infrastructure services, training or statements about specific requirements. Each of these categories is further refined to different degrees and we have developed tools to produce coding sheets as well as graphical representations of the hierarchy.

Method and Tool Usage

When dealing with large corpora of qualitative data, the use of electronic tools to manage, index, code and analyse data is essential so that the amount of information does not overwhelm the capacity of the analyst. Standard social science tools such as Computer Assisted Qualitative Data Analysis (CAQDAS) packages (Lewins and Silver 2007) offer a range of functions for this purpose and they generally support a range of different analytical styles. In principle, such tools would be adequate choices for analysing the data collected by the e-Uptake project. However, they suffer from a number of limitations that caused us to consider alternative approaches. Two main aspects are their use of largely proprietary file formats (which might inhibit the linking of the corpus to other sources and building on the corpus in the future) and lack of effective support for collaborative editing (which represents a potentially major handicap for a distributed research team).

As Fielding (2003) points out, e-Social Science holds great promise for qualitative research, both in terms of breaking new ground methodologically and in terms of overcoming limitations of current tools. While CAQDAS packages provide roughly similar baseline functionality, there are significant differences when it some to the handling of audio and video data, metadata, quantitative data as well as the analysis of these within the tools or the export of data for further analysis. As Fielding (*ibid*.) puts it: "no software contains all these features, nor are there independent evaluations of them. [e-Social Science offers] the opportunity to test the procedures, combine them, and apply them to larger data sets […] CAQDAS may dominate current approaches to computer-supported text analysis but XML and HTML applications increasingly offer similar functionality and can do so in a Web environment." Aspects that we are particularly interested in are:

- collaborative annotation and coding of qualitative data,
- integration of qualitative, quantitative and meta-data,
- complex queries of the coded data beyond the simple reports provided by CAQDAS tools,
- semi-automatic markup, meta-data generation and anonymisation using advanced analytical tools such as text mining tools,
- dynamic online presentation of the data in a number of different forms for different purposes and stakeholders and
- archiving in non-proprietary formats and subsequent reuse of the data and analytical tools.

We are currently exploring the use of an XML-based representation for interview data developed in the SQUAD project (Cummings 2006; Milosavljevic, Grover and Corti 2007). The project has developed schemas based on TEI (www.tei-c.org) to provide content markup for transcriptions of interview data. Using a standard XML application[2] such as TEI as a basis has a number of important advantages: the underlying schema is mature and supported by a significant community and tools developed for TEI coded data as well as generic XML tools can be applied. For example, before applying text analysis or text mining tools, we might want to extract the utterances made by respondents while ignoring those made by the interviewer. This can be achieved using a simple XML transformation.

---

[2] The term 'XML application' does not refer to an executable program but to an XML-based format to represent specific kinds of data, together with any associated schemas or tools.

Figure 4 shows an extract of interview data marked up with tags defined in the e-Uptake coding scheme. It shows that the coding covers a number of different aspects and that important information can be derived from the relationship between codes. For example, we may want to ask of our data what barriers are mentioned in the context of discussions of a particular service and the coding gives us a way to extract this information automatically based on the proximity of codes[3]. Drawing on the questionnaire data, we might also want to ask what barriers are mentioned by researchers from particular disciplines.

The metadata contained in the TEI header consists of information about the participants in the interview (interviewers and respondents), the consent given by respondents (to be approached again, to data sharing between the community engagement projects and to be quoted verbatim) as well as details of the interview itself such as modality (face-to-face or telephone) and the date. In addition, we have incorporated a simple extension mechanism for including arbitrary XML data which we use for the results of the questionnaire.

```
<u who="#interviewer"> What are the projects you are working on at the moment?</u>
<u who="#respondent">
    <seg type="DescOfWork">We are working on a distributed database with the [anonymised], we don't really
        have a name for it, but what we have done is distributed [anonymised] terabytes of data between the
        NGS and the university of [anonymised], and we have been using <seg type="Software">OGSA-DAI</seg>
        and <seg type="Software" subtype="OGSA-DQP">DQP</seg> which is the distributed query programme (or
        protocol, cannot remember what P is stand for) to run distributed queries between these two
        databases, one in Manchester and one in [anonymised]. The other project is much more computational
        problem, we have been developing a particular computational and statistical tools for doing very
        fast analysis on large datasets, so it could be called data -mining, but we know what we are looking
        for but we are interested in developing algorithm that are very quick.</seg>
</u>
<u who="#interviewer"> The first project where you used NGS?</u>
<u who="#respondent"> Well, overall it wasn't easy using them. There are several technical problems in that
    it was quite difficult, because we were quite a <seg type="Barrier" subtype="ResourcesAvailable"> large
        user, we wanted two terabytes of disc space on NGS</seg> and we wanted a <seg type="Barrier"
        subtype="SpecialisedSoftware">particular software, for example, we wanted to ingest the data into
        the oracle database and we wanted to use <seg type="Software">OGSA-DAI</seg></seg>. That required
    significant work on their part and also significant work on our part, so what happened was, it took
    months for us to work with <seg type="Service">NGS</seg> people to get to the point where we could give
    them the data and put it into the oracle database. Also one of the biggest problem we had was <seg
        type="Barrier" subtype="Performance">transferring the data to Manchester because of certain web
        protocols it was incredibly slow, it took us months to take the 2 terabytes into Manchester.</seg>
</u>
```

Figure 4: Marked up interview data

Our fieldwork is still ongoing and the sample is currently somewhat biased towards the early adopters and members of the e-Research community. This naturally limits the extend to which we can draw conclusions from the data collected so far, as well as the nature of findings presented. In the following section we will therefore focus on training requirements as we have access to a data collection that is more mature than the interview data collected and available for analysis to date.

Awareness Raising, Education and Training

We have conducted an analysis of data gathered in a series of semi-structured face-to-face interviews with attendees at a set of UK and international e-Research events, complemented by email follow-up from training events and an online survey. The analysis draws on fieldwork conducted since 2004 supported by a number of different projects[4] that have used the data on a year to year basis to direct the planning of their training provision. Respondents were asked to estimate their personal training requirements but also to extend these to their research group where possible. To date about 250 responses have been gathered.

---

[3] Of course, this is just a heuristic and there will be a need for manual checking.

[4] EGEE, EGEE II, OMII-Europe, ICEAGE, the PPARC training project and e-Uptake.

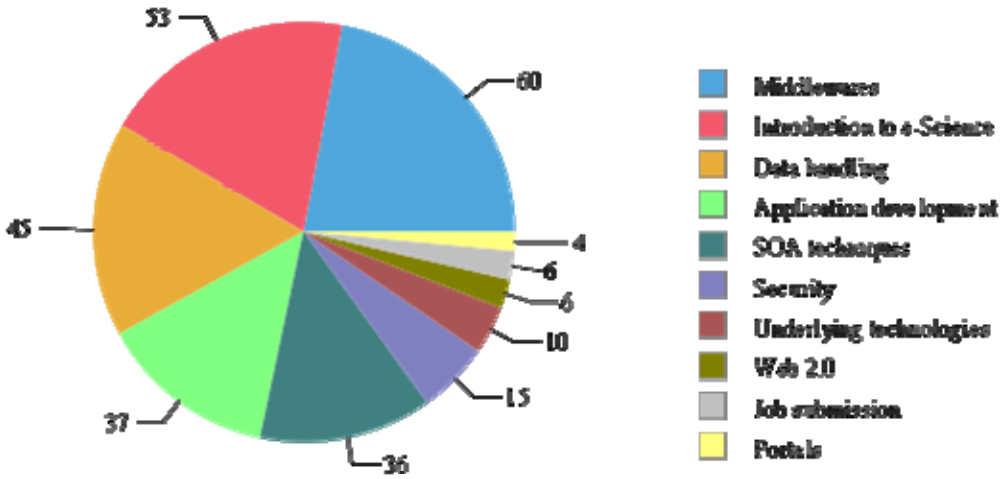

Figure 5: Training Requirements

The results (see Figure 5) indicate that there is a requirement for education about the overall concepts of e-Research, its provision and the cost/benefit balance for groups considering the adoption of e-Infrastructures and e-Research approaches. This trend is present both in the UK and internationally. This view is confirmed by our work on barriers to adoption – there are clear signs of a lack of awareness of e-Research methods, pathways to adoption and long-term implications. This view is further supported by evidence that courses offered are either not heavily subscribed or that the interest comes predominantly from early adopter disciplines. This is true both for specific training programmes and for education programmes. This suggests that a vicious cycle of lack of awareness exists. A related observation is that knowledge of e-Research is clustered around established centres of excellence.

Of the 49 respondents in our current sample in e-Uptake, less than half (43%) stated that they had received training on topics relating to advanced IT. While 81.63% said that their institutions were providing support for advanced IT (12.24% did not know and 6.12% responded with no), the detailed responses show that the level and kind of support available (or known to the respondents) can vary widely. Previous work, such as the SUPER (Newhouse *et al*. 2007a) and AVROSS studies (Barjak *et al*. 2007, 2008), have found that there are still gaps in the effective provision of technical information, consultancy and training. Newhouse *et al*. (2007a) note that there is a "lack of information about what is truly available and what tools might be better suited to a particular use case". Based on the interviews conducted by the e-Uptake project to date, we can say that similar issues have been raised by our respondents.

These findings also need to be seen against the background of an ICT skills shortage that affects e-Research activities as much as other areas of the economy (cf. Atkinson *et al.* 2008). Overall, it seems fair to say that while some level of support is available to researchers, it is not provided in a way that is consistent and reliable enough to help the interested through the initial hurdles of uptake and that there is a lack of sufficient background knowledge that compounds this problem.

Consequently, we have started to develop interventions targeted at the wider lifecycle of e-Research activities. The development of a UK 'one-stop-shop', led by the National e-Science Centre, will help to tie together the various interventions that can help to inform decision making, project formation and initiation, development as well as deployment, operation and usage of services and applications. The model for uptake we work with is illustrated in Figure 6. It should be noted that the model presented is only one of many

possible ways to depict the uptake process and that other formulations of similar virtuous cycles existed before (and have been found to be helpful).

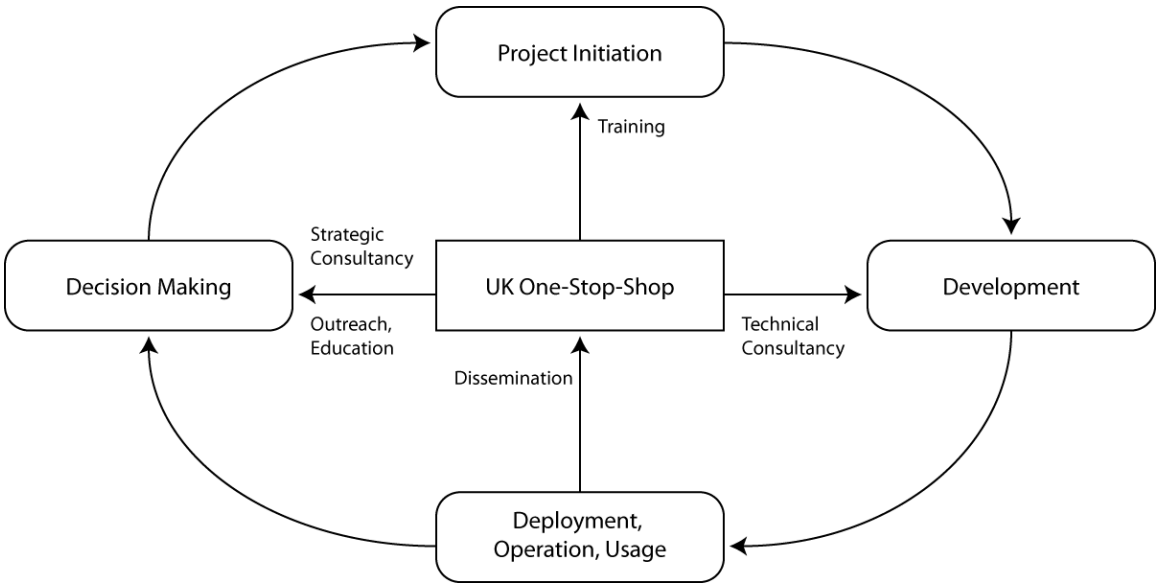

Figure 6: Uptake virtuous cycle

Based on this model, we have started to develop training and consultancy events that aim to take researchers through parts of this cycle. A first set of three related events aimed at widening the uptake of OGSA-DAI in the social sciences and the arts and humanities has confirmed the value of events that link the development of e-Research project ideas to specific training and consultancy supporting development and deployment. Provision of the various elements of support in the uptake cycle is still relatively patchy but important initiatives are underway to help address this issue. For example, in order to break the vicious cycle of lack of awareness, JISC is funding a programme of roadshows, carried out by the National e-Science Centre's Training Education and Outreach team. These roadshows will introduce concepts of e-Science and will be run at a range of different institutions, in particular those that do not already host significant e-Science activities.

Discussion

It is clear that our preliminary findings resonate with some of those of earlier studies such as SUPER (Newhouse *et al*. 2007a, 2007b) and AVROSS (Barjak *et al*. 2007, 2008). The former, which was a cross-disciplinary study of UK e-Science, identified the need to support researchers through training. The latter study, which investigated factors influencing the adoption of e-Infrastructure by European social scientists, reported, *inter alia*, issues related to the lack of qualified staff and funding. The breadth and depth of the e-Uptake study, informed, as it is, by these earlier studies and an extensive literature review, is such that it is likely to confirm and elaborate on these findings but also to uncover issues which have not previously come to light.

The uptake of e-Infrastructure in UK is in the state of flux. Barriers get overcome and new and perhaps more complex barriers appear. What the e-Uptake project does is to capture and document those barriers that have been overcome as well as suggest enablers for those barriers that have not been overcome yet. Furthermore, it identifies barriers that are hidden and have not yet been discussed or recognised. Based on this, we develop interventions in close collaboration with ENGAGE and eIUS, the two other projects in the JISC community engagement strand whose activities complement those of the e-Uptake project. eIUS aims to

document pathways to adoption and good practice in the use of e-Infrastructures in research through the development of use cases and service usage models (Voss *et al*. 2007). The ENGAGE initiative seeks to build on the SUPER report, further data collection as well as input from e-Uptake and eIUS to disseminate a set of best practice information through its website (at www.engage.ac.uk) and to promote ongoing discussion, leading to the development of domain specific systems or gateways in collaboration with scientific communities. Together, the three projects form a significant and coordinated effort to promote the greater engagement of researchers in the UK with e-Infrastructure services.

A common framework of understanding has been established between them that allows for coordination of activities, a common consenting process, data sharing (where consent is given by the respondents) and a common approach to analysis and dissemination. One important element of an effective community engagement process is a 'triage' system whereby contact can be established between interested researchers and other parties such as infrastructure providers, application developers, computer science researchers and intermediaries such as e-Research centres or OMII-UK. Linking the different elements of engagement together in an effective way will be of crucial importance for the success of e-Research endeavours and programmes.

Conclusions

In the e-Uptake project and in collaboration with our partners in ENGAGE and eIUS, we have begun to develop the most comprehensive and consistent programme of community engagement and promotion of e-Research to date, including the collection of an unprecedented body of evidence, rigorously analysed and presented in a way that will be useful to all major stakeholders. Based on this evidence, we are developing interventions that will help to build a virtuous cycle of uptake that will widen the usage of e-Infrastructures beyond the current user-base.

Acknowledgments

The research reported in this paper has been funded by the UK Joint Information Systems Committee (JISC) under the community engagement strand of its e-Infrastructure programme. We would like to thank the participants of the workshop on Fostering e-Infrastructure: from User-Designer Relations to Community Engagement, especially Ross Gardner for his comments on our formulation of the virtuous cycle of adoption and all the participants in our interviews. Finally, we would like to thank the reviewers for their helpful comments on earlier versions of the paper.

References

Atkinson, M., van der Meer, E., Fergusson, D. and Artacho, M. (2008). *e-IRG Education and Training Task Force Draft Report*. Long Report Version 4, 11th March. Available at: http://www.iceage-eu.org/v2/documents/ETTF%20Long%20Report%20draftMarch08.doc (accessed 28th May 2008).

Barjak, F., Weigand, G., Lane, J., Kertcher, Z., Poschen, M., Procter, R.and Robinson, S. (2007). Accelerating Transition to Virtual Research Organization in Social Science (AVROSS): First Results from a Survey of e-Infrastructure Adopters. In *Proceedings of e-Social Science '07*. Ann Arbor, October.

Barjak, F., Lane, J., Kertcher, Z., Poschen, M., Procter, R. and Robinson, S. (2008). Case Studies of e-Infrastructure Adoption. *Proceedings of e-Social Science'08*.

Charmaz, K. (2008). Reconstructing Grounded Theory. Chapter 27 in Akasyytari, P., Bickamn, L. and Brannen, J. (eds.) *The SAGE Handbook of Social Research Methods*. Sage Publications Ltd.

Cummings, J. (2006). *Smart Qualitative Data: Methods and Community Tools for Data Mark-Up (SQUAD): ESDS Qualidata project recommendations*. Available at: http://quads.esds.ac.uk/projects/squad_results.asp

Daw, M., Procter, R., Lin, Y., Hewitt, T., Jie, W., Voss, A., Baird, K., Turner, A., Birkin, M., Miller, K., Dutton, W., Jirotka, M., Schroeder, R., de la Flor, G., Edwards, P., Allan, R., Yang, X. and Crouchley, R. (2007). Developing an e-Infrastructure for Social Science. In *Proceedings of e-Social Science '07*. Ann Arbor, October.

Fielding, N.G. (2003). *Qualitative research and E-Social Science: appraising the potential*. University of Surrey.

Lewins, A. and Silver, C. (2007). *Using Software in Qualitative Research: A Step-by-Step Guide*. Sage Publications.

Milosavljevic, M., Grover, C. and Corti, L. (2007). Smart Qualitative Data (SQUAD): Information Extraction in a Large Document Archive. *Proceedings of RIAO 2007*, Pittsburgh, PA, USA.

Newhouse, S., Schopf, J., Richards, A. and Atkinson, M. (2007a). Study of User Priorities for e-Infrastructure for e-Research (SUPER). In proceedings for the UK e-Science All Hands Meeting, September.

Newhouse, S., Schopf, J., Richards, A. and Atkinson, M. (2007b). *Study of User Priorities for e-Infrastructure for e-Research (SUPER).* UK e-Science Technical Report UKeS-2007-01.

Strauss, A. L. (1987). *Qualitative Analysis for Social Scientists*. Cambridge University Press.

Voss, A., Mascord, M., Argüello Castelleiro, M., Asgari-Targhi, M., Procter, R., Fraser, M., Jirotka, M., Halfpenny, P., Fergusson, D., Atkinson, M., Dunn, S., Blanke, T., Hughes, L. and Anderson, S. (2007). e-Infrastructure Development and Community Engagement. *Proceedings of e-Social Science '07*. Ann Arbor, October.