This work is distributed as a Discussion Paper by the

STANFORD INSTITUTE FOR ECONOMIC POLICY RESEARCH

SIEPR Discussion Paper No. 08-11

Mapping e-Science's Path in the Collaboration Space: Ontological Approach to Monitoring Infrastructure Development

By
Paul A. David
Stanford University

Matthijs den Besten
Oxford e-Research Centre

June 2008

Stanford Institute for Economic Policy Research
Stanford University
Stanford, CA 94305
(650) 725-1874

Mapping e-Science's Path in the Collaboration Space: An Ontological Approach to Monitoring Infrastructure Development

By

Matthijs den Besten[1] & **Paul A. David**[2]

[1]*Oxford e-Research Centre, University of Oxford:* matthijs.denbesten@oerc.ox.ac.uk

[2] *Stanford University:*pad@stanford.edu *; Chaire Innovation & Regulation (l'Ecole Polytechnique & Telecom-ParisTech):* http://wwwinnovation-regulation.eu *; UNU-Merit/Collaborative Creativity Group (Maastricht,NL):* http://www.merit.unu.edu

Version 3.3: 4 June 2008

Accepted for the Proceedings of the 4th International Conference on e-Social Science, held in Manchester, England on 19th June 2008

Abstract

In an undertaking such as the U.S. Cyberinfrastructure Initiative, or the UK e-science programme, which spans many years and comprises a great many projects funded by multiple agencies, it can be very difficult to keep tabs on what everyone is doing. But, it is not impossible. In this paper, we propose the construction of ontologies as a means of monitoring a research programme's portfolio of projects. In particular, we introduce the "virtual laboratory ontology" (VLO) and show how its application to e-Science yields a mapping of the distribution of projects in several dimensions of the "collaboration space." We sketch out a method to induce a project mapping from project descriptions and present the resulting map for the case at hand. What the map suggests is that the UK's e-Science programme so far has remained very "data-centric". Apart from methodological bias, two hypotheses could account for this focus: distributed databases are of central importance to a wide array of science and engineering fields, and tools for federation, annotation and facilitated access form a logical priority in middleware development (H1); there is a preference for organizing projects that involve dyadic interactions with a research facility that requires no intervening human agency, and an aversion to undertaking contracts for collaborative work among research groups that would transcend institutional or organizational boundaries (H2). Further studies that would make use of ancillary information, including interviews with principals in the formation of the UK's e-Science core program, will be needed to throw light on the validity of these or still other potential explanations. Be that as it may, this paper shows that the proposed mapping approach to be informative as well as feasible, and we expect that its further development can prove to be substantively useful for future work in *cyber-infrastructure-building*.

Introduction

Collaboration across groups has long been recognized as an important condition for research in science and engineering. Collaboration also is increasingly associated with research and development activities that lead to innovation (e.g. Carlile 2002). Moreover, Inasmuch as advances in information and communication technologies are seen to play a vital role as an enabler or facilitator of collaboration (e.g. Castells 1996), it is not surprising that research policy on both sides of the Atlantic has sought to promote the creation, deployment and diffusion of digital "collaboration technologies" (see David and Steinmueller 2003, David 2005). In the US, the NSF instigated programmes to develop Collaboratories, the Grid, and now the Cyberinfrastructure; the EU is pursuing e-Infrastructure; and in the UK, EPSRC coordinated the e-Science Programme and presently is working towards the establishment of a subsequent programme to enable the "digital economy" (see Finholt 2003; EPSRC 2007; NSF 2007). All these programmes consists of a multiplicity of distinct projects – sometimes well over a hundred – and, with a common infrastructure often stated as the programmes' primary goal, monitoring the projects to establish the coherence and completeness of a programme so as to be able to adjust where necessary would seem useful. In this paper, we propose the "virtual laboratory ontology" (VLO) as a tool that could facilitate this kind of project portfolio coordination – which could be hierarchically managed, or self-organized by providing participating sub-projects with access to intelligible information about the evolving configuration of the ensemble. In our exposition, we focus on the UK e-Science programme. Yet, the approach is generic enough that the approach thereby illustrated could be applicable to other programmes as well.

Portfolio Coordination in the UK e-Science Programme

Collaboration requires coordination. This truism holds not only at the level of projects (Swan and Scarborough (2005), but by extension also for the research programmes who maintain a portfolio of projects. Also within the UK e-Science programme, an effort to develop the next generation infrastructure for global collaboration in science (see http://www.research-councils.ac.uk/escience/), the need to map the activities of its projects was soon recognized (Hey et al. 2002). The programme was a collaboration of several research councils, which pursued the development of "middleware" through its "core" programme and tried to establish a "proof of concept" by means of pilot projects and adoption through the support of additional application-oriented projects. The "middleware" of the collaboration technology was conceived of as a common infrastructure that linked a wide variety of resources to an even wider variety of applications. Consequently, the programme's "roadmap" was intended to give everyone involved an overview of the components on which the various projects were working, so that people would know what pieces of the jigsaw they could reuse and what pieces they might usefully contribute. Today, the map that resulted from that exercise is still available (see http://www.research-councils.ac.uk/escience/), but it does not seem to be used as much as had originally envisaged – perhaps because the continued funding of the programme as a coherent entity appears in doubt.

Nevertheless, the mapping of projects in this field remains a fundamentally good idea. In much the same way that the gene ontology (Bada et al. 2004) gives genomics researchers (and funding agencies) an overview of what is currently known about genes and allows them to develop research questions on a more fully informed basis, a "virtual laboratory ontology" (VLO) could help developers of cyber-infrastructure determine where most work has been done and help identify existing gaps in the infrastructure and tool sets. In this way, research programmes could be in a position to benefit from the tools for "knowledge management" whose development they have already been promoting in other contexts.

The Virtual Laboratory Ontology

In the field of advanced computing initiatives the employment of ontologies has been proposed as a means to assist the management of the Grid – a core tool for e-Science. Those proposals concerned day-to-day resource allocation in the operation of that infrastructure (Goble and De Roure 2002; Tangmunarunkit, Decker and Kesselman 2003). The VLO approach advanced here has a more high-level purpose, however, that of supporting planning, funding, and monitoring functions in the management of major research and development initiatives for the middleware and applications software layers of the "cyber-infrastructure."

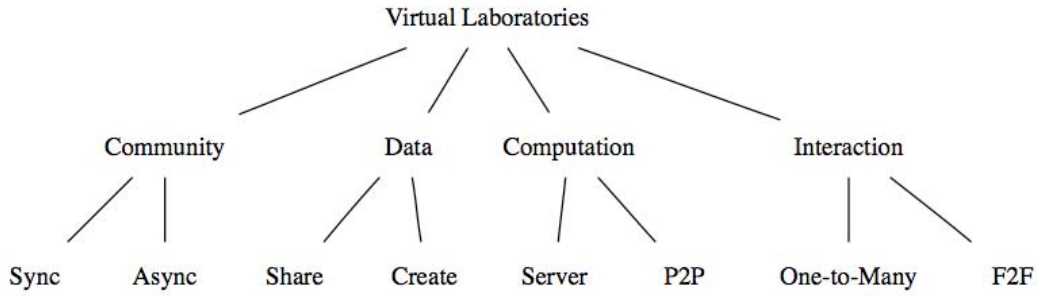

Fig. 1 : A rudimentary virtual laboratory ontology

The way in which the VLO could support a research programme is similar to the project mapping that was attempted in the e-Science programme, with a number of important differences. The specific VLO that will be utilized here has been built upon work that was first presented an early study of e-Science infrastructures for the Joint Information System Committee of the UK Research Councils (David and Spence, 2003: Appendix 2). Its first distinctive feature of the VLO is that it does not proceed from an explicit list of requirements for a collaboration infrastructure; instead, it asks what aspect of distributed research collaboration each of the projects in the portfolio is addressing (in line with Allen et al 2000; 2003). Secondly, our VLO promotes a tree-like representation of the project-space in which leaves represent projects and branches aspects of collaboration. Thirdly, this VLO approach (unlike the original e-Science "roadmap") does not call for self-reporting by the projects, or demand a consensus on terminology. Rather, it can be implemented on the basis of information already provided by principal investigators' descriptions of their proposed and ongoing activities.

Figure 1 shows a basic version of VLO. Utilizing information on e-Science projects that has been compiled by the National Centre for e-Science in Edinburgh, one can classify a small group of projects by hand, and then use standard tools for machine learning and statistics to obtain a map of the programme as a whole. Our claim is that this method could have been used to refocus efforts in the past and might be beneficial still for similar programmes in the future.

UK e-Science Projects

The National Centre for e-Science (NeSC) has compiled information on projects that benefited from the e-Science programme as part of the project mapping exercise that was undertaken by the e-Science programme. For each project, this dataset consists of the following elements:

A project description of about 200 words each;

Metadata about the project such its name, the investigators involved, application area and start- and end-date;

A list of middleware-components to which the project contributes.

There are numerous ways to “map” the information that NeSC has compiled, and a lot can be learned even without the help of ontologies. The figures on the next pages will serve to illustrate some of the many available forms of mapping (data display), while also hinting at the limits of both the data and these non-ontologically structured maps.

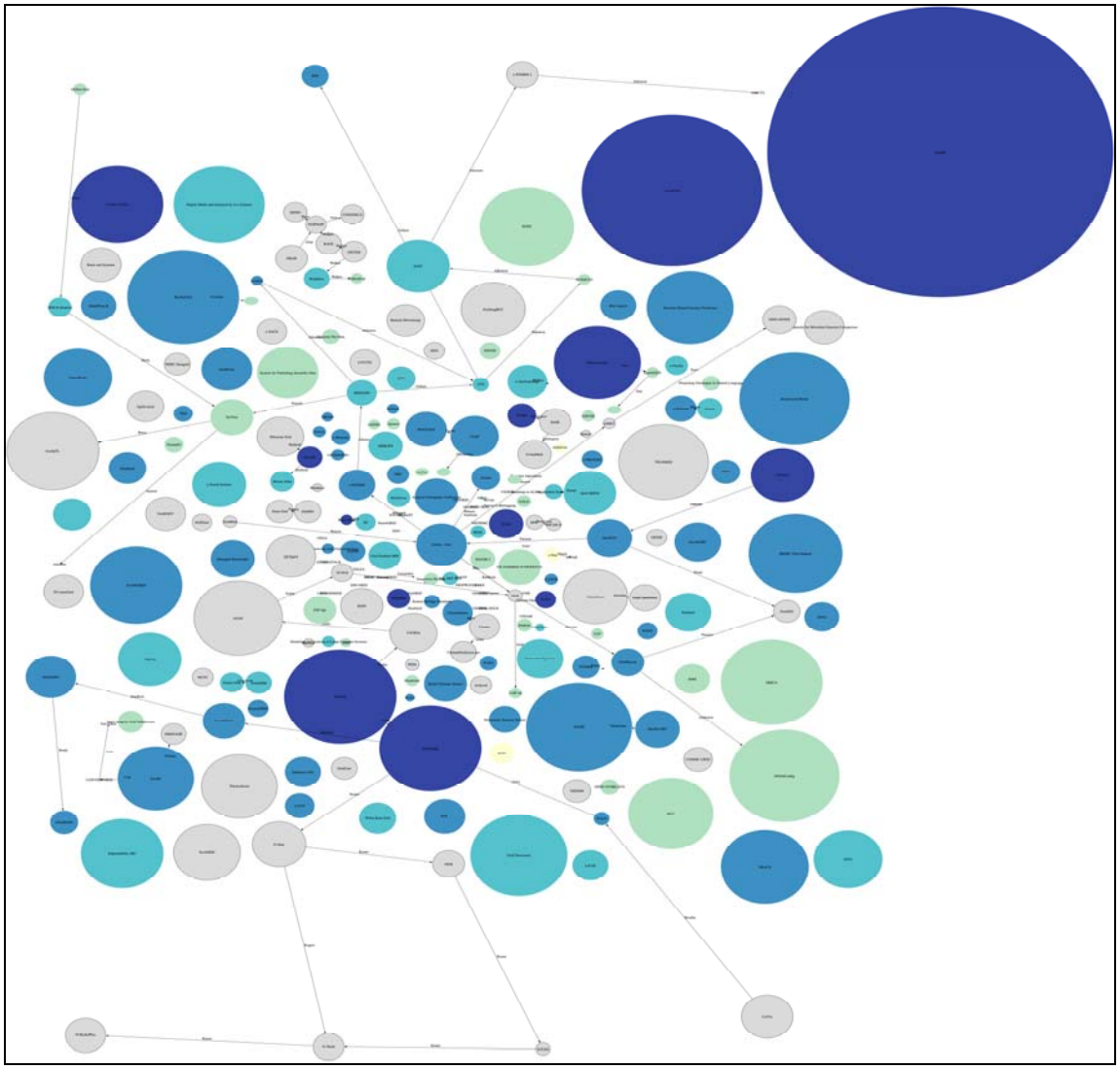

Fig. 2 : Investigator links and relative grant size of UK e-Science projects

Figure 2 is a map of project metadata. Each project is represented by a circle and the surface of this circle is proportional to the amount of funding that the project has received. The circles are colour-coded with dark colours representing an early start-date and light colours a more recent start-date. Gray indicates that no start-date information is present. Moreover, for each investigator who has been involved in multiple projects, a line links the subsequent projects that this investigator has been involved in. The circles are organized in a way that minimizes the amount of space and the positioning of the circles has no particular meaning apart from that.

Several features about the UK e-Science programme are exposed by Figure 2. First, there is considerable variety in the magnitudes of the grants awarded to projects, with a few very big

ones, like for instance GridPP, and many more small ones. Second, while most projects did not give rise to a follow-on projects, or, at least, not to a project that involve one or more of the initial investigators, there is a group of over 30 projects that are connected – due mainly to the efforts of two researchers: Carol Goble (subsequent connections from MyGrid and GEODISE to GOAT, Utopia and iXGRID, all but the last quite large projects), and Malcolm Atkinson (subsequent connections from GRIDPP, to DAIT, CFG, SCADEVO, GRIDNET, BRIDGES, e-STORM and e-STORM-2, and OGSA, after DAIT most of which were quite small). Third, the database is not perfect: There is no start-date or grant-size for quite a few projects and close inspection suggests that some projects, for instance ClimatePrediction.net, have registered multiple entries.

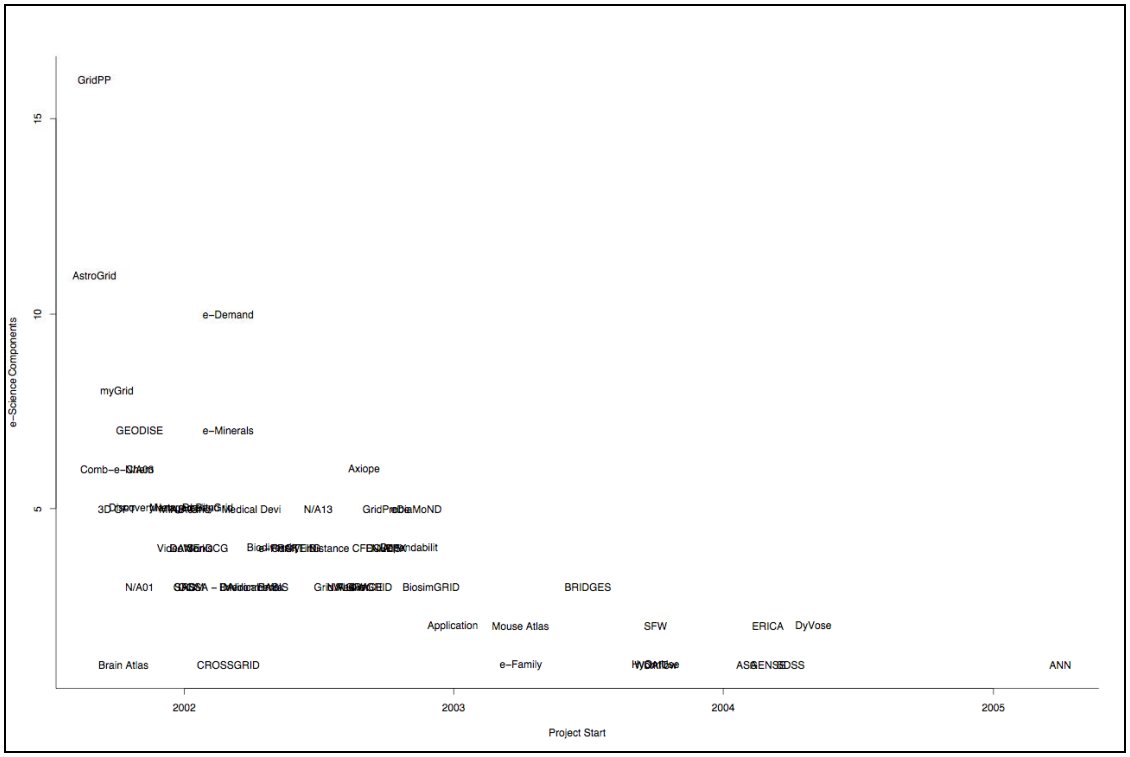

Fig. 3 : Number of e-science components promised by UK projects

Figure 3 plots the number of components that projects contribute to (according to the database) against time. It shows a decrease from as many as sixteen components in some projects during the programme's initial phase to only one or two in subsequent years. This might signal that e-Science Core program had been an astounding success, in the sense that projects were able to turn to other tasks that were not on the original list of infrastructure components. Alternatively, it might be a sign that those components were no longer seen as relevant, or it might simply reflect the overall decrease in the willingness of funding-agencies to collect all the metadata that NeSC desired. The latter, more pessimistic reading resonates with a practical consideration in the design of a VLO: it should not be too vulnerable to the unwillingness of others to conscientiously tick boxes or select keywords from a predefined lists whenever asked.

Like the "tag clouds" that have been appearing on web sites in recent years, it is possible to derive a picture of the e-Science projects on the basis of the project descriptions alone. Figure 4 shows the result of such an exercise. In order to obtain the figure, we listed all words employed in project descriptions and counted how frequently they were mentioned. From the 100 most mentioned, we removed stop-words such as "the" and "and". The remaining 40

words were then ordered alphabetically and resized in proportion to the number of projects in which they were mentioned.

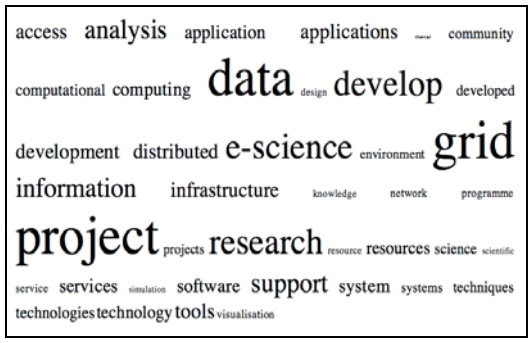

Figure 4: Top 40 most frequently mentioned terms in e-Science project descriptions (excluding stop-words)

Figure 4 conveys the message that many e-Science projects in the UK were concerned about "data" and the Grid, and comparative few were concerned with "community" or even with "collaboration." More detailed analysis would be needed to confirm whether or not that actually was the case, and whether that conclusion also holds if one considered not just the numbers of projects, but the relative magnitudes of the resources they deployed.

Towards a New Map of e-Science

In order to make statements about the activities of projects and the overall balance of a programme, it helps to have some kind of a priori notion as to what the array of its activities might include and where their focus might fall. To take the words of Dr. John Taylor, who as Director General of the Research Councils launched the Programme in 2000, e-Science is (or was to be) about "global collaboration in key areas of science and the next generation of infrastructure that will enable it". From that it would seem reasonable to have presumed that the activities of the projects were to be concerned directly or indirectly with the development of collaboration technologies. Hence, the VLO we have presented in rudimentary form, and which recognizes the variety of forms of research collaboration that might be enhanced by the e-Science programme, can be said to offer an appropriate template for mapping the portfolio of projects that emersed.

Obviously, the VLO is first and foremost a classification scheme. In the e-Science study by David and Spence (2003), a set of 23 e-Science pilot projects were classified using this scheme, based on close reading of the project descriptions collected by NeSC and the information presented on the projects' respective web-sites. This "expert analysis" yielded a distribution of projects in which 16 were qualified as data-centric, 4 as computation-centric, 2 as interaction-centric and 1 as community-centric (See appendix 2 of David and Spence 2003 for the details). In order to get an impression of the focus of the e-Science programme as a whole, one could repeat the exercise for all UK e-Science projects. That, however, would required a lot of work, not just for the classifiers, but also for those who want to scrutinize the resulting classification and the sensitivity of the overall assessment of the programme to variant classification decisions. Therefore, we have explored an alternative procedure, in which the a relatively small set of projects were classified (arduously) by hand, and, on the basis of the association between their respective classifications and various objective indicators that could also be readily obtained for all projects, a general machine-implemented classification routine could be devised.

We have taken the classification of e-Science pilot projects from David and Spence (2003) as the initial set for this exercise, and used it as a training-set for machine-learning and the

subsequent application of the resulting routine in classifying all the later UK e-Science projects.

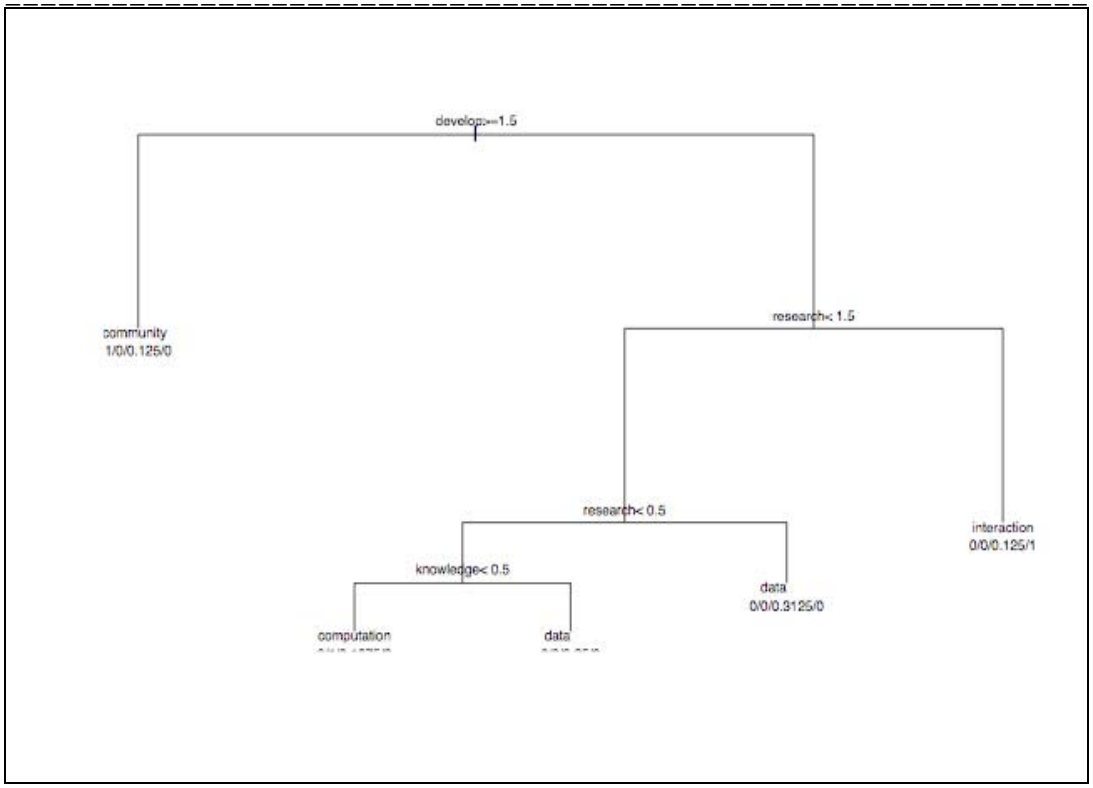

Fig. 5 : Regression tree to classify pilot projects
(Values at the leaves indicate weight of the four classes)

Figure 5 shows the regression tree that was derived by a machine-learning method known as "recursive partitioning" that was asked to predict the classification of the pilot projects on the basis of the number of time that these projects mentioned the words listed in Figure 4 in their project description. In order to arrive at the tree in Figure 5, weights were assigned to all 23 cases so that the prior likelihood for each of the four classes at the first level of VLO would be ¼. Without this, the default classification for all projects would be "data-centric." In addition, the requirement for the number of cases that must exist in a node in order for a split to be attempted was reduced from the default level of 20 to 15. Without this relaxation, the decision tree induced would have only one level predicting that if the word "research" occurs a project is likely to be community- or interaction-centric, whereas if "research" does not occur, it is likely to be computation- or data-centric. The regression tree in figure 4 is slightly more interesting than that. Still, it would seem a bit primitive to classify projects only on the basis of their use of the words "research", "develop", and "knowledge."

A drawback of using word-counts to characterize projects is that single words do not have much expressive power. In addition, words often co-occur and are seldom independent. To deal with both these drawbacks, we performed a principal components analysis on the matrix of word-counts for the top-100 most often mentioned words in all the project-descriptions collected by NeSC. The components returned by this analysis are composites of the word-counts that are mutually independent of each other. Recursive partitioning of the pilot projects on the basis of these components, using the settings noted above, yields a regression tree in which the projects are classified on the basis of their valuation in component 1 and 54.

Figure 6 on the next page plots each class of pilot projects along the axes of component 1 and 54; Figure 7 shows the weights or loadings that both components assign to the 40 non-stop-word words of the top 100 on which the analysis is based; and Figure 8 indicates how the whole portfolio of UK e-Science projects is classified using the decision tree that was derived in this manner.

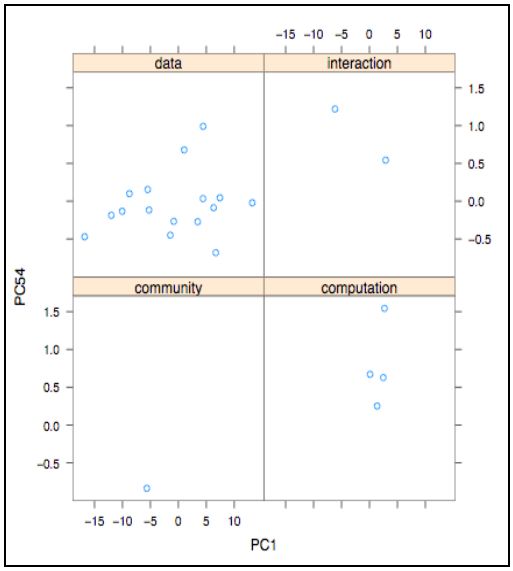

Fig. 6 : e-Science pilot projects mapped on principal components from project descriptions.

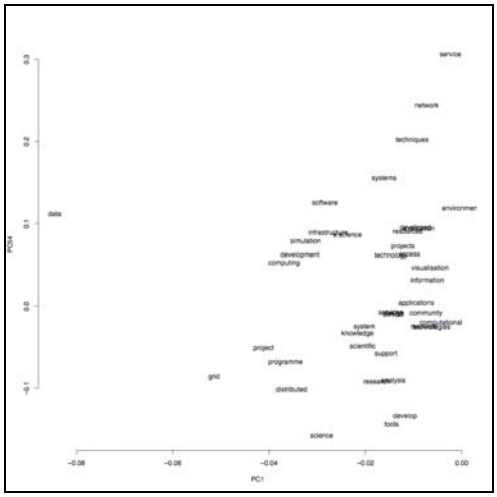

Fig. 7 : Weights given to words (variable loadings) for principal components 1 and 54.

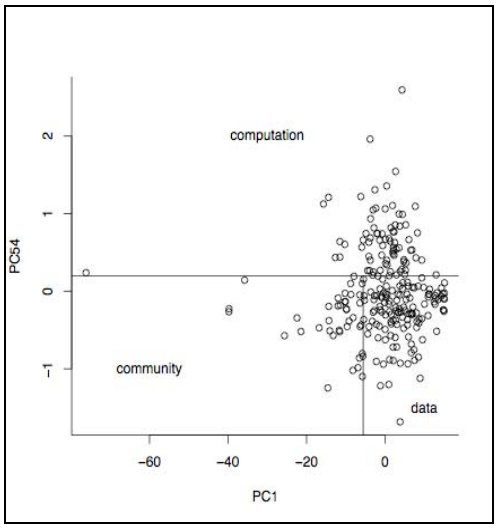

Fig. 8: Mapping of e-science projects on principal component axes 1 and 54.

Conclusion

The results presented in Figure 8 suggest that the design of the e-Science middleware infrastructure has been oriented primarily toward data-centric or computation-centric applications from its inception, and that this orientation has been pursued at the expense of developing other possible means of supporting the conduct of collaborative e-Science activities, namely those centred around the construction of multidisciplinary research communities, or access to shared large-scale facilities and instruments.

Such an orientation, or "bias" might be supposed to reflect the "the priorities of the e-Science programme" (Hey and Trefethen 2002). Yet, rather than jumping immediately to that inference, one should entertain the possibility that the emergence of a strong data-centric orientation reflected the acknowledgement of constraints arising from the difficulties and delays that were likely to attend the negotiation of inter-organisational agreements, e.g., among groups based in different universities or schools within one university, or between academic and industry research units. Thus, rather than being an expression of "top-down" priorities, the skewed distribution of the projects in the space described by the VLO may have been a response to practical, "bottom-up" considerations of the "delays and transactions costs" that could be expected in forging contractual agreements for the conduct of close and continuing collaborative interactions across organisational boundaries.

It has been argued (see David and Spence (2003, 2008), David (2006)) that considerations of that sort, unless addressed by the formation of flexible institutional arrangement the facilitate inter-organizational contracting, and remove existing institutionalized impediments to direct collaboration among distributed researchers, will most likely exert powerful forces shaping the evolution of the U.S. "Cyberinfrastrure" program and the EU's "e-Infrastructure." But the direction they would impart would perversely work to limit the realization of the available technology's potentially transformative contribution to enhancing the global conduct of scientific research.

Further work therefore is needed to examine collateral data from UK e-Science projects, and those of other kindred programmes, firstly to establish that a general data-centric orientation persists after differences in project size (both in terms of average flow rates of expenditure and durations) have been taken into account. Once that '"phenomenon" has been proved to be reasonably robust, sufficiently so to form a source of broader concern, further research

surely should turn to the task of identifying the source or sources of that “bias”. Two hypotheses woulde present themselves for initial examination in that regard, as either could account for the revealed data-centric focus: (H1) -- distributed databases are of central importance to a wide array of science and engineering fields, and tools for federation, annotation and facilitated access form a logical priority in middleware development; (H2) – there is a preference for organizing projects that involve dyactic interactions with a research facility that requires no intervening human agency, and an aversion to undertaking contracts for collaborative work among research groups that would transcend institutional or organizational boundaries Further studies that would make use of ancilliary information, including interviews with principals in the formation of the UK’s e-Science core program, obviously would be required to throw light on the validity of these or still other plausible explanations.

References

G. Allen, T. Goodale, M. Russell, E. Seidel, and J. Shalf (2003): 'Classifying and Enabling Grid Applications', in F. Berman, G. Fox, A. J. G. Hey, and F. Berman (eds.): *Grid Computing: Making the Global Infrastructure a Reality*, John Wiley and Sons, 2003, pp. 601–613.

M. Bada, R. Stevens, C. Goble, Y. Gil, M. Ashburner, J. A. Blake, J. M. Cherry, M. Harris, and S. Lewis (2004): 'A short study on the success of the gene ontology', *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, Feb. 2004, pp. 235–240.

P. R. Carlile (2002): 'A pragmatic view of knowledge and boundaries: Boundary objects in new product development', *Organization Science*, vol. 13, no. 4, 2002, pp. 442–55.

M. Castells (1996): *The Rise of the Network Society*, Blackwell, Oxford, UK.

P. A. David (2006): 'Towards a cyberinfrastructure for enhanced scientific collaboration: Providing its 'soft' foundations may be the hardest part', in *Advancing Knowledge and the Knowledge Economy*, B. Kahin and D. Foray (eds.), MIT Press, Cambridge, MA, 2006.

P. A. David and M. Spence (2003): 'Towards institutional infrastructures for e-science: The scope of the challenge.' A Report to the Joint Information Systems Committee of the Research Councils of Great Britain, *Oxford Internet Institute Report No. 2*. September 2003. [Available at: http://www.oii.ox.ac.uk/publications]

P. A. David and M. Spence (2008), "Designing Institutional Infrastructures for e-Science," Forthcoming in *Legal and Policy Framework for e-Research*, Brian Fitzgerald, ed., Sydney, Australia: University of Sydney Press, 2008 [Available as SIEPR Discussion Paper No. 07-023 (December 2007) at: http://siepr.stanford.edu/papers/pdf/07-23.html]

P.A. David and W.E. Steinmueller (2003): 'The Economics of Research Collaboration and Collaboration Technologies', *Economics of Innovation and New Technologies*, vol. 12, no. 1-2 [Special Issue edited by P. A. David and W. E. Steinmueller], January 2003.

EPSRC (2007): 'Connecting Communities for the Digital Economy Workshop', available at http://www.epsrc.ac.uk/CMSWeb/Downloads/Calls/ConCommWrkshp.doc.

T. Finholt (2003): 'Collaboratories as a new form of scientific organization', *Economics of Innovation and New Technology*, vol. 12, no. 1, 2003, pp. 5-25.

C. Goble and D. D. Roure (2002): 'The Grid: An application of the semantic web', *ACM SIGMOD Record*, vol. 31, no. 4, Dec. 2002, pp. 65–70.

T. Hey (2002): 'Towards an e-Science Roadmap', available at http://www.nesc.ac.uk/news/ukroadmap180402/TonyHeyTowards_an_eScience_Roadmap.ppt.

T. Hey and A. Trefethen (2003): 'The data deluge: An e-science perspective', in F. Berman, A. Hey, and G. Fox (eds.), *Grid Computing - Making the Global Infrastructure a Reality*, John Wiley & Sons, 2003, pp. 809–824.

T. Hey, A. Trefethen, J. Fleming, K. Bartoszewska, C. Becker, R. Browne, L. Vousden, and J. Whalley (2002): 'The UK e-Science core programme', Annual Report 1, National e-Science Centre, 2002.

J. Swan and H. Scarbrough (2005): 'The politics of networked innovation', *Human Relations*, vol. 58, no. 7, July 2005, pp. 913–943.

H. Tangmunarunkit, S. Decker, and C. Kesselman (2003): 'Ontology-Based Resource Matching in the Grid – The Grid Meets the Semantic Web', in *The SemanticWeb - ISWC 2003*, Springer, Berlin, Germany, 2003.

National Science Foundation, Cyberinfrastructure Vision for 21st Century Discovery, Washington D.C., March 2007, available at: http://www.nsf.gov/od/oci/CI_Vision_March07.pdf.