Oxford Internet Institute, Research Report No. 18, 5 January 2009

Data and Information Access in e-Research: Results from a 2008 Survey among UK e-Science Project Participants

Matthijs den Besten

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

Paul A. David

Stanford University, UNU-MERIT (Maastricht, Netherlands),
and All Souls College, University of Oxford
pad@stanford.edu

Questions concerning the actual extent of "openness" of research processes identified with contemporary e-science should 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 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. Therefore, our project designed a fuller, more specific set of questions (inspired by that source) to gather data about the issues of information access arising in the conduct 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.[1] 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.

[1] 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, the result of which are reported by den Besten and David (2008). The complete set of survey questions may be consulted in the Appendix of this paper (Figures 1, 3, 5-13).

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.

Contract terms and "open-ness in research": survey findings 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 UK 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.[2] 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

[2] 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.

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. Table 1 displays the results of a simple analysis of the responses obtained from participating researchers about these agreements.[3]

Table 1 Participants' reports in an inventory of e-science projects' contractual agreement terms and governance rules affecting "open-ness in research": 30 UK projects surveyed in 2008

Does the project agreement, or its internal governance rules:	Yes	No	Don't Know	Not Applicable	Responses
Restrict research participation (faculty, student, others) based on country of origin or citizenship?	3.7% (2)	77.8% (42)	9.3% (5)	9.3% (5)	54
Require research participation in EU-citizen-only meetings?	1.9% (1)	79.6% (43)	7.4% (4)	11.1% (6)	54
Prohibit the hiring of non-EU citizens to be involved in the proposed research?	1.9% (1)	75.5% (40)	9.4% (5)	13.2% (7)	53
Grant the sponsor a right of prepublication review for purposes other than the preparation of patents or the exclusion of proprietary data?	5.6% (3)	55.6% (30)	14.8% (8)	24.1% (13)	54
Provide that any part of the sponsoring, granting, or establishing documents may not be disclosed?	5.6% (3)	57.4% (31)	24.1% (13)	13.0% (7)	54
Contain language referring to or mandating compliance with government regulations restricting the export of certain materials or software programs?	5.6% (3)	57.4% (31)	22.2% (12)	14.8% (8)	54
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?	9.3% (5)	66.7% (36)	9.3% (5)	14.8% (8)	54

Source: Underlying data from the Oxford Internet Institute/OeSS Project Survey of e-Research Open Information Access. See Appendix Fig. 6, Question 8.

The overall impression one draws from these survey responses is, once again, quite broadly congruent with the impressions that Fry, Schroeder and den Besten (2008) report on the basis of their 12 in-depth interviews. That applies also in regard to their vagueness as to the way that their project's governing agreement(s) had been arrived at. More than one third of the projects' P.I. and non-P.I. members either could not or would not say whether it was the lead scientists, or university administrative and service offices, or funding agency staff that had framed the initial project agreement; nor could they say who—if anyone—subsequently had sought contract modifications, whether before or after the funding contract(s) had been signed and the project was launched officially. The latter aspect of the results predominantly reflects the reality that in many instances a university-based project's scientific

[3] The specific survey questions that are referred hereinafter by their number (in the text and notes) are reproduced in Appendix Figures (1-11), below. Unlike other survey findings discussed in this section (4), the results given in Table 1 and discussed in the following text are based on the complete tabulation of answers to the 7 items in survey Question 8 (Fig. 4) from all 54 survey respondents—including both the 3 cases of reports on multiple projects by a single P.I., and the 21 non-P.I.'s.

activities already are underway well before of the completion of the initial template of a legal agreement, let alone the signing of a contract. Furthermore, the responsibility for producing an agreement that will fund and govern the collaboration, typically will be in the hands of actors that are not directly engaged in the project or involved in any way with its scientific work: staff in the host universities' research services offices (sometimes their legal counsel's offices), or officers of public funding agencies, or both. When multiple partners are involved, the role of the funding agency in the formal framing of the project—and hence in the framing much of its governing agreement, tends to be augmented *vis-à-vis* that of both the academic host institutions and sponsoring business companies.[4]

With a few notable exceptions, involving restrictions on the uses of proprietary data and publication of findings (where a collaboration had industrial partners), the terms of the agreement governing their project about which respondent P.I.'s could respond were not such as would breach "openness in research" guidelines modelled on those of Stanford University guidelines (1996). Excluding the respondents who either found the question "not applicable" to their project or "did not know" the answer, between 96 and 98 percent of the replies reported that the terms of the agreement governing their project neither restricted research participation on the basis of country of origin or citizenship, nor required participation in EU-citizens-only research meetings, nor prohibited the involvement of research personnel from outside the EU.[5]

When asked whether their project agreement gave a sponsor the right of pre-publication review for purposes other than the preparation of a patent application, or the exclusion of proprietary data—i.e., the right to suppress findings that (presumably) were simply deemed "commercially sensitive"—92 percent among those replying definitively said "No." Although approximately one-quarter of all the respondents did not give a definitive reply because this was not applicable to their project (one may suppose there was no sponsor that would have such interest), 19 percent of those who accepted the question as relevant did not know whether to give a "yes", or a "no" answer. Almost as high a proportion (87 percent) among the definitive (yes or no) responses, reported that their project placed no restrictions on access to proprietary data that would have the effect of significantly blocking the work of a participating researcher. But, in the latter case, there was a considerably lower fraction of "don't know" responses (11 percent) from P.I.'s who accepted the question to be applicable to their respective projects.

The highest proportions of "don't know" responses were elicited by the questionnaire items concerning the existence of project contract terms and sponsorship agreements that were to be kept confidential, or provision that mandated project

[4] Funding bodies sometimes seek to form larger joint projects by bringing together academic that have submitted separate (competing) proposals, especially where there are opportunities to exploit differences the applicants' respective areas of special expertise. Where industrial partners are included as well, such multiparty agreements can become very complicated and require protracted, indeed tortuous negotiations. Whether that is the case, however, depends upon the nature of the collaboration tasks and the extent to which the resources being brought to it by the several parties can be organizationally partitioned, rather than requiring substantial joint-ness and actual co-mingling of the rights and responsibilities of the parties in regard to collaboration inputs and outputs. For further discussion of the formal legal context of collaborative e-science, see David and Spence (2008: sect. 2.3), and Fitzgerald (2008: Chs. 6, 11, 12).

[5] Among all the respondents who found these 3 questionnaire items (in Q 8) applicable to the circumstances of their respective projects, approximately 11 percent said they did not know the answer to the question. each of those questions.

compliance with government regulations restricting the export of material or software (deemed sensitive for national "defense" purposes). The latter represented between 26 and 28 percent of those respondents who did not dismiss these specific issues as irrelevant to the circumstances of their project. Of course, it is to be expected that quite a few participants would not be uninformed about contract provisions that were supposed to be confidential, *a fortiori* when a substantial share of them were not project P.I.'s. Nonetheless, among those who thought they could give a definitive answer to the question declared that their project's agreement contained no such restrictive provisions.

The survey results just reviewed suggest that these e-science projects 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. That a substantial fraction of project members appear not to be informed about the specifics of the project agreements under whose terms they are working 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. That more between 20 and 30 percent of participants remain uninformed about the details of contract terms that appear germane to the conduct of their research projects therefore 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.

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

The survey asked respondents asked (see App. Fig.3, Q.6) to classify their respective projects with regard to two taxonomic principles. Firstly, with which of the following functional scientific tasks was the project mainly engaged?: (1) generic tool development, (2) application development, (3) end-use application. Secondly, towards which among the main collaborative e-science forms was their project's work principally oriented to furthering? (i) grid access to distributed computing capacity, (ii) access to remote hardware instruments, (iii) access to specialized software, (iv) access to linked datasets or federated databases, (v) collaborative research with non-co-located teams. Although with these two axes and the resulting fifteen taxonomic combinations a more elaborate taxonomy may be constructed (den Besten and David, 2008b), for purposes of empirical analysis of the present small survey, the project classifications were collapsed into 3 broader purpose-engagement categories: (I) developing generic middleware tools for access to

distributed computing resources and instruments (8 projects), (II) combining application development with database resources (11 projects), and (III) combining end-use for collaborative research (7 projects). A residual category absorbed (4) projects characterized by mixed purposes and activities that resisted simple summary description. In the following, we therefore focus on findings relating to the project-purpose clusters that can be concisely labelled as (I) *middleware*-, (II) *database*-, and (III) *end-user community*-oriented.[6]

From responses to survey questions about measures actually undertaken to provide access to data and information relating to project results to researcher within the project and to outside researchers (specifically from [App. Fig. 7, Q.10; Fig.11, Q.13, and Fig. 11, Q.13]) it is possible to form some sense of the relative importance of these goals among the projects. What emerges is that when projects are grouped by main purpose category (I, II, or III), the distributions of responses differs noticeably from group to group. One simple measure of relative importance is the ratio for the group between "yes" (Y) responses, signifying that specific access-enhancing facilities were being provided, and "no" (N) responses.[7]

The overall pattern in this (Y/N) response ratio displays systematic variation along two axes. Along the first axis, there is rather wider attention to providing external researcher with information access, compared with concerns about within-project access provision by means of working paper and publication repositories, databases, and regular data-stream access. Thus, the external-vs-internal access differences in the Y/N ratio holds within each of the main project-purpose categories: for the 3.0 vs 1.0 for projects in the *database* group, 0.91 vs. 0.44 for those in the *end-user community* group, and 0.35 vs. 0 in the *middleware* group. These figures also display the second axis along which there is systematic variation: attention to providing information access (both to outside and to inside researchers) is relatively more widespread among the database projects, less so among the end-user community projects, and least evident among the middleware development projects.

The existence of a separate institution created by the UK e-Science project that is dedicated to improving robustness and distributing open middleware, namely the OMII (discussed previously), may well account for the latter feature of the pattern. That comparatively lower priorities appear to be attached to the internal provision of formal information access facilities among all 3 project-purpose categories, may well reflect the fact that only two-fifths of the survey responses pertain to projects that involved more than 2 consortium members, and another two-fifths of them had no other participating team. The management of inter-team information flows and data exchanges therefore may not be perceived among these projects as presenting major challenges.

Looking at the project start dates for the projects one may group these project into three cohorts whose relative sizes in the aggregate reflect the marked recent deceleration in the funding dynamics of the UK's e-science program as a whole: the pre-2003 cohort accounts for about 40 percent of the survey sample, the 2003–2004 cohort another 40 percent, leaving 20 percent in the post-2004 cohort. Within that temporal framework, something further may be said in regard to the specific

[6] Considering only the "classifiable" group of projects, their percentage distribution among the broad "purpose-engagement" clusters is seen to be: 31 percent with middleware (I), 42 percent with database applications (II), and 27 percent with end-user communities.

[7] In compiling the results reported in the text, counts of instances where respondents said the particular question was not applicable, or that they did not know, have been omitted.

information access repositories that have been provided, the extent to which projects providing them also require their members to deposit materials therein, and also about the trends in the diffusion of information management practices. From the following table (see below) it is evident that common repositories for projects' research outputs in the form of working papers and software code were established very generally from the inception of the e-Science program, but among the more recent of the three project cohorts (those launched after 2004) there has been some relative decline in the ubiquity of working paper depositories. On the other hand, comparison of the pre-2003 and post-2004 cohorts shows a rise in the proportion of projects that are providing common repositories and requiring the deposit of software code. In the case of data, however, common repositories are found only about half as frequently, and there is no evident secular movement on the part of projects that do provide them to also require that participants deposit their data.

It should be clear that access to the "common" repositories that are maintained by these e-science projects may be restricted in many ways, and it is therefore of particular interest to turn to the data about "open access" repositories that is displayed in the table's lower panels. One immediately sees that in the case of data there are essentially no "open-access" repositories in the sense in which that term is understood currently. The spread of institutionally maintained (department or university-wide) repositories for "OA publications" is noticeably strong, although there has been no increase in the proportion of cases in which participants are required to deposit material in them. The opposite pattern of change appears for pre-print repositories: their ubiquity has risen less markedly, but where these facilities have been set up, deposit requirements have become universal.

With regard to the various types of repositories for software, it seems clear that the proportion of open access repositories has approached the 30 percent share of middleware development projects in the total, and the relative frequency of adoption of version-control systems (with their archives) has more-or-less matched the relative share of middleware projects in the total—at least among the initial and most recent cohorts. Open access repositories for applications software have been established less frequently among the projects in the survey sample, but, where they do exist among the more recent product cohorts, the requirement mandating deposit of project-created computer code is widespread as is in the case among projects engaged in developing middleware.

What stands out most clearly from the findings reviewed in this section 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 UK 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 (as has been noted before by David, den Besten, and Schroeder (2006))—even through there has been an evolution away from the original expectations of open source release of these output under GNU General Public Licences once they had passed through the OMII's enhancement and repacking process.

Table 2. UK e-Science Projects' Common and Open-Access Repositories: 2008 Survey Responses from Principal Investigators

	Project start: pre-2003		**Project start: in between**		**Project start: post-2004**	
	Projects provid-ing (%)	**Projects providing that also require deposit (%)**	**Projects provid-ing (%)**	**Projects providing that also require deposit (%)**	**Projects provid-ing (%)**	**Projects providing that also require deposit (%)**
"Common" repositories for project-generated:						
(i) Working papers and memos	77	90	88	59	71	100
(ii) Software code	69	78	62	61	100	71
(iii) Data	38	100	25	48	43	67
"Open access" repositories for project-generated:						
(i) Publications (department- or university-wide)	36	60	50	50	71	60
(ii) Pre-prints	46	83	50	76	57	100
Source code for:						
(i) version-controlled development	23	35	12	100	29	48
(ii) middleware	23	100	25	100	29	100
(iii) applications software	15	53	25	100	14	100
Data	8	0	0	0	0	0

Source: Underlying data from the Oxford Internet Institute/OeSS Project Survey of e-Research Open Information Access. See Appendix Fig.7, Question 9.

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 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 Fig. 13, Questions 16, 17).[8] Among the projects 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

[8] 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.

expressed as well by one-third respondents from projects involved with *user-communities* and *database resources,* especially the latter group.[9]

The responses concerning "obstacles encountered by the project in achieving "openness" (see Fig. 13, 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.[10] 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.[11]

References

David, P.A. and Spence, M. (2008) Designing Institutional Infrastructures for e-Science. In: B.Fitzgerald (ed.) Legal and Policy Framework for e-Research: Realizing the Potential [Ch. 5] (University of Sydney Press: Sydney, Australia).

David, P.A., den Besten, M. and Schroeder, R. (2006) 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. Available 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

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

Stanford University, Openness in research (2006) In: *Stanford University Research Policy Handbook,* ch. 2.6 (Stanford University: CA). Available at: http://www.stanford.edu/dept/DoR/C-Res/ITARlist.html

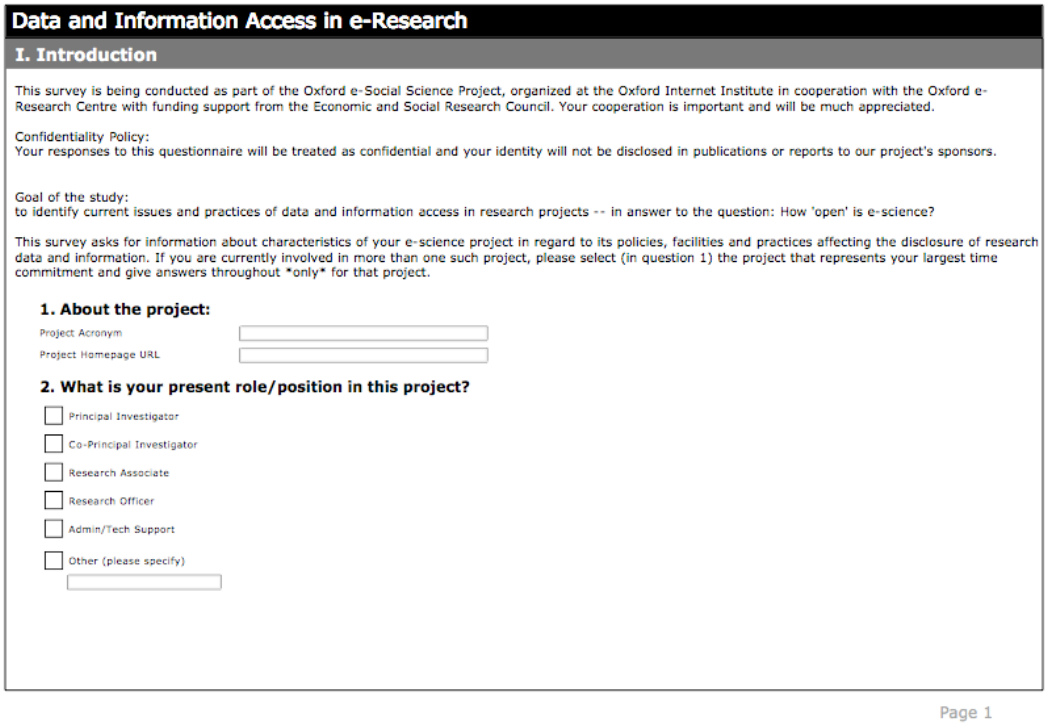
[9] Specifically, providing access to researchers outside the project was a significant concern for almost two-thirds of the data-centric projects and a third of community-centric projects.

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

[11] 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.

Appendix

Figure 1. Layout of questions 1 and 2.



Data and Information Access in e-Research

I. Introduction

This survey is being conducted as part of the Oxford e-Social Science Project, organized at the Oxford Internet Institute in cooperation with the Oxford e-Research Centre with funding support from the Economic and Social Research Council. Your cooperation is important and will be much appreciated.

Confidentiality Policy:
Your responses to this questionnaire will be treated as confidential and your identity will not be disclosed in publications or reports to our project's sponsors.

Goal of the study:
to identify current issues and practices of data and information access in research projects -- in answer to the question: How 'open' is e-science?

This survey asks for information about characteristics of your e-science project in regard to its policies, facilities and practices affecting the disclosure of research data and information. If you are currently involved in more than one such project, please select (in question 1) the project that represents your largest time commitment and give answers throughout *only* for that project.

1. About the project:

Project Acronym

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)

Page 1

Figure 2. Distribution of Respondent roles per project.

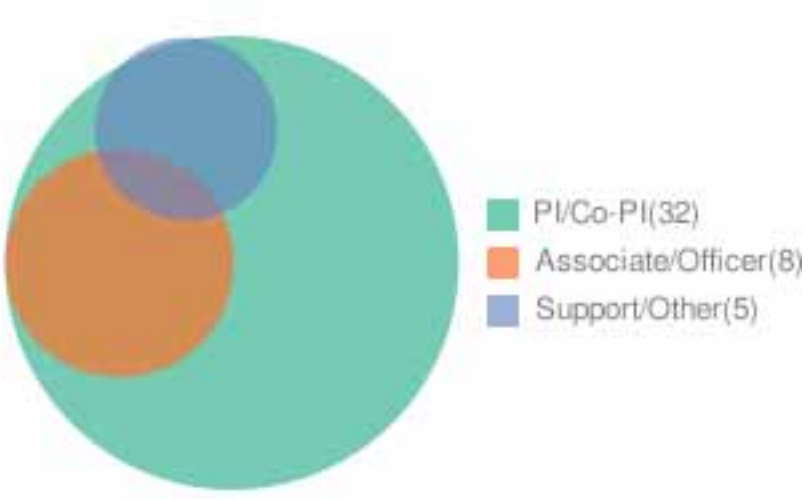
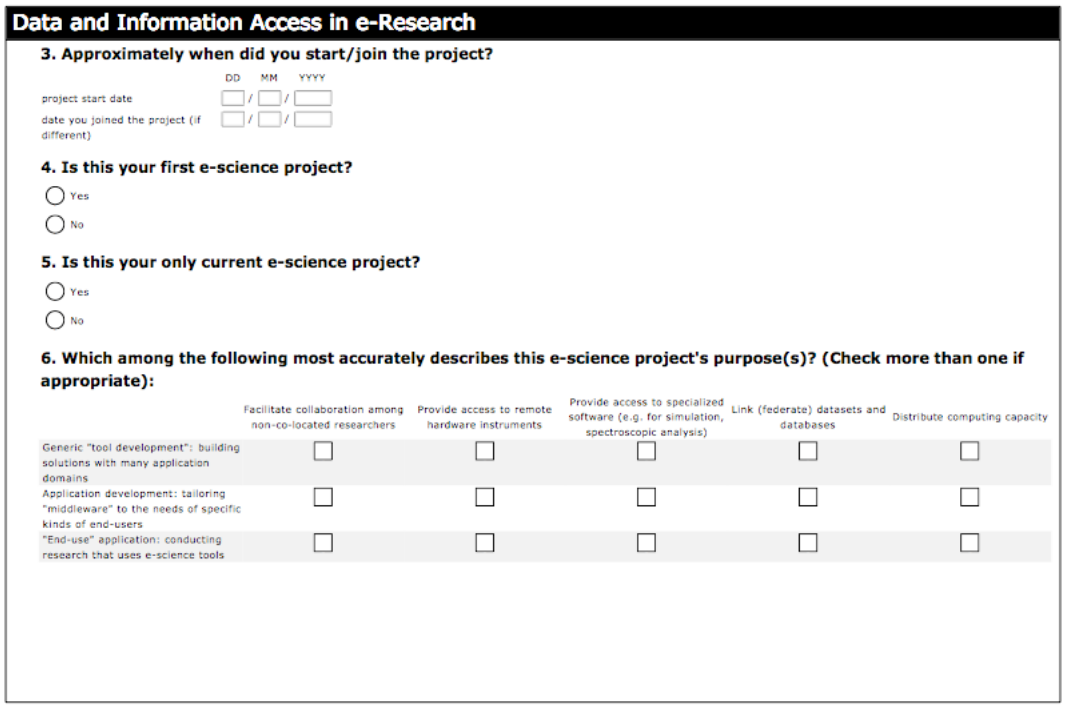


Figure 3. Layout of questions 3-6.



Data and Information Access in e-Research

3. Approximately when did you start/join the project?

	DD	MM	YYYY
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):

	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
Generic "tool development": building solutions with many application domains	☐	☐	☐	☐	☐
Application development: tailoring "middleware" to the needs of specific kinds of end-users	☐	☐	☐	☐	☐
"End-use" application: conducting research that uses e-science tools	☐	☐	☐	☐	☐

Page 2

Figure 4. Classification of answers relative to project start date.

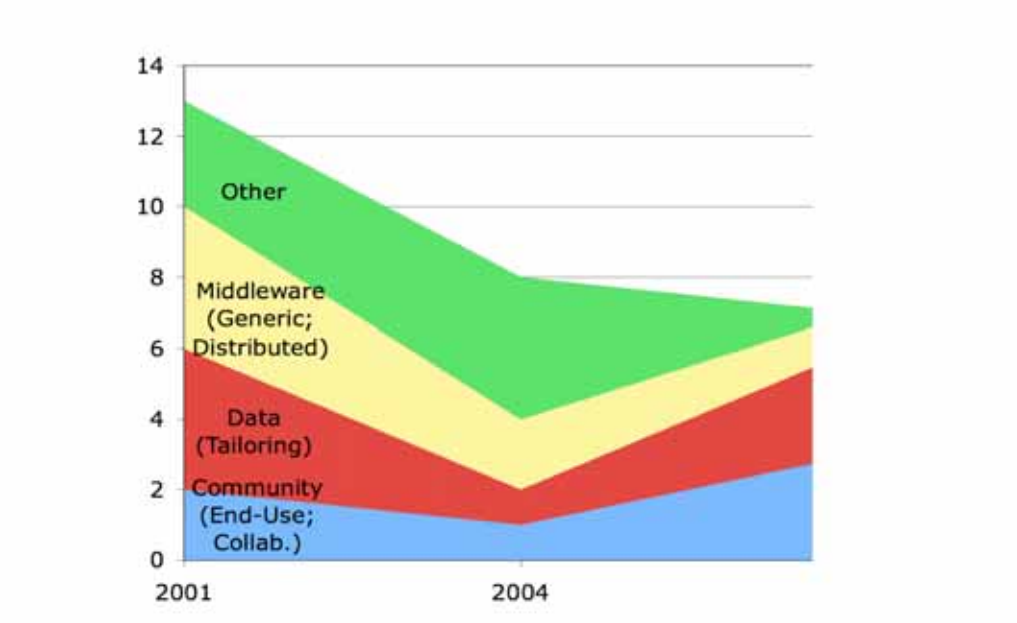


Figure 5. Layout of question 7.

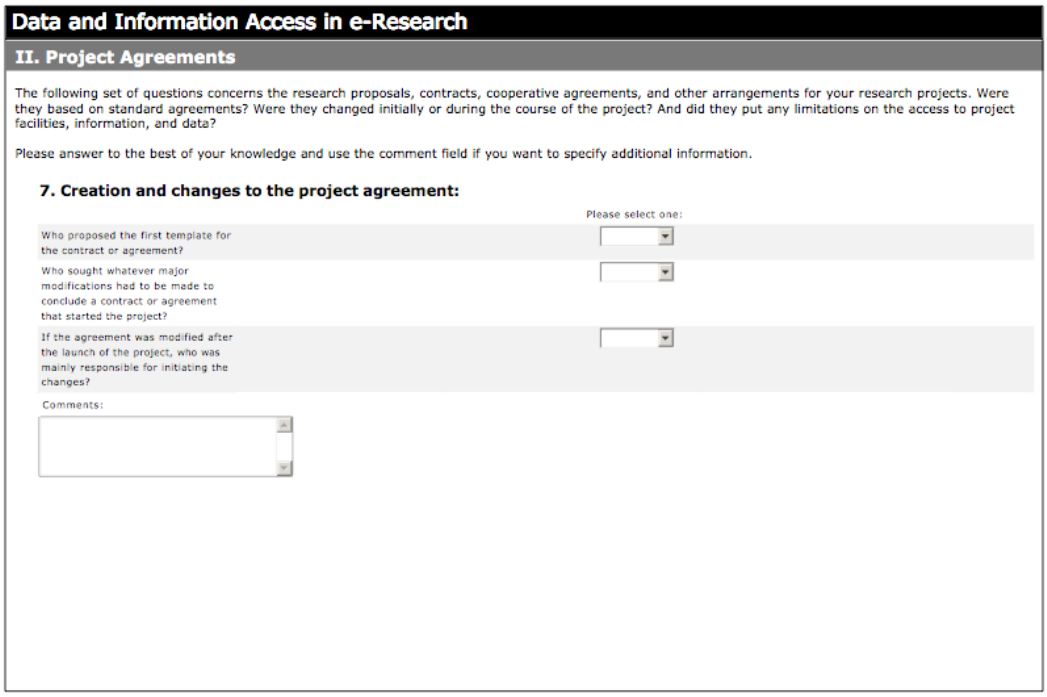


Table 3. Cross-tabulation of PI-responses to questions 7a and 7b.

		Who proposed the first template or for the contract or agreement?							
		Blank	**Don't know**	**Funding Agency**	**Industrial Partner**	**Not Applicable**	**Other**	**University Office**	**Total**
Who sought major modifications that were needed to conclude a contract or agreement that started the project?		**1**	0	0	0	**2**	0	**2**	5
	Don't know	0	**1**	0	0	0	0	0	1
	Funding Agency	0	0	0	0	0	0	0	0
	Industrial Partner	0	0	0	0	0	0	**2**	2
	Not Applicable	0	**1**	**6**	0	**5**	0	**1**	13
	Other	0	0	0	0	0	**1**	**1**	2
	University Office	0	0	**2**	0	**1**	0	**4**	7
	Total	1	2	8	0	8	1	10	30

Figure 6. Layout of question 8.

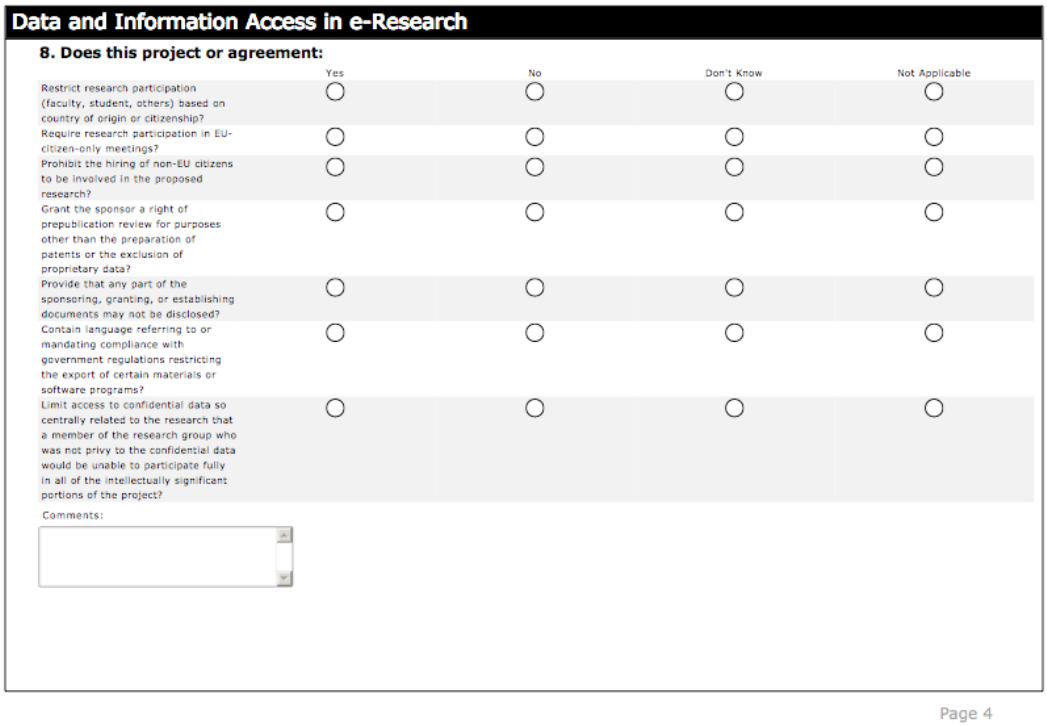
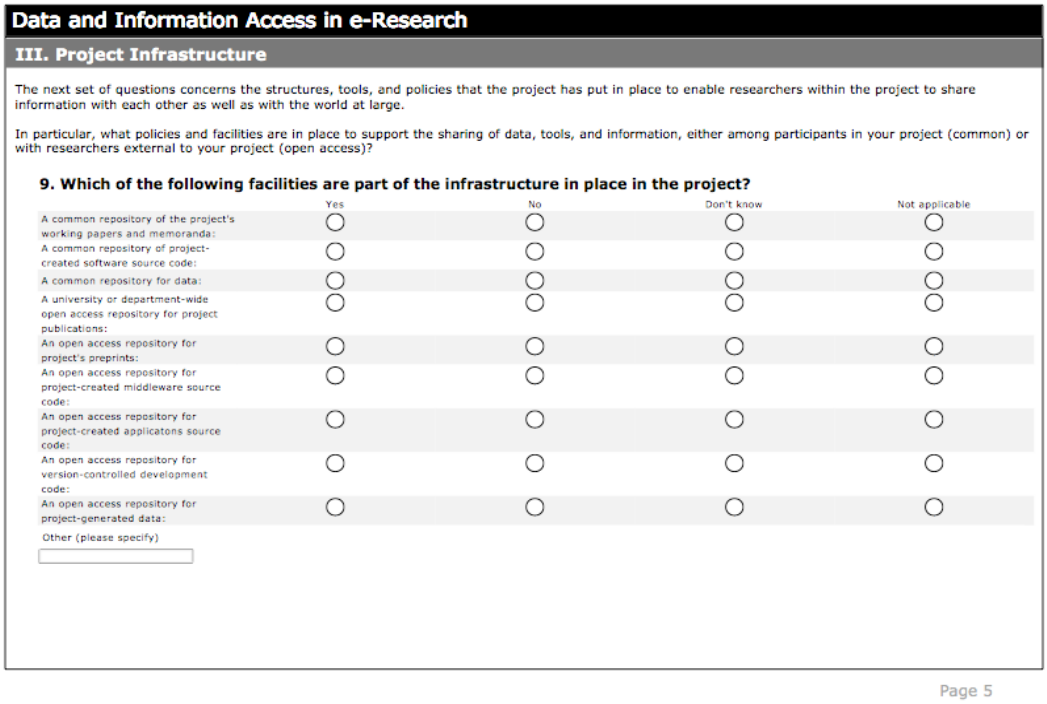


Table 4. PI-respondents' answer patterns to question 8.

n	Restrict	Require	Prohibit	Grant	Provide	Contain	Limit
1	Yes	Yes	No	No	Yes	No	Yes
1	Yes	No	No	No	Yes	No	No
1	No	No	No	Not Applicable	Not Applicable	No	Not Applicable
1	No	No	No	Not Applicable	No	No	No
1	No	No	No	Don't Know	No	Don't Know	No
1	No	No	No	No	Don't Know	Don't Know	No
1	No	No	No	No	No	Yes	No
1	No	No	Not Applicable	Not Applicable	No	Yes	Yes
1	No	No		Yes	No	No	No
1	Don't Know	Don't Know	Don't Know	Don't Know	Don't Know	Don't Know	Don't Know
1							
2	No	No	No	Not Applicable	Not Applicable	Not Applicable	Not Applicable
2	No	No	No	No	Don't Know	No	No
3	Not Applicable	Not Applicable	Not Applicable	Not Applicable	Not Applicable	Not Applicable	Not Applicable
13	No	No	No	No	No	No	No

Figure 7. Layout of question 9.



Data and Information Access in e-Research

III. Project Infrastructure

The next set of questions concerns the structures, tools, and policies that the project has put in place to enable researchers within the project to share information with each other as well as with the world at large.

In particular, what policies and facilities are in place to support the sharing of data, tools, and information, either among participants in your project (common) or with researchers external to your project (open access)?

9. Which of the following facilities are part of the infrastructure in place in the project?

	Yes	No	Don't know	Not applicable
A common repository of the project's working papers and memoranda:	○	○	○	○
A common repository of project-created software source code:	○	○	○	○
A common repository for data:	○	○	○	○
A university or department-wide open access repository for project publications:	○	○	○	○
An open access repository for project's preprints:	○	○	○	○
An open access repository for project-created middleware source code:	○	○	○	○
An open access repository for project-created applicatons source code:	○	○	○	○
An open access repository for version-controlled development code:	○	○	○	○
An open access repository for project-generated data:	○	○	○	○

Other (please specify)

Page 5

Table 5. Availability of repositories across projects according to PI-responses.

	Yes	No	Don't Know	Not Applicable	Blank
comprint	22	4	1	1	2
comcode	22	4	2	1	1
comdata	13	9	0	6	2
openprint	13	9	0	5	3
openpre	14	9	1	4	2
openmid	13	11	3	1	2
openapp	6	13	1	7	3
opendev	4	19	1	4	2
opendat	7	11	1	8	3

Figure 8. Layout of question 10.

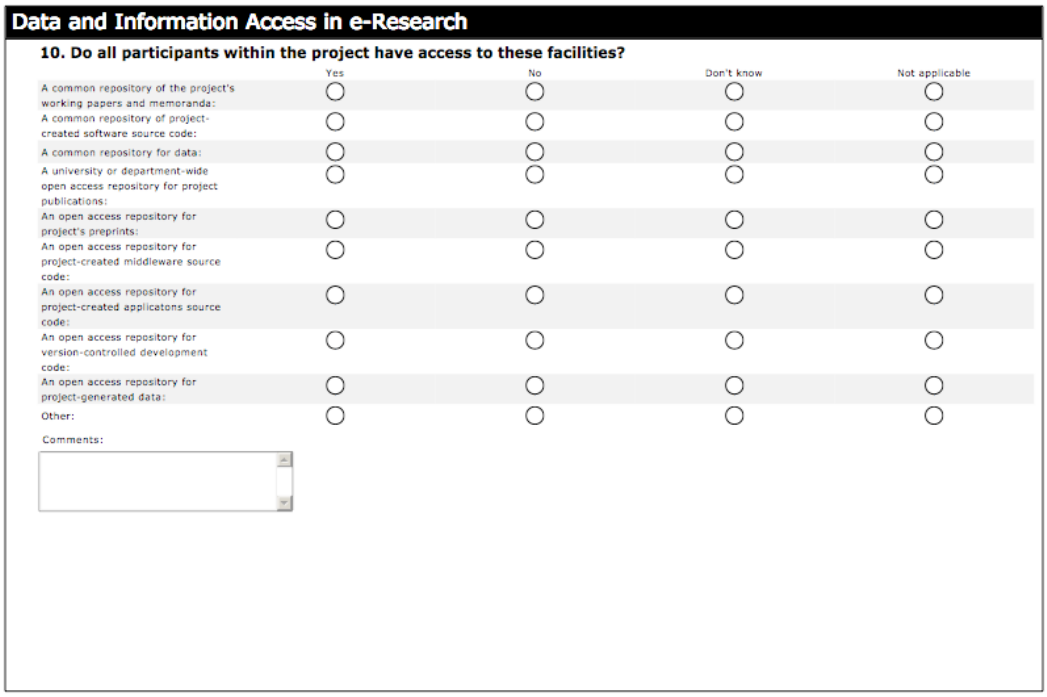


Table 6. Access to repositories across projects according to PI-responses.

	Yes	No	Don't Know	Not Applicable	Blank
comprint	23	1	0	2	4
comcode	22	0	0	4	4
comdata	11	2	0	12	5
openprint	15	1	0	9	5
openpre	15	1	0	10	4
openmid	12	2	1	10	5
openapp	8	2	1	14	5
opendev	6	3	1	15	5
opendat	7	1	1	16	5
other	1	0	0	12	17

Figure 9. Layout of question 11.

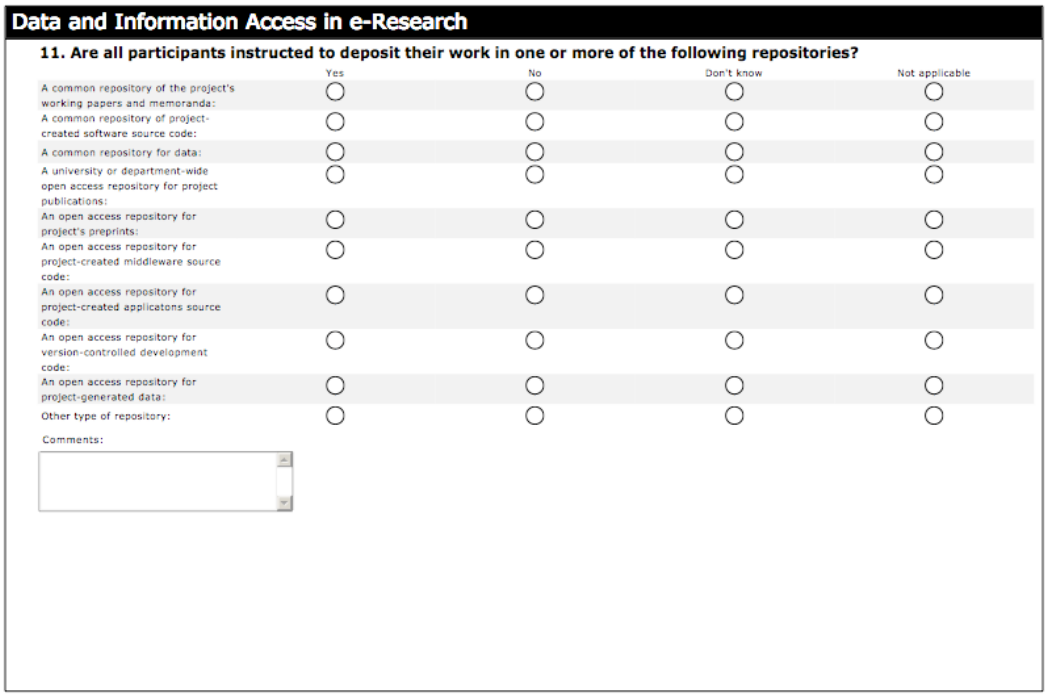


Table 7. Mandated depositing in repositories across projects according to PI-responses.

	Yes	No	Don't Know	Not Applicable	Blank
comprint	19	5	0	3	3
comcode	15	5	0	6	4
comdata	10	2	0	13	5
openprint	8	9	0	8	5
openpre	12	7	0	7	4
openmid	8	8	0	10	4
openapp	4	6	0	15	5
opendev	4	8	0	13	5
opendat	4	6	0	16	4
other	0	3	0	16	11

Figure 10. Layout of question 12.

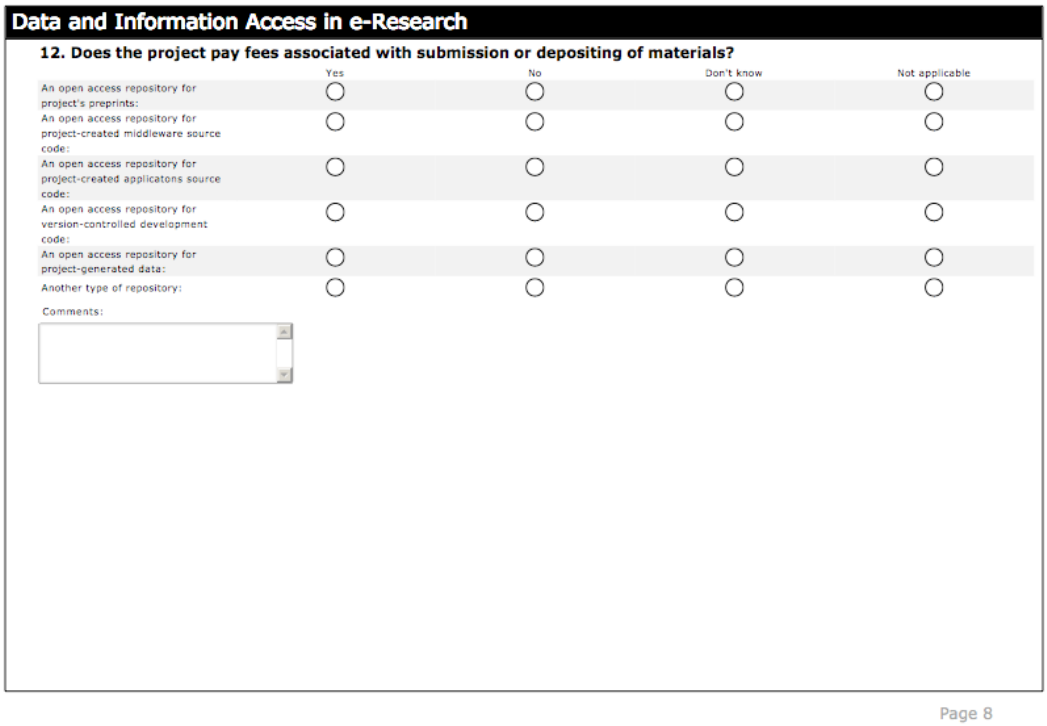


Table 8. Fee paying policies across projects according to PI-respondents.

n	preprints	middleware	applications	development	data	other
14	No	No	No	No	No	No
4						
2	No	No	No	No	No	Not applicable
2	No	No	No	Not applicable	Not applicable	
1	No	No	No	No	Not applicable	Not applicable
1	No	No	No	Not applicable	No	Not applicable
1	No	No	No	No	No	
1	No	No	Not applicable	No	Not applicable	Not applicable
1	No	No	Not applicable	Not applicable	Not applicable	Not applicable
1	No	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
1	No		No	No	No	No
1	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable

Figure 11. Layout of question 13.

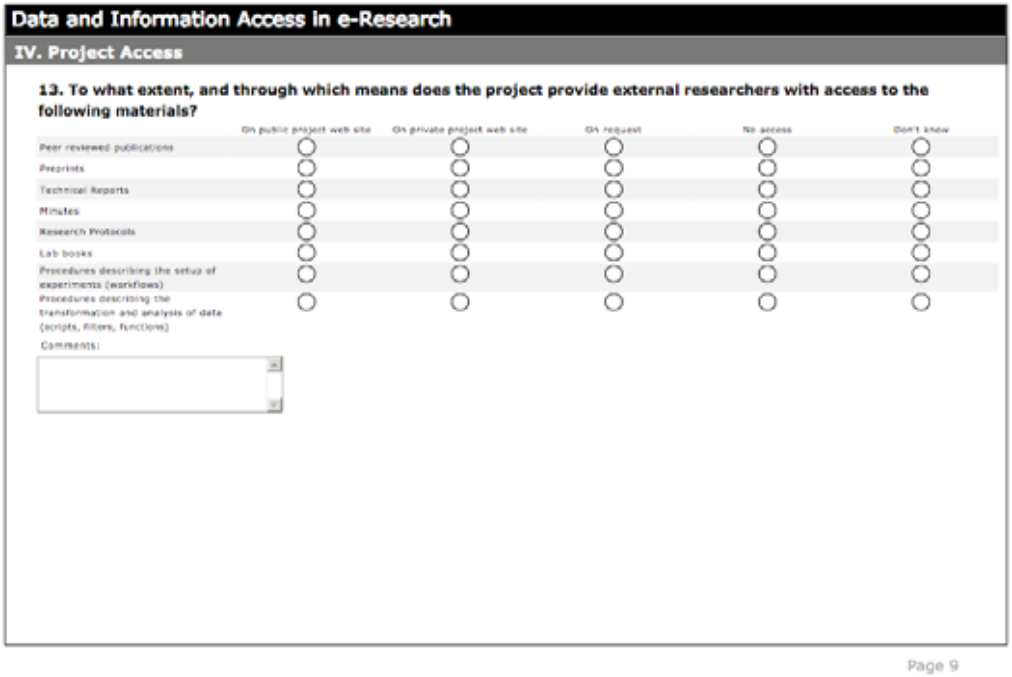


Table 9. Data and Information access across projects according to PI-respondents.

	On public project website	On private project website	On request	No access	Don't know	Blank
Peer reviewed publications	18	2	3	2	1	4
Preprints	15	2	4	3	1	5
Technical reports	19	2	2	3	0	4
Minutes	1	3	1	16	1	8
Research protocols	2	0	4	11	3	10
Lab books	0	0	1	17	1	11
Workflows	4	0	2	11	2	11
Scripts etc	3	2	4	9	3	9

Figure 12. Layout of questions 14 and 15.

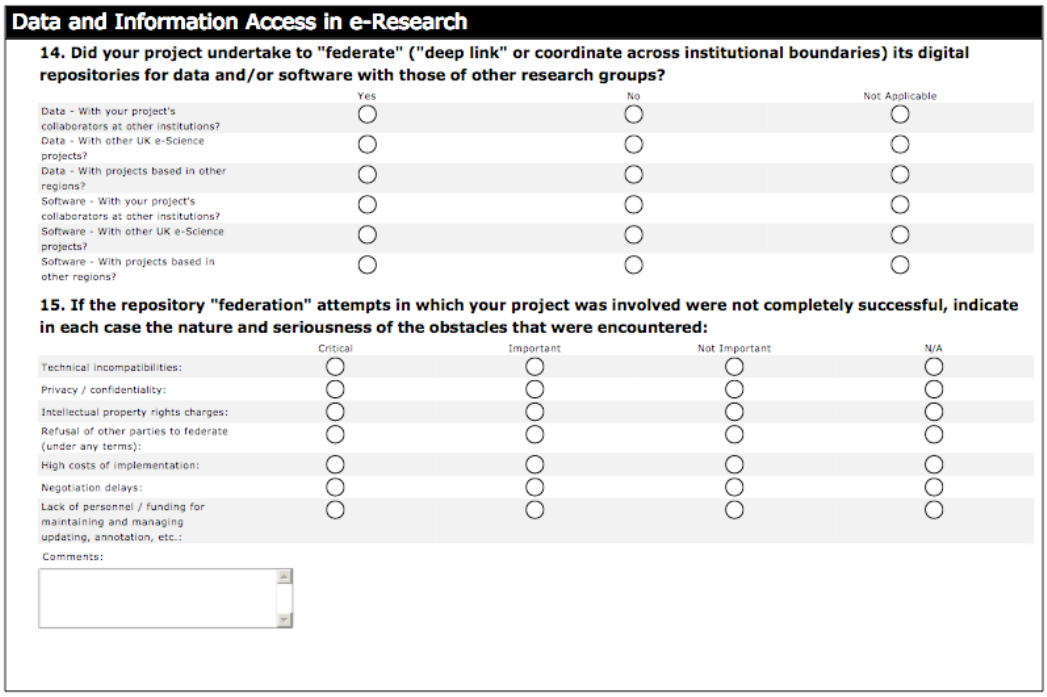
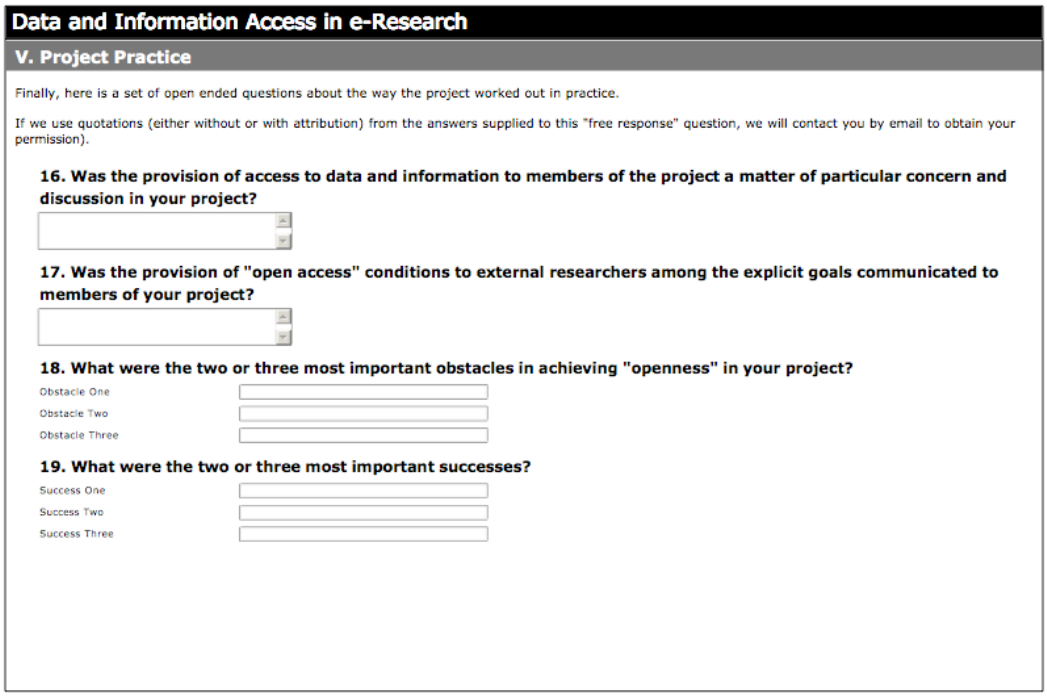


Table 10. Number, nature, and severity of obstacles that project run into when they attempt to federate resources; the answers critical, important, and not important correspond to scores of 2, 1, and 0, respectively.

		Data			Software		
		within project	within UK	outside	within project	within UK	outside
Fed. attempts		11	6	7	10	7	5
avg. no. obstacles		2.36	1.83	1.71	3.3	2.57	0.8
No. ticked	**technical**	6	3	4	7	4	2
	privacy	4	2	2	5	3	1
	property	3	1	1	4	2	0
	obstinance	3	1	1	4	2	0
	cost	3	1	1	4	2	0
	delays	3	1	1	4	2	0
	resources	4	2	2	5	3	1
Total Score	**technical**	6	3	3	5	3	2
	privacy	3	2	2	3	2	1
	property	1	1	0	1	1	0
	obstinance	0	0	0	0	0	0
	cost	4	1	1	3	1	0
	delays	2	1	1	0	1	0
	resources	6	3	2	7	5	1

Figure 13. Layout of questions 16-19.



Data and Information Access in e-Research

V. Project Practice

Finally, here is a set of open ended questions about the way the project worked out in practice.

If we use quotations (either without or with attribution) from the answers supplied to this "free response" question, we will contact you by email to obtain your permission).

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?

Obstacle One
Obstacle Two
Obstacle Three

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

Success One
Success Two
Success Three

Page 11

Table 11. Cross-tabulation of coded PI-respondents' answers to questions 16-19.

Internal access of concern:		**No**		**Yes**	
Outside access of concern:		**No**	**Yes**	**No**	**Yes**
Obstacles:	**0**	5	3	1	0
	1	0	2	0	0
	2	1	1	2	1
	3	0	1	1	2
Successes:	**0**	2	3	0	0
	1	1	1	1	0
	2	3	1	1	1
	3	0	2	2	2

Table 12. Coding of answers to questions 16 and 17.

	Yes	No	Neither
Internal	Discussion, not concern; It was a concern in that it was paramount that all members have access to the data and information; No it was assumed open from the outset; No, we had an open policy to data and information in the project. The only restrictions being those placed on the data by licence issues.;Not especially. We set up a Subversion repository as a matter of course. The project was about data and information access ...; There were issues of commercial confidentiality in that we were linked with a commercial company who gave us data we could use for the project but which was confidential; We used the SRB successfully. People were keen to share data.; Yes - using patient medical data so confidentiality of data was important; Yes but as most researchers were bases in Southampton it was not as serious as it could have been; Yes, with an industrial partner, information sharing was important.; Yes. Communication/availability of document versions and code critical and managed via a CVS system; Yes. This was fundamental from conception.; Yes. We used a great deal of personal data from Sillitoe's ethnographic research in Papua New Guinea. He was very concerned that no individuals could be identified from the publicly available data. Similarly, my own data, from Punjab, Pakistan, had previously been coded with several categories of sensitivity so while there was no need to implement a new protocol for dealing with that data, it was important that those restrictions remain in place. ; yes; yes, well-defined protocols were established at the start of the project, and implemented rigorously	No; No. Everything in our project is open-source with the possible exception that some work is performed on the boundary with a particular experiment collaboration and that might be restricted to members of that collaboration.	Do not understand the question. See previous comment; N/A; No. The project was of sufficiently small scale to permit this kind of activity, when it was required, to take place informally.; Project was just focused on feasibility and did not move to the data stage; no, but we were a small project with just four participants at any given time.
External	At this moment all the material (papers, data, software) is open to a large (of the order of a hundred members) European scientific community involved in the same European project. Alter launch the material will be made available to everybody.; It would be unusual for research in our area (theoretical computer science) to be restricted in any way.; No it was obvious; Not in the initial stages, at least as OA has now become viewed but open availability to data corresponding to published material was a major plank of the original proposal; The project was to produce a smart data warehouse for public access to cross-species anatomical and associated data; We work with two international projects EGEE and LCG and to enable this work to happen documents have to open access; Yes; Yes—see earlier comment.; Yes, everyone was expected to use the software repositories, wiki etc.; Yes.; Yes. This was required as members of the Worldwide LHC Computing Grid.; Yes: We established our open-source policy early in the project.; yes	Codes for the construction of Vision Recognition software were kept in-house but the software is freely available on the NCeSS website.; NO; No; No - and it wasn't the purpose of the funding; No.; No. When we started escience was in such a primitive state that we were feeling our way forward. The eventual toolset we developed could not have been properly anticipated.; They can have access to some aspects of the source code but other areas were restricted because of commercial necessities; We were not concerned about open access to all datasets (though obviously we needed some level of openness in order to demonstrate aspects of the project to one another). We put in place tools for anonymizing personal data and permitting certain properties of the data to be freely available without compromising individual privacy.; no; only the final dBase will be public - everything else is private	;N/A; Not known; n/a

Table 13. Coding of answers to questions 18 and 19.

Obstacles	Anonymization that didn't render the data meaningless; DISPARITY OF SECURITY TECHNOLOGIES (ATHENS, SHIBBOLETH, GSI); Ensuring data security; Ensuring data was open but that security of the infrastructure was not compromised; Failure to adequately understand what we wanted to do with the middleware once we had developed it.; Finding the right tools to make openness easy and not an additional burden.; Industrial sponsor's contract; Industrial vs Academic goals; Industry confidentiality requirements; Industry does not want material to be public; Limited vision of team members and thereby not understanding that you can achieve more in a team than on your own; Need to protect commercial IPR; Opinions; People; Poor organization of source code; Propensity to use private email lists. Very hard to make folk use open email lists.; Some difficulty in communicating with researchers in more standard types of e-science; Technical issues e.g. software bugs/incompatibilities; Technical: software development was not a primary goal, so we had limited resource to \"deep-link\" software; costs of implementation; lack of personnel to maintain web site; met office sensitive to data access; working out the how to present the tool; IPR; Lack of funding; Minor: culture clash across the UK e-Science programme; Sheer volume of material produced meant we made stuff public even we couldn't find later ...; Technical/Financial; Time; Training collaborators
Successes	\"Learning from failure\" i.e. project was too technologically ambitious; \"Outsourcing\" to existing open software projects. e.g. Scientific Linux; Acceptance by and integration into the OMII software stack is a currently emerging success.; Advancement in Grid technology; Collaborative outputs across disciplinary borders (computer science & biology); Community building.; Deploying Trac as a combined wiki, bug tracker, subversion client.; Developing an understanding of how to represent and manage data within a grid computing environment; Developing and deploying part of the largest Scientific Grid in the world; Easy access to webserver and support; Everything except the code is in one place; Higher visibility; INTEGRATION OF DISPARATE TECHNOLOGIES; Implementation of new technology; It produced a new form of database structure; RAVE is being adopted by OMII-UK; Raise the importance of metadata for chemistry; Software is available on GridEngine community web site; Software widely downloaded and used by a number of project; Successful and harmonious collaboration; Supportive end user; User-led requirements; Very good collaboration within the project; We are active in most areas of research in Grids in both LCG and EGEE; Working software for dynamic reconfiguration; all results and code published openly; cross-platform support; delivering on-time and on-specification!!; first-time provision of a data bank for multi-modal corpus interactions; getting met office data out to community; some important cross-species data; the cloudtag tool; A common goal; Capacity building within project team; Continued support via OMII funding for public release; Introducing standards based data access methodologies to the atmospheric and oceanographic sciences; Publication of the above results; Remote Access and Monitoring of Chemical Experiments; Technically excellent and mutually supportive team members; Tools for using XML for data representation; Unexpected application to natural language dialogue; a cleaver website; and won a prize for best paper in a conference; tools for the integration of disparate data types (audio, video, image, text)