The International Journal of Digital Curation
Issue 1, Volume 2 | 2007

User Priorities for Data: Results from SUPER

Jennifer M. Schopf,
Argonne National Laboratory,
UK National eScience Centre

Steven Newhouse,
University Southampton

July 2007

Summary

SUPER, a Study of User Priorities for e-infrastructure for Research, was a six-month effort funded by the UK e-Science Core Programme and JISC. Its aim was to inform investment in order to provide a usable, useful, and accessible e-infrastructure for all researchers and a coherent set of e-infrastructure services that would increase usage by at least a factor of ten by 2010. Through a series of unstructured face-to-face interviews with over 45 participants from 30 different projects, an online survey, together with a day-long workshop at NeSC, we have observed recurring issues relating to the provision of e-infrastructure. In this article we focus on the data-related issues identified during these interactions. We conclude with a prioritised list of future activities for research, development, and adoption in the data space.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

SUPER, a Study of User Priorities for e-infrastructure for Research, was a six-month effort funded by the UK e-Science Core Programme and JISC to inform investment in order to:

- provide a usable, useful, and accessible e-infrastructure for researchers across a wide variety of disciplines;
- integrate existing and new e-infrastructure facilities and services into a coherent whole; and
- provide the basis to increase the use of the existing e-infrastructures by a factor greater than ten by 2010.

E-infrastructure encompasses the facilities and services that support more advanced or effective research through support of access to information, data, distributed collaboration and computation resources. Since multidisciplinary teams often span national as well as institutional boundaries, support for international collaboration and consistent provision across these resources must be considered. Inevitably, different organisations will be engaged in constructing and operating an e-infrastructure; therefore the recommendations must permit autonomous providers to collaborate in order to provide coherent and consistent facilities.

The early vision of Grids – running applications on remote, shared distributed resources ranging from dedicated clusters to shared-use, cycle-stealing desktop machines – is now a reality in many communities. The availability and accessibility of these resources, as a side-effect of the UK e-Science programme, have brought new user communities to larger-scale science than was previously feasible and have even promoted paradigm-shifting science techniques in some domains, such as large-scale ensemble and parametric studies.

In this article we focus on the data-related requirements that emerged from these interactions. More information relating to issues other than data and input from other sources can be found in the SUPER technical report (Newhouse, Schopf, Richards, & Atkinson, 2007).

Community Inputs

As a first step in carrying out the SUPER Project, we spent several months in late 2006 meeting with approximately 30 groups across the UK who represent a cross-section of the research community and which were engaged in a variety of roles and projects. The work continued with two additional data sources, a day-long workshop and an online survey.

The groups interviewed included:

- current and potential end-users of e-infrastructure: generally those conducting research, design, analysis or diagnosis in projects funded by one of the UK Research Councils;
- technologists: or developers who take generic middleware and adapt it for a specific use case;
- generic tool developers: those who were building solutions that could be used in many application domains;
- and service providers, including universities and publicly funded research institutes that are setting up hardware and services.

Interviews

The face-to-face meetings lasted anything from half an hour to half a day and covered a wide variety of topics, concentrating on current use and needs. We considered performing structured interviews based on a standard questionnaire, but differences in the background and knowledge of those interviewed made this approach ineffective. Instead, we asked what functionality the groups had tried in the past, what their applications needed today from the current Grid infrastructures, and what functionality the groups were considering in the short term. Most meetings concluded with our inquiring into what functionality the group thought was most important but which was deficient in today's tools or services.

The interviews were conducted in the following institutions: Newcastle, Glasgow, Edinburgh, Oxford, Cambridge, UCL, and Reading, covering 45 people from over 30 projects. A third of the projects were funded by EPSRC and a further third by BBSRC, MRC, and JISC. The remaining projects were funded by DTI, EU, Wellcome, AHRC, ESRC, NERC, and PPARC, with a significant number of campus activities being funded directly by the local university. A full list is available (Newhouse, Schopf, Richards, & Atkinson, 2007, Appendix A).

Workshop

In support of our interviews and this report, a workshop was held at the UK National e-Science Centre on February 16, 2007, to discuss the draft report and offer clarifications, which have been included in the main body of the report. The participants at this workshop were principally policy makers and project managers, e-Science strategists, and e-Infrastructure providers. This contrasts with the focus on direct users which was chosen for the interviews.

The workshop itself consisted of an introductory discussion of the SUPER interviews and report, followed by breakout sessions to discuss in more detail the data, VO management, and support. Each breakout session was asked to consider the following points:

- Who are the early adopter/active communities?
- How uniform are the requirements within the community? Are there gaps? Do we need to revise the emphasis?
- What are the targets over the next 12 months and then for the longer term?

Talks and summary slides are available online[1].

Online Survey

We posted a Zoomerang survey online from December 2006 to March 2007. It was advertised over a dozen UK e-Science mailing lists and newsletters in order to reach members of the community we could not interview in person. We received between 24 and 26 responses for a large set of questions.

Details for the complete survey results can be found in the final SUPER report (Newhouse, Schopf, Richards, & Atkinson, 2007).

[1] http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=743

E-Infrastructure Today

At a high level, today's e-infrastructure users go well beyond the risk takers who started work in this area; and as the end-user community expands, there arises a need for *broad outreach and evangelising* of the benefits provided by adopting the emerging campus, national and international e-infrastructure deployments. Many new project participants simply do not know what is possible or probable; and many new communities do not know what is easy and available as opposed to difficult, or even an open research question. There is no baseline of capabilities agreed upon in the broader community, and this situation is significantly hampering additional uptake.

In addition to promoting the adoption of available software and services, *training* needs to be provided for the different stakeholders within a project: the end-users who will be using the e-infrastructure services through the tools and applications; the developers (both of generic and domain-specific services) who will be using the deployed e-infrastructure services to build other services and tools; and the deployers (in both a local and national context) who will have the responsibility for managing the required e-infrastructure. Training materials are needed in many forms: formal instructor-led courses, standalone self-help material, worked examples, reference systems, and so forth. Ironically, many infrastructure projects noted the need for training and had funds from organisations such as the National Grid Service (NGS) or the National e-Science Centre (NeSC) but were ignorant of existing training materials (from software providers) or ongoing courses. As a consequence we observed considerable duplication of activity.

Both of these issues identified in the broad sense during the interviews were brought to the forefront at the workshop, where several work items were identified as essential first steps to making progress in this area.

First among these was the need for an evaluation of commonly available network file systems (GPFS, PVFS, etc.) in comparison to distributed file management tools (SRB, SRM, dCache, etc.). Such an evaluation should include an assessment of the criteria, for example, collaboration, deployment, ease of use, and cost. NGS/ETF was identified as a group that could follow up on this area and then host a work shop.

A question was raised about why common tools available in the digital repositories space (e.g., from the Digital Curation Centre, (DCC)[2]) were not in common use by e-scientists. This question will be followed up by the DCC group.

Several repository systems had already undergone evaluation, but the results were not broadly known to the community. It was suggested the JISC might fund an effort here to make the previous evaluations and ongoing project information more widely available.

Several participants were also concerned that current open source or commercial solutions for distributed file systems were not in more common use. The ETF said it might be able to follow up this evaluation.

[2] Digital Curation Centre http://www.dcc.ac.uk/

Sharing Data

By far the largest concern of the users with whom we spoke centred on the use of large-scale data stored in structured file sets, i.e. *how to share data with colleagues* within their project or their wider community. Of particular importance is the management of data stored in files, whether software, results or other data, as opposed to data stored in databases. Users are concerned about the data's long-term storage and curation, about means of accessing these files from a local desktop, and about seamless data transfer to remote resources for further analysis. End-users also want to be able to annotate the files with metadata about the contents and provenance, in order to support search and reanalysis at a later date. *Metadata is the key to being able to share the results*. Many of the groups have explored use of the Storage Resource Broker; however, this is seen as a heavyweight solution that is hard for a project to deploy and maintain.

To suppport the easy curation and annotation of data files, additional tools are needed to *autogenerate metadata* about the data and how, where, and by what means those data were generated, i.e. their *provenance*. Once provenance data are collected, there will also be a requirement to navigate and analyse such data. It was noted that if users are responsible for the annotation of their data, the task is generally left undone.

A separate issue concerns how the data are annotated. Many groups have begun to create their own metadata after failing to find acceptable community standards. Standards exist for common basic properties, sometimes many in number (for example, timestamps), and users were generally happy with the existing frameworks for this higher-level information. Lower-level and domain-specific metadata, however, have yet to be standardised for many communities. The quality of such annotations is often uneven, however, and some communities therefore have started developing common metadata schemas to drive annotation standards. Automated collection of basic metadata is seen as a large step forward from current practice; but for some domains, specialists still may be required to do the full job by hand; almost always, some human input is required (e.g., the purpose of an experiment).

These concerns were echoed at the workshop. It was recommended that JISC put out a call to create a best practice document for current metadata and annotation practices and possible policies. It was noted that in order to contribute data to a common collection, a policy will be needed to identify how annotation and metadata creation will be performed, preferably employing standards which support interoperability.

In addition, it was suggested that the DCC could help document what standards were currently available for successful data curation, and follow this with a dissemination workshop.

Access to Data

Easier *access of data* was also an issue. Many groups now have to manage the file output from computations across multiple locations, including national resources such as the NGS, as well as campus and desktop resources. Researchers would like to access their local files seamlessly when running an application remotely, so that they can edit locally the input files that form the basis of the simulation. Likewise, the

output files residing on a remote resource need to be accessible for future processing and analysis on the local desktop or on other remote resources. This situation leads to requirements for the registration and discovery of files held in different locations.

This information from the interviews was also seen in the online survey. The two data services of greatest importance to the respondents were file transfer services and data access services. Other tools such as replication and provenance were listed as second-tier concerns.

Research has experienced a paradigm shift with respect to the changing uses of data as well as the changing use of compute resources. More groups now need to share data more widely under defined access policies, to merge data, and to retain that data for longer. There was an identified need for *standard policies and tools for data curation* across RCUK-funded projects. Those policies are currently not well defined, but they need to be – both for user roles and temporal constraints. For example, several groups mentioned a "guarded period of access," during which only a limited group of people could see the data, followed by wider access to the community after a set period, perhaps coinciding with the publication of results.

For some groups "wider access" implies being open to the community only, perhaps within a well-defined virtual organisation or set of virtual organisations; whereas for other groups it means being readable by anyone. Some groups have quite strict access limitations, including even what resources were able to host the data. Any access control structure and its underlying authentication mechanisms must be able to support both controlled access to any external collaborator and eventually unrestricted access to others. But in general, there is a shift towards much *longer-term storage of data*, some for pragmatic experimental use, and some at the behest of the funding agencies.

Recommendations

As a result of this work, we recommend investment in three broad areas: *software*, *policy* and *support*, with items listed in no particular order. We list only the data-related items here. Sustained investment in these areas will provide a set of structured tools, services and environments to support access to e-infrastructure, and a support infrastructure to enable the adoption of e-infrastructures by new user groups.

Software:

- Automatic data annotation and provenance tools to support domain-specific schema
- Mechanisms to support controlled and convenient sharing of files between groups
- Reliable documented software base to enable virtual organisations built around individuals to gain access to services and resources, and collaborative mechanisms to facilitate research between these individuals

Policy:

- Development of a best practice document to support research groups in developing their own data curation and file management policies
- Development of common annotation schemes for individual communities to support consistent metadata labelling within these communities

Support:

- Better technical consultancy to end-users who wish to move their applications to use e-infrastructure services, developers who wish to use best practice to build e-infrastructure services, and deployers who need to configure e-infrastructure services for maximum performance and reliability – in general, better information backed by a human. (This needs to be a funded, coordinated service with experts in support. Simple lists are worse than nothing, as they are always out of date and frequently misleading.)

We hope that these recommendations will influence the individual roadmaps and activities of organisations charged with supporting collaborative multidisciplinary science in the UK (e.g. OMII-UK, NGS, DCC) and their funding bodies – the UK Research Councils and JISC.

Acknowledgements

This work was supported in part by EPSRC, JISC, OMII-UK, and also by U.S. Department of Energy, under Contract DE-AC02-06CH11357.

We gratefully acknowledge all of the people we talked to; their time and frank discussions are appreciated. Malcolm Atkinson was fundamental in supporting this ongoing project, and Andrew Richards accompanied us for the interviews. In addition we received valuable feedback and comments from Dave Berry (NeSC), Richard Sinott (University of Glasgow), and David Wallom (University of Oxford).

References

Newhouse, S., Schopf, J.M., Richards, A., & Atkinson, M. (2007). *Study of user priorities for e-infrastructure for e-research (SUPER)* (UK eScience Technical Report No. UKeS-2007-01). University of Edinburgh. Retrieved July 6, 2007 from http://www.nesc.ac.uk/technical_papers/UKeS-2007-01.pdf