SIEPR Discussion Paper No. 08-10

Will e-Science Be Open Science?

By
Paul A. David
Stanford University

Matthijs den Besten
Oxford e-Research Centre

Ralph Schroeder
Oxford Internet Institute

December 2008

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

A chapter contributed to
World Wide Research: Reshaping the Sciences and the Humanities in the Century of Information
William H. Dutton and Paul Jeffreys (Eds.), Forthcoming from MIT Press in 2009

Will *e-Science* Be *Open Science*?

Paul A. David
Stanford University & UNU-MERIT (Maastricht, Netherlands)
& All Souls College, Oxford
pad@stanford.edu

Matthijs den Besten
Oxford e-Research Centre
matthijs.denbesten@oerc.ox.ac.uk

Ralph Schroeder
Oxford Internet Institute
ralph.schroeder@oii.ox.ac.uk

This Version: 14 December, 2008

Abstract

This contribution examines various aspects of "openness" in research, and seeks to gauge the degree to which contemporary "e-science" practices are congruent with "open science." Norms and practices of openness are held to have been vital for the work of modern scientific communities, but concerns about the growth of stronger technical and institutional restraints on access to research tools, data and information recently have attracted notice – in part because of their implications for the effective utilization of advanced digital infrastructures and information technologies in research collaborations. Our discussion clarifies the conceptual differences between e-science and open science, and reports findings from a preliminary look at practices in U.K. e-science projects. Both parts serve to underscore the point that it is unwarranted to presume that the development of e-science necessarily promotes global open science collaboration. As there is evident need for further empirical research to establish where, when, and to what extent "openness" and "e-ness" in scientific and engineering research may be expected to advance hand-in-hand, we outline a framework within which such a program of studies might be undertaken.

1. Introduction

Anyone enquiring about "e-science" is bound to be led to a quotation from John Taylor's (2001) introductory description of this movement's essence as being "about global collaboration in key areas of science and the next generation of infrastructure that will enable it." Although much that has been written about e-science is occupied with the engineering and application of an enhanced technological infrastructure for the transmission, processing and storing of digital data and information (Hey, 2005), this paper steps back to consider other, non-technological requirements for attaining the ostensible goal of e-science programs – augmenting the scale and effectiveness of global collaboration in scientific research.

Global scientific collaboration takes many forms, but from the various initiatives around the world a consensus is emerging that collaboration should aim to be "open" -- or at least that there should be a substantial measure of "open access" to the data and information

underlying published research, and to communication tools. For example, the Atkins Committee, in a seminal NSF report that set the stage for research on "cyber-infrastructure" in the natural sciences and engineering in the US, advocated "open platforms" and referred to the grid as an "infrastructure for open scientific research" (Atkins, et al., 2003:pp. 4, 38). In a follow-up report expanding that vision to include the social sciences, Berman and Brady (2005:pp.19) likewise stress the need for a "shared cyber-infrastructure." In the UK, the e-Science Core Program has required that the middleware being developed by its projects be released under open source software licenses, and established an Open Middleware Infrastructure Institute (OMII). The e-Infrastructure Reflection Group (a high level European body formed in 2003 to monitor and advise on policy and administrative frameworks for easy and cost-effective shared use of Grid-computing, data storage, and networking resources) has gone further, issuing an "e-infrastructure roadmap" (Leenaars, 2005:pp.15-17, 22, 27) which calls for open standard grid protocol stacks, open source middleware, "transparent access to relevant [grid] data sources, and sharing of run-time software and interaction data including medical imagery, high-resolution video and haptic and tactile information"; and for public funding of scientific software development, because "current Intellectual Property Right solutions are not in the interest of science" (p. 16).

Provision of enhanced technical means of accessing distributed research resources is neither a necessary nor a sufficient condition for achieving open scientific collaboration (David 2005, David and Spence 2008). Collaboration technologies – both infrastructures and specific application tools and instruments – may be used to facilitate the work of distributed members of "closed clubs," including government labs engaged in secret defense projects, and corporate R&D teams that work with proprietary data and materials, guarding their findings as trade secrets until they obtain the legal protections granted by intellectual property rights. Nor do researchers' tools *as such* define the organizational character of collaboration. This is evident from the fact that many academic researchers who fully and frequently disclose their findings, and collaborate freely with colleagues on an informal, non-contractual basis, nonetheless employ proprietary software and patented instruments, and publish in commercial scientific journals that charge high subscription fees.

At the same time, it should be acknowledged that the availability of certain classes of tools, and the ease with which they may be used by researchers within and across scientific domains, is quite likely to affect organizational decisions and shape the ethos and actions of the work groups that adopt those tools. Some basic collaboration technologies -- notably e-network infrastructure such as grid services and middleware platforms -- are particularly potent enablers of distributed multi-participant collaborations; they may significantly augment the data, information and computational resources that can be mobilized by more loosely organised, the "bottom-up" networks of researchers engaging in "open science." The availability of access to those resources on "share-and-share alike" terms can induce researchers' participation in passive as well as active collaboration arrangements, acquainting them with benefits of cooperation and thereby reinforcing the ethos of open science.

The sections that follow present our understanding of the term "open science," its significance for epistemologists, sociologists and economists studying the relationships between institutional structures, working procedures and the formation of scientific knowledge, and discuss ways that this concept may be applied to assess the "open-ness" of certain structural features and organizational practices observable in programmatic e-science initiatives and particular projects. We then consider some results from preliminary empirical enquiries, intended primarily to illustrate the empirical implementation of our proposed conceptual framework. Although only a limited sample of U.K. e-science projects (to date) have been selected for study from this perspective, the recent findings based on structured interviews and responses to a targeted email survey of research project directors display noteworthy consistencies and support our contention that further investigation along the conceptual and methodological lines explored will prove to be both feasible and illuminating.

2. Open Science

Many of the key formal institutions of modern science are quite familiar not only to specialists concerned with the economics and the sociology of science, technology and innovation, but equally to academic researchers of all disciplinary stripes. It is a striking phenomenon, well noted in the sociology of science, that there is high degree of mimetic professional organization and behavior across the diverse cognitive domains of academic endeavor. Whether in the mathematical and natural sciences, or the social sciences or the humanities, each discipline has its professional academies and learned societies, journal refereeing procedures, public and private foundation grant programs, peer-panels for merit review of funding applications, organized competitions, prizes and public awards. The outward forms are strikingly similar, even if the details of the internal arrangements may differ.

2.1. The norms of "open science"

The norms of "the Republic of Science" that were so famously articulated by Merton (1942, 1973) are summarized compactly by the mnemonic device "CUDOS": communalism, universalism, disinterestedness, originality, and skepticism.[1] These five key norms constitute a clearly delineated ethos to which members of the academic research community generally subscribe, even though their individual behaviors may not always conform to its strictures. ***C****ommunalism* emphasizes the cooperative character of enquiry; ***U****niversalism* emphasizes the need to keep entry into scientific work and discourse open for all persons of "competence;" ***D****isinterestedness* emphasizes the neutrality of researchers vis-à-vis the nature and impact of the knowledge that they contribute; ***o****riginality* is the basis on which collegiate reputations are built and rewards are based; and consequently ***S****kepticism* is the appropriate attitude towards all priority claims that are made.

Separately as well as systemically, these norms lead to the functional allocation of resources in an idealized research system. This is to say that a complete functionalist explanation can be provided for the existence of the "open" part of the institutional complex of modern science, by focusing on its economic and social efficiency properties in the pursuit of knowledge, and making explicit the supportive role played by norms that tend to reinforce cooperative behaviors among scientists (Dasgupta and David 1987, 1994; David 1998, 2003). This rationale highlights the "incentive compatibility" of the key norm of disclosure within a collegiate reputation-based reward system grounded upon validated claims to priority in discovery or invention. In brief, rapid disclosures abet rapid validation of findings, reduces excess duplication of research effort, enlarge the domain of complementarities and yield beneficial "spill-overs" among research programs. Without delving deeper into the details of this analysis, it may be noted that it is the difficulty of monitoring research effort that make it necessary for both the open science system and the intellectual property regime to tie researchers' rewards in one way or another to priority in the production of observable "research outputs" that can be submitted to "validity testing and valorization" – whether directly by peer assessment, or indirectly through their application in the markets for goods and services.

The specific functionality of the information-disclosure norms and social organization of open science rests upon the greater efficacy of data and information-sharing as a basis for the cooperative, cumulative generation of eventually reliable additions to the stock of knowledge. Treating new findings as tantamount to being in the public domain fully exploits the "public goods" properties that permit data and information to be concurrently shared in use and re-used indefinitely, and thus promotes faster growth of the stock of knowledge. This contrasts with the information control and access restrictions that generally are required in order to appropriate private material benefits from the possession of (scientific and

[1] The nmenonic *Cudos* was introduced by Merton's 1942 essay on the normative structure of science, but the association of the "O" with originality was a subsequent modification that has become conventional (see Ziman 1994).

technological) knowledge. In the proprietary research regime, discoveries and inventions must either be held secret or be "protected" by gaining monopoly rights to their commercial exploitation. Otherwise, the unlimited entry of competing users could destroy the private profitability of investing in research and development.

The relationship between the conduct of the scientific research process as seen from the epistemological perspective, and the norms perceived by Merton to both underlie and receive reinforcement from the institutionalized organization and stratified social structure of scientific communities, is a subject with which sociologists and philosophers of science have continued to wrestle. Indeed, one that has occasioned some internal disciplinary struggles as well as the usual difficulties in cross-disciplinary communication. Still, reviewing the evolving literatures in the philosophy and social science of scientific research, one may say that it is now broadly recognized that there is a reciprocal interdependence between the ethos and normative structures of research communities, and the informal and institutionally reinforced conditions of access to research findings, underlying data and methodologies.[2] Together, the norms and rules affecting communications through personal networks and broadcast channels, and the interchange of personnel among scientific workgroups, shape the possibilities of coordination and effective collaboration. They thereby impinge upon the efficiency of scientific projects internal use of resources, and of resource allocation among the members of separate projects who constitute "invisible colleges" that are distributed across academic departments, institutes, universities, transcending national and regional boundaries – extending even into research laboratories of business corporations and government agencies.[3]

2.2. Questions about the degrees of "openness" of the organization of research – the practice of "open science"

The foregoing considerations have given us strong reason to regard the formal and informal *institutional* arrangements governing access to scientific and technical data and information, no less than to physical research facilities, instruments, materials, as critically influential among the factors determining how fully e-science will be able to realize the potentials for the advancement of reliable knowledge that are being created by advances in digital information technologies.

Questions concerning the actual extent of "openness" of research processes identified with contemporary e-science therefore ought to address at least two main sets of issues pertaining to the conduct of "open science." The first set concerns the terms on which individuals may enter and leave research projects. Who is permitted to join the collaboration? Are all of the participating researchers able to gain full access to the project's databases and other key research resources? How easy or hard is it for members and new entrants to develop distinct agendas of enquiry within the context of the ongoing project, and how much control do they retain over the communication of their findings? What restrictions are placed (formally or

[2] See e.g., Quinne (1969), Kuhn (1962/1970), Fuller (1994), Kitcher (1993), for epistemology of science; Cole and Cole (1967), Crane (1972), Cole (1978), Ben-David (1984) for sociology of science after Merton; Barnes (1974, 1977), Bloor (1976), Knorr-Cetina (1981)Latour and Woolgar (1979), Shapin (1994), and the survey by Callon (1995). On the relationship of the "new economics of science" to the foregoing disciplinary developments, see David (1998) and den Besten, David and Schroeder (2009).

[3] See Fry, Schroeder and den Besten (2008) and Schroeder (2008) for discussion of the distinction between generic research-technologies and narrowly defined research tools, and its bearing on the potential for openness in e-science. It has been suggested that generic-research technologies traverse traditional disciplinary boundaries and draw divergent disciplinary specialities together through a common language and approach, whereas research tools are embedded within differentiated, highly specialized research domains and tend to impose epistemic boundaries between fields. Standardization of open platforms supporting particular applications tools and annotated databases would appear in this light be facilitated by mobility of researchers between workgroups and the formation of inter-group collaborations within specific areas of research. Generic middleware infrastructure services, e.g. grid services would be more likely to facilitate coordination and distributed trans-disciplinary collaborations.

informally) on the uses they may make of data, information and knowledge in their possession after they exit from the research collaboration?

The second set of questions concerns the norms and rules governing disclosure of data and information about research methods and results. How fully and quickly is information about research procedures and data released by the project? How completely is it documented and annotated--so as to be not only accessible but also useable by those outside the immediate research group? On what terms and with what delays are external researchers able to access materials, data and project results? Are findings held back, rather than being disclosed in order to first obtain intellectual property rights on a scientific project's research results, and if so, then for how long is it usual for publication to be delayed (whether by the members or their respective host institutions)? Can research partners in university-business collaborations require that some findings or data not be made public? And when intellectual property rights to the use of research results have been obtained, will its use be licenses to outsiders on an exclusive or a non-exclusive basis? Do material transfer agreements among university-based projects impose charges (for cell lines, reagents, specimens) that require external researchers to pay substantially more than the costs of making the actual transfers? In the case of publicly funded research groups, are the rights to use such legally "protected" information and data conditional on payment of patent fees, copyright royalties such that the members of the research group has any discretionary control, or is control exercised by external parties (in their host institution, or the funding sources)?

Ideally, these and still other questions may be formulated as a simple checklist such as the one devised by Stanford University (1996) to provide guidelines for faculty compliance with its "openness in research" policy. The Stanford checklist, however, having initially been designed primarily to implement rules against secrecy in sponsored research, actually is too limited in its scope for our present purposes, and a fuller, more specific set of questions (inspired by this source) has been designed for data-gathering data in the context of contemporary U.K research projects. This empirical framework has been "field-tested" both in a small number of structured interviews, and a subsequent more extensive email-targeted survey of e-science project-leaders.[4] It is not intended to be comprehensive, and, instead, focuses on salient aspects of "openness and collaboration in academic science research" that could be illuminated by implementing systematic surveys of this kind on a much wider scale.

Of course, to pursue a substantially expanded program of inquiry into evolving e-science practices along these lines would necessitate some substantive modifications of the questionnaire in order to appropriately "customize" the interview protocols and the survey template, which been designed for exploratory, "proof-of-concept" investigations. Conducting research of this kind across a widened international survey field certainly would require adjustments to allow for the greater diversity of institutional and organizational forms, research cultures, languages and technical nomenclatures. Furthermore, practical considerations might call also for abridging the questionnaires, so as to reduce the burden upon respondents and obtain a reasonably high response rates from an internationally administered survey – while avoiding costly individual email-targeting and follow-up requests for cooperation from potential respondents.

3. e-Science as Open Science: Evidence from Structured Interviews and a Survey of U.K. e-Science Project P.I.'s

Researchers in public sector science and engineering organizations historically have been at the forefront of many basic technological advances underlying new paradigms of digital

[4] For a report on the structured interviews, see Fry, Schroeder and den Besten (2008). David, den Besten and Schroeder (2006) presents a preliminary version of the framework of questions from which were developed both the structured interview protocol and subsequent on-line survey questionnaire. For the latter, see text Box 1 and the web layout of the survey instrument that is reproduced in the Oxford Internet Insitute OeSS project report by den Besten and David (2008). .

information creation and dissemination. Their pressing needs for more powerful information processing and communication tools have led to many of the key enabling technologies of the "Information Society," including its mainframe computers, packet-switched data networks, the TPC/IP protocols of the Internet and the World Wide Web, its proliferation of markup languages, the Semantic Web and many more recent advances that facilitate distributed conduct of collaborative research. For essentially the same reasons, scientific and engineering research communities throughout the world now are active in developing not only technical infrastructure tools like the grid and middleware platforms, but a new array of shareable digital data and information dissemination resources, including public-domain digital data archives and federated open data networks, open institutional repositories, "open access" electronic journals, and open-source software applications. (David, 2005; Dalle et al., 2005; David and Uhlir, 2005, 2006; Uhlir and Schröder, 2006; Schroeder 2007b). Here, we focus on one of these efforts in particular: the U.K. e-Science programme (cf. Jeffreys, this volume).

The U.K. e-Science programme has given rise to several high-profile projects. By looking more closely at these projects we can get a first impression of the degree of openness in e-Science as a whole. Besides, the questions about degrees of openness that we outlined in section 2.2 could be answered, at least in some part, by the people involved in the research and development projects associated with the e-Science program in the UK. Fry, Schroeder and den Besten (2008) have carried out a series of structured interviews with a small group of the principal investigators of U.K. e-Science projects, designed to assess perceptions and practices relating to aspects of "openness" of the projects for which they had leadership responsibilities. A related questionnaire, suitable for implementation in an on-line Internet survey was developed on the basis of this experience and implemented in an email targeted survey of a larger population of U.K. e-science projects P.I.'s. The results obtained from the latter survey by den Besten and David (2008a) are broadly congruent with the detailed and more nuanced impressions drawn from the structured interviews.

3.1 e-Science Research Projects

Let us first look at three projects in more detail: (1) e-DiaMoND, a Grid-enabled prototype system intended to support breast cancer screening, mammography training, and epidemiological research; (2) MiMeG, which aims to produce software for the collaborative analysis and annotation of video data on of human interactions; and (3) Comb*e*-chem, an e-science test-bed that integrates existing sources of chemical structure and properties data, and augments them within a grid-based information and knowledge environment. Although none of these quite different projects have developed income-generating activities that might conflict directly with their adherence to open science norms, it is striking that all three have confronted other difficult issues related to "control rights" over data and information.

For e-DiaMoND the problem of control of mammography images remained unresolved when this "proof of concept" project reached its scheduled end. The researchers' original intentions to distribute standardized images for research and diagnostic purposes over electronic networks, clashed with the clinicians' concerns about their professional responsibilities to patients, protecting patient privacy, and assuring ethical uses of the data. Convincing clinical practitioners to trust the researchers, and engineering a comprehensive, adequately flexible security system proved to be less straightforward than had been expected (Jirotka et al., 2005). Even "to develop a clear legal framework that fairly accounts for the needs of patients, clinicians, researchers and those in commerce"— one that the project's diverse partners would be able to work with – has been surprisingly difficult (Hinds et al., 2005).

MiMeG, an ESRC funded e-social science project, encountered similar problems: the researchers who employed the tool for collaborative analysis of video-streams felt that the trust of the persons whose images they were studying would be violated by archiving the collaboration's data and making it available for re-use by other researchers, possibly for purposes other than the one for which consent originally had been obtained. It remains to be

seen whether or not the ethical *desiderata* of privacy and informed consent of experimental subjects can be satisfied in future projects of this kind that plan sharing research data via the grid.

For the present, however, MiMeG has abandoned the project's initial intention to analyze video collaboratively via e-networks, and is focusing on the development of video analysis tools that other researchers can use. In that connection it is significant that the research software created by MiMeG is being released under the GNU GPL license (and hence distributed at minimal cost for non-commercial use). This policy resulted at least in part from the use of some GPL components (such as the MySQL relational database) to build the project's software tools. In addition, however, MiMeG's is encouraging external users to participate in further developing it recently released video analysis software tools. In these respects, the project has been able to go forward in the collaborative "open science" mode.

The Comb*e*-chem project at Southampton University is funded under the EPRSC's e-science program and includes several departments and related projects. Only a few organizational features of this complex collaboration can be considered here, but several important aspects of its activities clearly are "open". One utilises the pre-existing EPSRC National Crystallographic Service, which has allowed remote "users" from UK universities to submit samples of chemical compounds to the laboratory at Southampton for x-ray analysis. Comb*e*-chem accepts submitted samples and returns them via a Globus-based grid and web services infrastructure (see Coles et al. 2005: appendix B). A present this service has some 150 subscribers who submit more than 1000 samples per annum (Frey 2004: 1031).

In addition to demonstrating and developing this grid implementation, a major project goal is to increase the archiving of analysed samples, thereby averting the loss of un-archived information and the consequently wasteful repetition of crystallographic analyses. Formerly, chemical analysis results yielded by these techniques were "archived" by virtue of their publication in research journals, most of which were available on a "subscription only" basis. Now it is possible to make results available in open access repositories via the open archive initiative (OAI), and deposited in *e*-BankUK archives and *e*Prints publications (Coles et al., 2005). Because they are put into RDF (Resource Description Framework) and other standard metadata formats, the archived results are searchable via the Semantic Web. With only 20 per cent of data generated in crystallographic work currently reaching the public domain (Allen 2004) and not all of it beings readily searchable, this service extension is an important open science advance. Combe-chem's interrelated e-science activities thus illustrate four facets of open science practice: (a) using the Globus and web services open source grid software, (b) providing web access to shared resources for a diverse research community, (c) open access archiving and dissemination of results through an open repository, and (d) formatting of information using open standards. Like other publicly funded academic research, the project interacts easily with the world of commercial scientific publishing: fee charging journals that adhere to "subscriber only access" policies provide readers with links to the Comb*e*-chem data archive Moreover, as is the case in other collaborative projects that fit the traditional open science model quite closely, Comb*e*-chem has been able nonetheless to draw some sponsorship support from industry -- IBM having been interested in this deployment of a grid service.[5]

3.2 Open science in e-science -- policy or contingency? Insights from in-depth interviews

Fry, Schroeder and den Besten (2008) report the findings from their use of a structured interview in conducting in-depth interviews about the relationships between collaboration in 'e-science' and 'open science,' with 12 individuals who had roles as principal investigators, project managers and developers engaged in UK e-Science projects during 2006.[6] The

[5] See Fry, Schroeder and den Besten (2008), based on J. Frey (P.I.,Combe-chem) interview, on 29.22.2005.

[6] Fry, Schroeder and den Besten's (2008) structured interview protocol elaborated and modified the extended questionnaire proposed by David, Schroeder and den Besten (2006).

interview questions focused on research inputs, software development processes, access to resources, project documentation, dissemination of outputs and by-products, licensing issues, and institutional contracts. A focal interest of the approach in this study was the authors' juxtaposition of research project leaders' perceptions and views concerning research governance policies at the institutional level, with the responses describing local practices at the project level. As a detailed discussion of the responses (along with related documentary evidence drawn from the respective project's websites) is available elsewhere, it will be sufficient here to summarize briefly the main thrust of Fry, Schroeder and den Besten's (2008) findings.

Their interviews suggest that the desirability of maintaining conditions of "openness" in "doing (academic) science" is part of a generally shared research ethos among this sample of university-based project leaders. More specifically, the latter were not only cognizant of but receptive to the U.K. e-Science Pilot Program's strong policy stance favoring open source software tools and sharing of informational resources. Nevertheless, there were many uncertainties and yet-to-be resolved issues surrounding the practical implementation of both the informal norms and formal policies supporting open science practices. Making software tools and data available to external users might mean simply putting these research outputs on-line, but that need not be the same thing as making them sufficiently robust and well-documented to be widely utilized.[7] It seems that for those with leadership responsibilities at the project level, the most salient and fundamental challenges in resolving issues of openness in practice and operating policies, and thereby moving towards coherent institutional infrastructures for e-science research, involve the coordination and integration of goals across the diverse array of e-science efforts.[8]

By comparison, much less concern is voiced about the resolution of tensions between IPR (intellectual property rights) protections and the provision of timely common-use access to research tools, data and results. This is not really surprising when the context of the survey is considered, even though these issues have been very much at the center of public discussions and debates about the effects of the growth of "academic patenting" on the "openness" of publicly funded research.[9] The U.K. e-science was strongly focused on the development of software tools in support of research, and even in areas of application it did not enter into life science areas, particularly biomedical and biotechnology research, fields in which patenting is especially important for subsequent commercial innovation. EU policy has circumscribed the patenting of software (without eliminating the patenting of embedded algorithms and a

[7] As Fry, Schroeder and den Besten (2008) point out: "The effort to make the tools or data suitable or robust enough to make them into a commonly used resource may be considerable, and thus represents a Catch-22 situation for researchers: a large effort can be made, which may not be useful, but if it is not attempted, then it cannot be useful in the first place. Nevertheless, all projects expressed the aspiration to contribute to a common resource, even if this was sometimes expressed as a hope rather than a certainty or foregone conclusion."

[8] Coordination and integration problems calling for solutions that take the form of interoperability standards posed particularly difficult challenges for on-going projects in the UK e-Science Pilot Programme, according to the Fry, Schroeder and den Besten (2008): whereas some new software tools required compatibility with existing tools (for example, CQeSS needed to be interoperable with Stata) and this might be technically difficult to implement, achieving integration with other tools that are currently under development confronts more fundamental uncertainties about the requirements for compatibility or interoperability. The same applies to complying with standards, ontologies and metadata that are still in the process of development, which suggests that the during the formative phases of an e-infrastructure-building program, the rhetoric of projects' goals being to contribute to seamless integration and ubiquitous access to scientific computing and toolsets can be so forward-looking as to be perceived as unrealistic and consequently a source of frustration.

[9] See e.g., David and Hall(2006);David(2007). Much of that discussion, however, has focused on the implications of the patenting of research tools, and sui generis legal protection of database rights (in the EU) in the areas of genomics, biogenetics and proteinomics, the patenting computer software (in the US) and computer implemented inventions (in the EU), and extensive patenting of nanotechnology research tools. While those have been very active fields of academic science research, and growing university-ownership of patents, they are not represented in the U.K.'s e-Science core program and so do not appear among the projects included in either the structured interview or the survey samples discussed here.

wider class of so-called "computer implemented inventions"), and in the U.K. itself, government agencies funding e-science projects have explicitly prohibited university grant and contract recipients from filing software patents that would vitiate the open source licensing of their outputs of middleware and applications software.

Most of the foregoing observations, although drawn from structured interviews conducted with a only a very small and non-random sample of project leaders, turn out to be quite informative -- in that these impressions are reinforced by the findings of a subsequent on-line survey that sought responses from the entire population of principal investigators on U.K. e-science projects.

3.3 Contract terms and "open-ness in research": survey finding on e-science projects

Systematic and detailed data at the individual project level about the openness of information and data resources remains quite limited, both as regards actual practices and the priority assigned to these issues among project leaders' concerns. A glimpse of what the larger landscape might be like in this regard, however, is provided by the responses to the online survey of issues in U.K. e-science that was conducted among the principal investigators that could be identified and contacted by email on the basis of National e-Science Centre (NeSC) data on the projects and their principal investigators (den Besten and David, 2008). Out of the 122 P.I.'s that were contacted, 30 responded with detailed information for an equal number of projects.[10] A comparison of the distribution of the projects for which responses were obtained and the distribution of the population of NeSC projects showed remarkable similarities along the several dimensions on which quantitative comparisons could be made -- including project grant size, number of consortium members and project start dates. This is reassuring, providing a measure of confidence in the representativeness of the picture that can be formed from this admittedly very restricted sample.

Formal agreements governing the conduct of publicly funded university research projects may, and sometimes do, involve explicit terms concerned with the locus and nature of control over data and publications, and the assignment of intellectual property rights based upon research results, especially when there are several collaborating institutions and the parties include business organizations. The survey sought to elicit information about project leaders' understandings of these matters and the importance they attached to such bearing as the terms of their respective project's agreement might have upon information access issues. It did so by posing various questions intended to probe the extent of participant's knowledge of the circumstances of the contractual agreement governing their project, namely, the identities of the parties responsible for its initial drafting and subsequent modifications (if any), as well as some of the contract's specific terms.

The results of the survey, which are presented in more detail elsewhere (Den Besten & David 2009), suggest that the projects surveyed generally are free from positive, contractually imposed restrictions on the participation of qualified researchers and significant restraints upon participants' access to critical data resources, and ability eventually to make public their research results. (See Box 1, below, for survey questionnaire.) A substantial fraction of project members appear not to be informed about the specifics of

[10] This number represented just over 10 percent of the projects listed by NeSC, implying a "project response rate" of 25 percent. The number of individual responses to this survey was larger, because P.I.'s receiving the email request were asked also to send it on to non-P.I. members of their project (which yielded an additional 21 responses that are not discussed here; also, in 3 cases more than one P.I. for a single project returned the questionnaire. The present analysis used only the one with the lowest frequency of "don't know" responses. The low apparent response rate from P.I.'s and projects may be due in some part to the relatively short time interval allowed for those who submitted survey replies to be eligible to receive a book-token gift. The existence of projects that appear more than once in the NeSC database and had multiple (co-) P.I.'s also would contribute to reducing the apparent rate of "project" responses.

the project agreements under whose terms they are working. This is not very surprising, as many scientists express disinterest if not impatience with such matters, wishing to get on with their work without such distractions, and therefore leaving it to others -- including some among their fellow P.I.'s -- to deal with legal aspects of governance if and when problems of that nature intrude into the scientific conduct of the project. Therefore, it could be taken as a healthy indication, namely, that issues involving restrictive provisions projects' contractual terms intrude upon the researchers' work only very infrequently, and so have remained little discussed among them.

Encouraging as that would be, the absence of formal, contractually imposed restraints on disclosure and access to scientific information and data resources leaves a substantial margin of uncertainty as to how closely the norms of "open science" are approximated by the operating practices and informal arrangements that are typically found within these projects. To probe into those important areas of "local" policy and practice, it is possible to examine the results obtained from a different set of the survey's questions.

4. Provision of information access in e-science projects: practices and policy concerns

What stands out most clearly from the findings of Den Besten and David (2009) is that high-level policy guidelines, set by the funding agency, can exert a potent influence on the pattern of adoption of open access archiving of scientific research products. In this instance there was an important early policy commitment by the U.K. e-Science core programme that middleware "deliverables" from its pilot projects would be made available as open source code, and this requirement for the research projects has been maintained – even through there has been an evolution away from the original expectations of open source release of these output under GNU General Public Licenses once they had passed through the OMII's enhancement and repacking process.[11]

The extent to which the provision of access to data and information is perceived *at the project level* to be matters of explicit policy concern varies with the projects' roles in e-Research. This is only to be expected, particularly in view of the varied nature of these projects' "deliverables" and the existence of higher level policy regarding the software that is being created. A clear pattern of co-variation is evident in the responses to the question: "Was the provision of access to data and

[11] See David, denBesten, and Schroeder (2006, 2009) on the evolution of OMII's policy on the licensing of its releases of middleware.

BOX 1 : Questions from the On-line 2008 Survey of U.K. e-Science Project Participants

[See den Besten and David (2008) for Web layout of the questionnaire (freely available under Creative Commons (non-commercial use, attribution only) license]

I. Introduction

1. About the project
1.a Project Acronym
1.b Project Homepage URL

2. What is your present role/position in this project?
Principal Investigator / Co-Principal Investigator / Research Associate / Research Officer / Admin/Tech Support / Other (please specify)

3. Approximately when did you start/join the project?
Project start date; date you joined the project (if different)

4. Is this your first e-science project?
Yes / No;

5. Is this your only current e-science project?
Yes / No;

6. Which among the following most accurately describes this e-science project's purpose(s)?
(*Check more than one if appropriate):*
6.a Generic "tool development": building solutions with many application domains
Facilitate collaboration among non-co-located researchers / Provide access to remote hardware instruments / Provide access to specialized software (e.g. for simulation, spectroscopic analysis) / Link (federate) datasets and databases / Distribute computing capacity
6.b Application development: tailoring "middleware" to the needs of specific kinds of end-users
6.c "End-use" application: conducting research that uses e-science tools

II. Project Agreements

7. Creation and changes to the project agreement:
7.a Who proposed the first template for the contract or agreement?
University Office / Funding Agency / Industrial Partner / Other / Don't know / Not Applicable
7.b Who sought whatever major modifications had to be made to conclude a contract or agreement that started the project?
7.c If the agreement was modified after the launch of the project, who was mainly responsible for initiating the changes?

8. Does this project or agreement:
8.a Restrict research participation (faculty, student, others) based on country of origin or citizenship?
Yes / No / Don't Know / Not Applicable
8.b Require research participation in EU-citizen-only meetings?
8.c Prohibit the hiring of non-EU citizens to be involved in the proposed research?
8.d Grant the sponsor a right of prepublication review for purposes other than the preparation of patents or the exclusion of proprietary data?
8.e Provide that any part of the sponsoring, granting, or establishing documents may not be disclosed?
8.f Contain language referring to or mandating compliance with government regulations restricting the export of certain materials or software programs?
8.g Limit access to confidential data so centrally related to the research that a member of the research group who was not privy to the confidential data would be unable to participate fully in all of the intellectually significant portions of the project?

III. Project Infrastructure

9. Which of the following facilities are part of the infrastructure in place in the project?
9.a A common repository of the project's working papers and memoranda:
Yes / No / Don't Know / Not Applicable
9.b A common repository of project-created software source code:
9.c A common repository for data:
9.d A university or department-wide open access repository for project publications:
9.e An open access repository for project's preprints:
9.f An open access repository for project-created middleware source code:
9.g An open access repository for project-created applicatons source code:
9.h An open access repository for version-controlled development code:
9.i An open access repository for project-generated data:

10. Do all participants within the project have access to these facilities?

11. Are all participants instructed to deposit their work in one or more of the following repositories?

12. Does the project pay fees associated with submission or depositing of materials?

IV. Project Access

13. To what extent, and through which means does the project provide external researchers with access to the following materials?
On public project web site / On private project web site / On request / No Access / Don't Know
13.a Peer reviewed publications
13.b Preprints
13.c Technical Reports
13.d Minutes
13.e Research Protocols
13.f Lab books
13.g Procedures describing the setup of experiments (workflows)
13.h Procedures describing the transformation and analysis of data (scripts, filters, functions)

14. Did your project undertake to "federate" ("deep link" or coordinate across institutional boundaries) its digital repositories for data and/or software with those of other research groups?
Yes / No / Not Applicable
14.a Data - With your project's collaborators at other institutions?
14.b Data - With other UK e-Science projects?
14.c Data - With projects based in other regions?
14.e Software - With your project's collaborators at other institutions?
14.g Software - With other UK e-Science projects?
14.h Software - With projects based in other regions?

15. If the repository "federation" attempts in which your project was involved were not completely successful, indicate in each case the nature and seriousness of the obstacles that were encountered:
Critical / Important / Not Important / N/A
15.a Technical incompatibilities:
15.b Privacy / confidentiality:
15.c Intellectual property rights charges:
15.d Refusal of other parties to federate (under any terms):
15.e High costs of implementation:
15.f Negotiation delays:
15.g Lack of personnel / funding for maintaining and managing updating, annotation, etc.:

V. Project Practice

16. Was the provision of access to data and information to members of the project a matter of particular concern and discussion in your project?

17. Was the provision of "open access" conditions to external researchers among the explicit goals communicated to members of your project?

18. What were the two or three most important obstacles in achieving "openness" in your project?

19. What were the two or three most important successes?

information to members of the project a matter of particular concern and discussion in your project?"; and a parallel question referring to "external researchers" (see Box 1, Questions 16,17).[12] Among theprojects engaged in *middleware development*, none expressed a concern for access within the project – presumably because the organization of the project and the ubiquity of open access code repositories meant that the matter one that had largely been settled. In contrast, however, the issue of external access was seen to be an important project concern by a third of the respondent P.I.'s from the projects developing *middleware*. That concern was expressed also by one-third respondents from projects involved with *user-communities* and *database resources*, especially the latter group.[13]

The responses concerning "obstacles encountered by the project in achieving "openness" (see Box 1, question 18) are consistent with the survey finding regarding actual practices and policy concerns at the project level, for they indicate that providing access to information to people *within* the project not found to be a problem deserving mention. All but two of the P.I.'s indicated at least one type of common repository to which participants were given access. Open access repositories are almost only provided where access for external research is seen as a concern within the project, which is the case for about one-third of the projects for which survey data is available. Project participants are not always instructed to contribute to the repositories when the latter are provided, and it appears to be generally assumed that they will do so. On the other hand, none of the respondents indicated that their project was paying fees for the maintenance of an institutional or external repository to which their researchers would be given access.[14] Among the respondents who stated that the provision of access to outsiders was an important project goal, almost two-thirds listed one or more obstacles that had been encountered in achieving it; whereas among those who stated that such provision was not a project concern, almost half volunteered that they had encountered practical obstacles to external dissemination of their research outputs.[15]

5. Conclusion

We have described both the rationale and key identifying characteristics of collaborative "open science," and have begun to explore ways to map the regions of practice where e-science and open science coincide. Although there are many e-science tools that could support distributed projects that conduct research in ways that accord more or less closely to open science norms, this does not assure that such is or will be the case where-ever collaborative research is pursued under the name of "e-science." Even academic e-science projects who leaders subscribe to the ethos of "open-ness in
research" and institute some concrete "open access" practices, fall short of those norms in one or more respects, especially in regard to effective sharing of data resources and timely external disclosure of research findings. But, as has been shown, e-science projects are far from homogeneous, and in order to understand the variations in their information sharing policies and practices it is necessary to take into account the diversity of their scientific purposes, the technical nature of their tasks and the details of their organizational structures. The review presented here of the empirical evidence pertaining to U.K.-funded e-science

[12] Over half of the projects having more diffuse purposes-- that is, purposes not preponderantly oriented toward either construction of middleware, research community usage, or applications and database resources -- failed to provide clear answers to questions 16 and 17. Responses from the "other purposes" group are not included in the analysis whose results are described in the text.

[13] Specifically, providing access to researchers outside the project was a significant concern for almost two-thirds of the *data*-centric projects and a third of *community*-centric projects.

[14] Perhaps this question should have been phrased differently, e.g.: "Would the project be willing to pay repository changes, and for the inclusion of open access journals?"

[15] 11 respondents listed external access among their project goals, 9 said it was not an important concern, and another 9 respondents left this question unanswered.

projects, has been able to draw upon recent studies that carried out a small number of in-depth (highly insightful) interviews with selected P.I.'s, and obtained quantitative data from the responses to an on-line survey of e-science project leaders and other participants. These efforts in data collection and analysis represent only a trial step in what is envisaged as a far broader and longer term program of systematic inquiries into the evolving global conduct of e-science.

Acknowledgements

Thanks to Anne Trefethen, Jenny Fry, Jeremy Frey and Mike Fraser for their contributions, and to Paul Jeffreys for editorial suggestions to improve the exposition. The preparation of this paper, and most of the research on the underlying data, has been supported by ESRC grant RES-149-25-1022 for the Oxford e-Social Science (OeSS) Project on Ethical, Legal and Institutional Dynamics of Grid-enabled e-Sciences. We gratefully acknowledge the institutional support provided by the Oxford Internet Institute and the Oxford e-Research Centre, and David also acknowledges the support of his research program on Knowledge, Networks and Institutional for Innovation (KNIIP) at the Stanford Institute for Economic Policy Research (SIEPR). We take full responsibility for the judgments expressed, as well as any mistakes and mis-interpretations that may be found herein.

References

F. H. Allen. High-throughput crystallography: the challenge of publishing, storing and using the results. *Crystallography Reviews*, 10:3–15, 2004.

B. Barnes, *Scientific Knowledge and Sociological Theory*, London: Routledge and Kegan Paul, 1974

. Barnes, *Interests and the Growth of Knowledge*, London: Routledge and Kegan Paul, 1977.

F. Berman and H. Brady. NSF SBE-CISE workshop on cyber-infrastructure and the social sciences. Final report, San Diego Supercomputing Centre, 2005.

D. Bloor, *Knowledge and Social Imagery* ,London: Routledge & Kegan Paul, 1976.

W. A. Brock. S. N. Durlauf, A formal model of theory choice in science, *Economic Theory*, 14(1), 1999: pp.13-130.

M. Callon, Four models for the dynamics of science.In *Handbook of science and technology studies*, Eds., S. Jasanoff, G.E. Markle, J.C. Petersen and T. Pinch (London: Sage Publications), 1995.

M. Callon, J.-P. Courtial, *Co-word analysis: a tool for the evaluation of public research policy* (Paris: Ecole Nationale Supérieure des Mines), 1989.

N. Carayol, The incentive properties of the Matthew Effect in academic competition, *Working Papers of BETA* 2003-11 ULP, Strasbourg, 2003.

N. Carayol, J -M. Dalle. Sequential problem choice and the reward system in Open Science, BETA *Working Papers of BETA* 2003-12, ULP Strasbourg, 2003. Revised and published in *Structural Change and Economic Dynamics*, 18(2), 2007:pp. 167-191].

N. Carayol, M. Matt, Does research organization influence academic production?: Laboratory level evidence from a large European university, *Research Policy*, 33(8), 2004: pp. 1081-1102.

N. Carayol, M. Matt, Individual and collective determinants of academic scientists' productivity, *Information Economics and Policy*, 18(1), 2006: pp. 55-72.

J. Cole, S. Cole, *Social Stratification in Science*, Chicago: Chicago University Press, 1973 .

S. Cole, Scientific reward systems: a comparative analysis," in *Research in Sociology of Knowledge, Sciences and Art*, Ed.,R. A. Jones (Greenwich, Conn.: JAI Press), 1978: pp. 167-190.

S. Cole, J.Cole, Scientific output and recognition, *American Sociological Review*, 32, 1967:pp.377-390.

S. J. Coles, J. G. Frey, M. B. Hursthouse, et al., The end-to-end crystallographic experiment in an e-science environment: From conception to publication. In *Proceedings of the Fourth UK e-Science All Hands Meeting*, Nottingham, UK, 2005(a).

S. J. Coles, J. G. Frey, M. B. Hursthouse, et. al., ECSES-Examining crystal structures using ”e-science”: a demonstrator employing web and grid services to enhance user participation in crystallographic experiments. *Journal of Applied Crystallography*, 38(819-826), 2005(b).

D. Crane, “Scientists at major and minor Universities: a study in productivity and recognition,” *American Sociological Review*, 30, 1965: pp.699-714.

D. Crane, *Invisible colleges: diffusion of knowledge in scientific communities*, (Chicago: The University of Chicago Press) 1972.

J.-M. Dalle, P. A. David, R. A. Ghosh and W. E. Steinmueller, Advancing economic research on the free and open source software mode of production. In *How Open Will the Future Be? Social and Cultural Scenarios based on Open Standards and Open-Source Software*, .Eds. M. Wynants and J. Cornelis, (Brussels: VUB Press), 2005. [Preprint available at: http://siepr.stanford.edu/papers/pdf/04-03.html.]

P. Dasgupta and P. A. David, Information disclosure and the economics of science and technology. Chapter 16 in *Arrow and the Ascent of Modern Economic Theory*, G. Feiwel, editor, pages 519–542. New York University Press, New York, 1987.

P. Dasgupta, P. A. David, Toward a new economics of science. *Research Policy*, 23:487–521, 1994.

P. A. David, Common agency contracting and the emergence of “open science” institutions, *American Economic Review*, 88(2), May 1998a.

P. A. David, Communication norms and the nollective nognitive performance of ‘invisible colleges’,” in *Creation and the Transfer of Knowledge: Institutions and Incentives*, G.Barba Navaretti et.al., eds, New York: Springer, 1998b.

P. A. David, The economic logic of ‘open science’ and the balance between private property rights and the public domain in scientific data and information,” in *The Role of the Public Domain in Scientific and Technical Data and Information: An NRC Symposium*, J. Esanu and P. F. Uhlir, eds., Washington, D. C.: Academy Press, 2003.

P. A. David, M. den Besten and R. Schroeder, How ‘open’ is e-science? In *e-Science '06: Proceedings of the IEEE 2nd International Conference on eScience and Grid Computing*, Amsterdam, v. Iss December 2006: 33ff.
[at:http://ieeexplore.ieee.org/iel5/4030972/4030973/04031006.pdf?isnumber=4030973âˆ=STD&arnumber=4031006&arnumber=4031006&arSt=33&ared=33&arAuthor=David%2C+P.A.%3B+den+Besten%2C+M.%3B+Schroeder%2C+R].

P. A. David and M. Spence. “Designing Institutional Infrastructures for e-Science,” Ch. 5 in *Legal and Policy Framework for e-Research: Realizing the Potential*, Brian Fitzgerald, ed., Sydney, Australia: University of Sydney Press, 2008.

P. A. David and P. F. Uhlir, Creating the global information commons for e-Science:Workshop Rationale and Plan. UNESCO, Paris, September 1-2, 2005.[Available at: http://www.codataweb.org/UNESCOmtg/workshopplan.html

P. A. David and P. F. Uhlir. Creating global information commons for science: An international initiative of the committee on data for science and technology (CODATA). Unpublised prospectus, 17 April 2006.

M. den Besten, P. A. David. Mapping e-science's path in the collaboration space. Paper presented to the conference on Oxford e-Research Conference 2008, 11-13 September 2008 [http://www.oii.ox.ac.uk/microsites/eresearch08/].

M. den Besten, P. A. David. Data Information Access in e-Research: Results from a 2008 Survey among U.K. e-Science Project Participants. OII research report X, 2008.

M. den Besten, P. A. David, R. Schroeder. Research in e-science and open access to information and data. In *The International Handbook of e-Research*, J. Hunsinger, M. Allen and L. Klastrup, Eds., New York: Springer Publishing, forthcoming in 2009.

J. G. Frey. Dark lab or smart lab: The challenges for the 21st century laboratory software. *Organic Research and Development*, 8(6):1024–1035, 2004.

J. Fry, R. Schroeder, and M, den Besten. Open science in e-Science: Contingency or Policy? *Journal of Documentation*, forthcoming in 2008.

S. Fuller, A Guide to the Philosophy and Sociology of Science for Social Psychology of Science, Ch. 17 in *The Social Psychology of Science,* W. Shadish and S. Fuller, Eds.(New York: The Guilford Press), 1994.

C. Hinds, M. Jirotka, M. Rahman, et al., Ownership of intellectual property rights in medical data in collaborative computing environments. In *First International Conference on e-Social Science*, 2005.

P. Kitcher, *The Advancement of Science: Science Without Legend, Objectivity without Illusions* (Chicago: University of Chicago Press), 1993.

K. Knorr-Cetina, *The manufacture of knowledge: An essay on the eonstructivist and contextual nature of science,* Oxford: Pergamon Press, 1981.

T. S. Kuhn, *The Structure of Scientific* Revolutions, First/ Second Edition (Chicago: University of Chicago Press), 1962/1970.

B. Latour, S. Woolgar, *Laboratory life* (Beverly Hills: Sage Publications), 1979.

M. Leenaars, e-Infrastructures Roadmap. *e-Infrastructure Reflection Group Technical Report,* 2005 [available at http://www.e-irg.org/roadmap/eIRG-roadmap.pdf].

L. Leydesdorff, *The challenge of scientometrics: the development, measurement, and self-organization of scientific communities*. Leiden: DSWO Press, 1995; 2nd Edition. Universal Publishers: uPUBLIS.com [available at: http://www.upublish.com/books/leydesdorff.htm].

J. Mairesse and L. Turner, Measurement and explanation of the intensity of co-publication in scientific research: an analysis at the laboratory level. Ch. 10 in *New frontiers in the economics of innovation and new technology*: *essays in honour of Paul A. David*. Eds., C. Antonelli et al. (Cheltenham, Eng.: Edward. Elgar. 2006. R. K. Merton. The normative structure of science [1942]. In

R.K. Merton, *The sociology of science: theoretical and empirical investigations*, Edited by N. W. Storer (Chicago: University of Chicago Press), 1973: pp. 267–278.

R.K. Merton, *The sociology of science: theoretical and empirical investigations*, Edited by N. W. Storer (Chicago: University of Chicago Press), 1973.

R. Schroeder. *Rethinking Science, Technology and Social Change* (Stanford: Stanford University Press), 2007 a.

R. Schroeder, e-Research Infrastructures and Open Science: Towards a New System of Knowledge Production?, in *Prometheus*, 25(1), 2007b: pp.1-17.

R. Schroeder. e-Sciences as Research Technologies: Reconfiguring Disciplines, Globalizing Knowledge, *Social Science Information*, vol. 47, 2, 2008: pp. 131-57.

S. Shapin, *A social history of truth* (Chicago: University of Chicago Press), 1994.

Stanford University, Openness in research, Iin *Stanford University Research Policy Handbook*, ch. 2.6, Stanford, CA, 1996; [Openness in research checklist at: http://www.stanford.edu/dept/DoR/C-Res/ITARlist.html].

P. F. Uhlir and P. Schröder. Promoting access to public scientific data for global science. *Data Science Journal*, 2006.