PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Data sharing, small science and institutional repositories

Melissa H. Cragin, Carole L. Palmer, Jacob R. Carlson and Michael Witt

Phil. Trans. R. Soc. A 2010 **368**, doi: 10.1098/rsta.2010.0165, published 2 August 2010

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2010) **368**, 4023–4038
doi:10.1098/rsta.2010.0165

Data sharing, small science and institutional repositories

By Melissa H. Cragin[1,*], Carole L. Palmer[1], Jacob R. Carlson[2] and Michael Witt[2]

[1]*Graduate School of Library and Information Science, University of Illinois at Urbana–Champaign, Champaign, IL, USA*
[2]*Purdue University Libraries, Purdue University, West Lafayette, IN, USA*

Results are presented from the Data Curation Profiles project research, on who is willing to share what data with whom and when. Emerging from scientists' discussions on sharing are several dimensions suggestive of the variation in both what it means 'to share' and how these processes are carried out. This research indicates that data curation services will need to accommodate a wide range of subdisciplinary data characteristics and sharing practices. As part of a larger set of strategies emerging across academic institutions, institutional repositories (IRs) will contribute to the stewardship and mobilization of scientific research data for e-Research and learning. There will be particular types of data that can be managed well in an IR context when characteristics and practices are well understood. Findings from this study elucidate scientists' views on 'sharable' forms of data—the particular representation that they view as most valued for reuse by others within their own research areas—and the anticipated duration for such reuse. Reported sharing incidents that provide insights into barriers to sharing and related concerns on data misuse are included.

Keywords: data sharing; small science; institutional repositories; scientific data curation; cyberinfrastructure

1. Introduction

If cyberinfrastructure is 'principally *about* data: how to get it, how to share it, how to store it, and how to leverage it' for scientific discovery and learning (Edwards *et al.* 2007, p. 31), then advancing cyberinfrastructure is dependent on our understanding of how to support data practices and needs. Sharing is at the heart of success, as collecting, storing and making use of data can only come after the means for sharing are in place. In sciences served by disciplinary or nationally scoped infrastructure initiatives, sharing research data is considered to be an inevitable trend. However, unlike fields such as physics and astronomy that tend to have standard data practices, data sharing in 'small science' is not common or expected. It functions largely as a cottage industry in which data are exchanged based on professional relationships and personal communication. Nonetheless, over the long term, small-science researchers, who span many fields

*Author for correspondence (cragin@illinois.edu).

One contribution of 15 to a Theme Issue 'e-Science: past, present and future II'.

and produce many different forms of highly valuable data, are expected to produce more data than researchers in big-science fields (Carlson 2006). These are also the scientists who are increasingly turning to their university libraries and institutional repositories (IRs) for assistance with their data problems. In response, IRs at many universities are now providing support for primary research data with varying architecture and service-model implementations (Choudhury 2008; Witt 2008; Wong 2009).

Although it is not expected that all academic institutions will provide IR or curation services or develop them in a uniform way, those that do will have a broad constituency with diverse data practices and needs. The research reported here provides empirical evidence of the kind of variations that will need to be accommodated by IRs to support small-science researchers. The unique qualities of researchers' data and their sharing requirements will necessitate tailoring of data curation services, while keeping development in alignment with the growing, global e-Science and curation infrastructure. At present, a number of research and development efforts are focused on solutions for small-science data (e.g. Karasti *et al.* 2006; Borgman *et al.* 2007; Rice 2009), but, in practice, management remains ad hoc (NSB 2005); descriptive standards, when they exist, are often not applied, and most data remain concealed locally and undiscoverable (Heidorn 2008).

The Data Curation Profiles Project (Witt *et al.* 2009) is investigating what data researchers are willing to share, when and with whom, and researchers' needs and requirements for sharing those data through an IR. Following on recommendations from the Research Information Network (RIN 2008), the project addresses the need for 'full account(s) of the different kinds of data that researchers create and collect, …of the significant variations in …behaviors and needs in different disciplines, sub-disciplines…; and make clear the categories of data that they wish to see…shared with others' (RIN 2008, p. 17). We focus on data types and subdisciplinary practices shown by Pryor (2009) to essential factors in the development of policy requirements for data sharing. This study contributes by including several disciplines not covered in the RIN study and also by elaborating further the specific data 'forms and varieties' scientists are willing to share. Specifically, we identify scientists' self-reported practices and views on 'sharable data'—data forms that are requested most often by others, but also those that are considered to have the most scholarly value, especially for reuse over time, and implications for curation services are considered.

(*a*) *Evolving small science*

Small science has traditionally been characterized as hypothesis-driven research led by a single principal investigator, in which progress and reward are contingent on generating and analysing one's own data. Research funding can be limited, and the day-to-day conduct of research is often dependent on a few graduate students (Peterson 1993), who carry out much of the data collection, and manage and process these data during the course of a project. In small science, data-management systems tend to be ad hoc, and if data standards exist, they are rarely applied (Heidorn 2008). However, these arrangements are not always static, and small science is expanding in two marked ways. For some laboratories, the traditional organizational configuration is extending into more community-oriented research networks. The other shift is the emergence of data-intensive science in some subdisciplines that continue to exhibit the traditional structure.

For many scientists, the traditional organizational structure tends to evolve into medium- or large-scale collaborative science, as described by Peterson (1993) from his own experience working in the environmental sciences. As a research group gets larger and more formally connected to other research groups, it begins to function more like big science, which requires production structures that support project coordination, resource sharing and increasingly standardized information flow. Of course, such changes in the localized organization and production of science are part of the long-recognized growth of science at large (De Solla Price 1963). What distinguishes the current state of the enterprise are the rapidity and amount of data generated by new methods and instrumentation, and an increasing focus on research questions that require interdisciplinary problem solving (e.g. Pachura & Martin 1991), and the use of the Web for communication and collaboration. These novel conditions alter the markers of traditional small science, especially in terms of data production and use, but they also allow and encourage many scientists to function more multi-modally, as illustrated by the participants enrolled in the Data Curation Profiles Project. Although all participating scientists work as solo Principal Investigators (PIs) on their own focused research, several were also working on larger collaborative projects. They were all consistently producing streams of digital data, and some were producing very large datasets relative to the size of their laboratory. As is increasingly the case, one participant was heavily involved in data management and sharing at the research community level.

Thus our conceptualization of small science does not exclude research that might, in earlier decades, have been considered large, certainly in terms of methods of data production and analysis, or the extent of data generated or analysed. However, the focus here is still on scientists who generate or gather data into privately held sets or collections that they analyse locally. Although most of these scientists continue to conduct hypothesis-driven research, some are developing or applying data-driven methods, such as earth and atmospheric scientists, who are developing large databases for computational modelling and the horticulturist who is mining databased genomic material for patterns. (For an example of data-driven approaches in the small-science context, see Murray-Rust 2007.) Still, these research communities tend to be heterogeneous in the methods and data types applied, without uniform or widely applied data standards, and are not currently well supported by disciplinary repository services. It is anticipated that these scientists will require access to a wider range of curation services to support deposit of data into shared repositories.

2. Disciplinary distinctions and Data Curation Profiles

There are two concerns with the current body of knowledge on data practices and curation needs in the academic context. First, the continuing use of high-energy particle physics and/or astronomy as a base of reference and comparison obscures the complexities of production and communication inherent in small science and across subdisciplines (e.g. Chompalov *et al.* 2002). Second, the focus on universal solutions for curation activities often applies to common and high-level service models. However, this does not generally account for research that functions at a more community level, and scientists' willingness to exchange data and

share publicly, which is influenced by differing research and publishing practices, data-generation methods and existing data infrastructures (Pritchard *et al.* 2005; RIN 2008). Many of the services and infrastructures implemented to support curation activities will necessarily be designed for general use—that is, it will not matter which academic discipline or university department the data come from; preservation workflows, for example, will be based on a set of general requirements and selection and appraisal criteria. However, identification of these practices, norms and conditions at the subdisciplinary level is necessary for developing effective data services that support research processes in academic institutions.

The success of new information and communication technologies is associated with their ability to support the research practices and culture of the target scholarly community. 'The range of technological possibilities may change rapidly, while the world views that dominate specific scientific communities are likely to change much more slowly' (Kling & McKim 2000, p. 1307); while communication practices will change over time, they 'are durable in the medium term'. Although data practices are generally masked during the publication process, they are an integral part of scientific communication and, therefore, subject to the same social and organizational constraints that shape disciplinary differences in the development and adoption of other communication practices and systems. The challenge is to determine which disciplinary and subdisciplinary distinctions need to be attended to in shaping curation requirements and services for IRs. Thus, our aim, in analysing who is willing to share what and when, is to associate vital differences in data types and practices with specific research communities.

Data Curation Profiles (Witt *et al.* 2009) support analysis of data characteristics, production and communication to capture the distinctions that make a difference for curation of research data. Drawing on this study and our previous work, table 1 provides a comparison of crystallography and geobiology, presenting features of data types and practices specifically documented in the data-profiling process. It represents a segment of the broader range of socio-technical conditioning variables covered by a complete profile and that need to inform the planning and implementation of IR-based data handling policies and curation services. Other variables included in a full profile that can impact data practices and IR services include organization and description of the data targeted for ingest, needs for discovery services and access control (http://wiki.lib.purdue.edu/display/dcp/Data+Curation+Profiles; Witt *et al.* 2009).

The 'type' row holds specific representations that show how data are transformed over the course of research, with indications of the sharing and reuse value of different forms. The case of crystallography is particularly interesting because of the dependency relationship with chemistry. As data producers, crystallographers assign high value to the 'raw' image file as that which holds the most complete information. As data consumers, chemists place higher value on crystal structures abstracted in the Crystallographic Information File (CIF; Hall *et al.* 1991); although it may represent only a small part of the raw data, the CIF is preserved via deposit into disciplinary repositories upon paper publication. The second, third and fourth rows, 'format', 'size' and 'accessibility', underscore the heterogeneity of the data types that IR services will encounter in the academic setting. The 'intellectual property/ownership' (IP) row lists examples of disciplinary contexts that may constrain or compel sharing and access. For example, when crystallography data are produced as a service, there is

Table 1. Selected profile elements.

data characteristics	crystallography	geobiology
types and value	'raw data':[a] most information-rich, with long-term value for reuse 'CIF' text file: most commonly shared data type	'reduced spreadsheet' has averaged values for multiple observations: most often requested by others
format	binary data: image CIF: text field-wide standard	EXCEL spreadsheet
size	each image or 'frame' = 0.25–1 MB set is approximately 2400 frames (approx. 1 GB) text file >500 kB	< 1 MB
accessibility	field-wide repositories many journals require deposit of CIFs OAI-PMH[b] tools becoming available for CIFs	difficult and ad hoc direct requests for data, often based on publications
intellectual property	service model: crystal structure solutions provided to chemists; ownership of the data is ambiguous, requiring negotiation before data 'hand-off'	variable depending on source of funding: primarily government and industry; ownership and rights vary from full to very limited; long-term embargoes common

[a]The table contains only one (geobiology) or two (crystallography) of the data forms generated during the research process.
[b]Open Archives Initiative Protocol for Metadata Harvesting (http://www.openarchives.org/pmh/)

an opportunity for the crystallographer and the chemist to negotiate control and release of the data. In the case of this subarea of geobiology, however, IP control is very much tied to the funding agency supporting the research, with industry generally more restrictive and government tending to encourage public access. The differences between crystallography and geobiology are clearly significant for curation services, but are also essential to understanding the varying uses and values of data forms, as seen in the case of crystallography. Data dependencies and relationships among different research communities form data communities that are meaningful for how data collections are developed and organized.

3. Methods

This Data Curation Profiles study was designed to investigate a range of disciplines and different forms of data that might be covered by curation services within an academic IR context, applying a disciplinary-practices approach as a means to compare how data-related scholarly activities vary among disciplines and research communities. The analysis of scientists' practices reveals how research work is performed and conditions within the disciplinary culture which influence that work. Both the practice and the related conditions need to be

Table 2. Research fields represented by the participants.

university 1	university 2
agronomy and soil science	agronomy and soil science
biochemistry	anthropology
biology	earth and atmospheric sciences
civil engineering	geology
earth and atmospheric sciences	horticulture/plant sciences
electrical engineering	kinesiology
food sciences	speech and hearing
horticulture/plant sciences	

understood for the development of technical capabilities and policies, including those that guide longer term management of multi-disciplinary data repositories (Palmer & Cragin 2008; M. H. Cragin 2009, unpublished data).

Data were collected through staged interviews and structured worksheets from a convenience sample of 20 scientists who conduct small-science research and have an interest in data management or sharing. The sample included researchers whose work is data intensive and who generate large digital datasets, as well as those who generate multiple kinds of data. Participants from two large research universities were recruited by investigators or librarians who already had professional relationships with some of the scientists. A small number of participants who were not previously known to the investigators were invited to participate based on referral from another participating scientist. Recruiting was guided by two main objectives: disciplinary coverage and targeted disciplinary overlap for purposes of comparison. The participants represented 12 disciplines, with one scientist from each field, except for the following disciplines where broader cases were developed: agronomy and soil science (five); anthropology (three); earth and atmospheric sciences (two); geology (three); and horticulture and plant science (two) (table 2).

Following Institutional Review Board approval for human subjects research from both sites, two stages of interviews were conducted using structured worksheets as a technique for focusing participants' attention on data issues. In the first stage, a Pre-interview Worksheet was distributed prior to the interview asking the participants to identify their research area and to describe two recent or on-going projects 'from the perspective of the data'. The interview sessions that followed were semi-structured and ranged from 60 to 120 min. The second stage included a Requirements Worksheet as part of the interview session, designed to gather details about curation needs and requirements for the types of 'sharable' data identified in the first interview. Follow-up interview questions were also customized based on first-stage interview results. Data and materials for each participant have been assigned a participant code, and any quotations or paraphrased quotes used hereafter will be referenced with that code and date of the interview.

All interviews were recorded and fully transcribed. Participant codes are used to manage interview files, and then also used to identify the source of quotes from those transcripts; for example, [U2GEO1B1_5_08_08] denotes a geologist

interviewed on 8 May 2008; quoted language in this text has been normalized for readability, such that 'um' and 'uh' have been left out, and vernacular terms modified (e.g. 'gonna' has been replaced with 'going to'). The analysis began with a categorical scheme developed by multiple team members through independent manual coding of selected interviews. To optimize intercoder reliability, the team worked together to develop a shared understanding of broad terms and their meanings. Transcripts were then coded using NVivo 8 qualitative analysis software, applying the initial broad categories followed by iterative micro-analysis of data related to strong emergent themes. Results from the data generated with the Requirements Worksheet were analysed to identify patterns and contrasts in the kinds of data the scientists were willing to share prior to publication, followed by further analysis of the interview data to draw out associated motivations and rationales. We were particularly interested in capturing how raw data are transformed throughout the research process and how these various (often refined) representations relate to 'share-ability'.

4. Findings

(*a*) *Forms of sharable data—sharing what and when*

The forms of data identified by participants as sharable were often also considered to have the most scholarly or reuse value, that is, potential for generating new results. However, there is a high level of variation in what is considered a sharable dataset. Table 3 provides excerpts from four reported cases that have been selected to illustrate the range of characteristics of datasets identified as sharable from the scientists' perspective. Across the 20 cases, image formats ($n = 4$), databases ($n = 4$) and tabular (spreadsheet) data ($n = 10$) were the most common, but many sharable datasets are constructed either from multiple files or from more than one format.

As noted earlier, composite datasets were also common; this type of dataset is derived from multiple data sources, integrated and analysed to produce a new complex dataset. For example, one earth and atmospheric science researcher spends upwards of 3 years gathering data from various sources on as many as 50 variables to build a dataset for modelling historic climate patterns. Below, we discuss this case further in terms of the impact of this practice on sharing.

Participants generally had positive views of data sharing and expressed openness to sharing their own data, particularly with people in their field to better advance their area of research. Not surprisingly, willingness to make data available increased as data were cleaned, processed, refined and analysed in the course of research. Key to understanding how this enthusiasm to share becomes action, however, requires uncovering what forms of data are intended for distribution, and the amount of work and time involved for the willing scientist to make these data available. There was, for example, only one participant willing to share raw data beyond immediate collaborators; while five participants stated that they were willing to share 'with everyone' once the data were normalized, few had ever actually done so.

In total, 60 per cent (12/20) of the participants identified a need to restrict some or all of their data from public access for some length of time. Of the 12 participants who stated that they were willing to share their data after

Table 3. Primary 'sharable' data formats for four research areas.[a]

field	specific research area	form to be shared	formats	type of dataset	size	shared when?
earth and atmospheric science	severe weather modelling	compressed output of the model	VIS5D	one file per dataset	10–100 MB	4–6 month embargo
agronomy	water quality, drainage and plant growth	cleaned and reviewed sensor and sample data	.xls	approximately 100 spreadsheets per dataset	approximately 1 MB each, up to 20 MB	after publication
geology	geobiology and microbes	averaged sensor and sample data; photographs	.xls and jpg	one file; images	< 1 MB	after publication
civil engineering	traffic movement	cleaned and normalized sensor data	MYSQL (postgresql)[b]	one database	approximately 1000 K d^{-1}	1 month to 1 year embargo

[a]For similar details on other research areas included in this study, see Data Curation Profiles available at http://datacurationprofiles.org.
[b]Two participants from different disciplines share the same database.

the findings were published, five would still require an embargo period first, ranging from 1–3 months to 2–5 years. For at least eight participants, sharing any data before publication or embargo was strictly limited to known and trusted individuals who were either immediate collaborators or known associates. In cases in which data had been shared, limitations and conditions were common, but any 'rules' for sharing were far from systematic. Actual dissemination of data to researchers beyond known colleagues or peers was limited to seemingly small numbers based on specific requests, and these were generally handled via e-mail or a posted CD-ROM.

Co-authorship was emphasized by several of the scientists (e.g. kinesiology, geology, plant science and agronomy), particularly in regard to data they shared that served as an essential component of an article. One agronomist explained that it was important for 'any article that might come out, because without…the authorship thing [it] really boils down to, can this work have, would it [have] been done without this person's contribution? And I think without the data set, they couldn't have' had a publication [U1A1J3_7_3_2008]. Interestingly, several also noted that expectations for co-authorship are generally anticipated, but arrangements often go unspecified. In fact, and in contrast to the sorts of legally binding agreements found in the governmental and industry sectors (Crompton *et al.* 2009), the coordination of authorship in these circumstances does not appear to be a regular part of the scholarly process.

Views on preservation are of particular importance for IR applications, and a need for long preservation periods was commonly expressed by participants. Two of the participants (10%) reported that their data have reuse value for 3–5 years, and four participants (20%) reported a preservation period of 5–10 years. Event-driven factors were reported that influence the shorter reuse periods, such as the introduction of new technologies (e.g. sensors) that result in 'better' datasets, or updated methods (e.g. algorithms) that alter data processing and analysis. Thirteen of the participants (65%) reported that their data had reuse value for a minimum of 10 years, and four of those (an agronomist, a soil scientist, a geologist and an anthropologist) reported their data having value for reuse for an indefinite period of time. The data with very long-term value tend to be observational data that have comparative value or potential for integration with other similar data for longitudinal analysis, data mining or modelling.

Finally, several scientists explicitly discussed the need for community-based data resources. For example, a climatologist and a geophysicist work in research areas where there are very few specialists in the world (approx. 10 and 100, respectively) and for which shared data resources could be a tremendous asset. In these research communities, individual scientists build detailed datasets to serve as either input for models or test sets against which to compare experimental results. In the first area, climatology, the input data are not shared at all, although the outputs from the model runs are shared; for geophysics research, databases are recreated for each new project and these are rarely shared, resulting in the duplication of effort and variance of error rates. In both the cases, scientists identified the need for standard, reference-type datasets or collections where datasets might be aggregated that will be accessible to everyone in their small research community.

(b) *Distinguishing private and public sharing*

When scientists talked about data sharing, they described practices that satisfied two fundamentally different needs—keeping data private or making it public. The instances considered here suggest that, beyond motivation, the effects of exposing or supplying one's data have differing implications for IR development. Supplying data involves either targeted transfer of data to and from current collaborators or close colleagues or distribution on request; these have the effect of keeping data private to some extent. Exposure of data is a more general dissemination activity that makes data accessible to the wider public. Targeted supply is a primary concern for several of the participants, who experience significant barriers to moving large sets of data from one institution to another. As one scientist reported:

> ...it gets pretty sticky when you try to move big chunks of data in and out of the university, ...university security, just getting through firewalls, you know. ...[I]f we could get people credentials, temporary credentials, it wouldn't be hard, but because we've got all this security in place, which makes sense, it makes it kind of hard to move things in and out. They're way beyond, way bigger than emailing files, ...they overwhelm our anonymous FTP servers. So as a result I'm setting up these little shadow, anonymous FTP servers. Again, you know, reinventing IT technology. [U1C1D2_5_08_08]

Another participant reported needing to provide a copy of a dataset to a colleague's student for analysis. The set of data files was too large for the file-sharing service provided to individuals by the university, and there was no departmental server for this task. Her solution was far from optimal: '[W]hat I think I may have to do is put it on the server back in (another state), so she can grab it from there and then I'll take it off' [U2A3L1_5_22_08]. Providing data to colleagues for processing or analysis is a standard part of the research process that many local technical infrastructures do not readily support.

However, data are also supplied to colleagues for evaluation purposes, rather than analysis or publication. A significant difference with respect to data sharing interactions in this collaborative context rests in the lack of formalized systems found in collaboratories (Bos *et al.* 2007). In these cases, there is a private exposure or exchange with the aim of obtaining feedback on the value, quality or potential of the data as part of the current research process. For example, one participant talked about bringing data to a conference specifically to get some advice from a close colleague, stating, 'I'm going to a conference' where there will be people from 'a lab that I used to be a member of, ...I have no qualms at all about sharing anything that's unpublished and I'll get feedback about it, which is pretty important' [U1B1B1_4_10_08]. This concern about trust is tied directly to concerns about data misuse, discussed further below; however, some scientists seem to use concern for 'misuse' of their data as a way to limit—or at least explain—their actual sharing practices.

(c) *Sharing with the wider public*

Providing data to wider public audiences, either through supply or exposure, is motivated by a range of individual concerns (e.g. data-analysis problems), scientific norms (contribution to one's field) and requirements (e.g. requirements

of funding agencies). Emerging from the scientists' discussions on sharing are several dimensions that are suggestive of the variation in both what it means 'to share' and how these processes are carried out. These variables include directionality (data flow in a single direction out to others or bi-directional exchange); the intended level of exposure (private versus increasingly public) and the level of control applied to the outflow. Sharing practices are constrained by data-management pressures and personal experience, which include the desire for personal control of one's research products and proper recognition or reward. For participants in this study, supplying data to others beyond bounded collaborations was often based on a decision process that weighs familiarity (i.e. how well known is the person requesting the data) with data-preparation complexity or cost (e.g. the time it takes to prepare data). As one agronomy/soil scientist explained, 'it just seems impossible for me to keep sending information to people once they request it, because it's just too taxing on my time as a researcher to be providing data and information to people. ...I'm starting to (get a lot of requests), and so my student is really busy with sending this out. ...[S]o just in the last two weeks we've got three requests that are probably going to take ten hours of time, and so if this keeps happening, it's really going to be burdensome on us, and we'll have to say no to a lot of these...' [U1A1P1_08_19_08].

Although many participants reported that they have or would provide data on request, several stated that, in practice, this has been limited to immediate or known colleagues and sometimes well-known people in their research area. For example, one geologist noted, 'I have done it...with people that I know through working with them in the past or through conferences or knowing their work in the area. ... or then working on pieces of work and emailing me or ringing me, and I've done the same to them, you know. I've been working on projects [and said] well, I know they're working on this area, would you be willing to let me to use your dataset for comparison, and they say 'well, yeah, I'm still working on this, but here it is'. So, yeah, I've done that, frequently' [U2G3J2_07_02_08]. For this scientist, the expectation of reciprocity influences sharing with a level of public beyond his immediate circle.

(*d*) *Avoiding misuse*

In the discourse on data sharing, risks of data misuse (and other barriers to sharing) have been prominent themes (e.g. Sieber 1989; Sterling & Weinkam 1990; Van House *et al.* 1998; Campbell *et al.* 2002; Gardner *et al.* 2003; Foster & Sharp 2007; Bertzky & Stoll-Kleemann 2009). To date, concerns have not been well supported with empirical evidence; however, our data suggest that this is an area of vital importance and in need of much deeper investigation. Misuse incidents experienced by scientists in this study influenced their views on the appeal of data sharing, decreasing their willingness to share and increasing their cynicism in data-sharing initiatives, but they also had a real impact on their behaviour. The most dramatic case unfolded during the course of staged interviews, where a geologist had experiences in which other scientists published data he had shared while he was still working on his own analysis. There had been no intention on his part that these data would be used by the other party in this manner. As a result, this previously enthusiastic advocate of early data sharing has shifted to a more conservative position. Now he will share only with close colleagues

just prior to publication and make data publicly available after publication of the related article. In another case, an agronomy/soil scientist discussed concerns about depositing data into a public repository because of the manner in which some industry groups might use the data. This scientist was aware of incidents where data had been 'cherry-picked' to make claims about the efficacy of certain products in marketing materials; this was seen as a breach of scientific protocol and also a lack of enforcement of the formal agreement between the institution and industry organizations. This practice has reduced general confidence on the part of some faculty scientists in the enforcement of these institutional agreements. A third participant recounted co-authoring and attribution problems that led to him withholding data among an active, bounded collaborative group.

A small number of participants also described incidents of, or concern about, wrong or inappropriate interpretation of data, and some have developed strategies to guard against this kind of misuse. In one case, for example, this agronomist was only willing to share data beyond immediate collaborators familiar with the data if she was allowed to approve the new application and interpretations:

> . . .my main concern is I don't want people to misuse it. . . and if I don't have some relationship of trust then I don't know whether they're going to, you know, just go off and do something and never check with me to see, well, was this a good interpretation. . . . there's all kinds of ifs, ands, and buts on a field site that it's extremely difficult to write down in a way that they're going to read all of that stuff, and I may have something that I remember, well, that plot in that year. Now, ideally all of that would be written down, but in practice, it's not. . . .so, I need to know that they would at least run something by me, so that I can say well this doesn't make any sense because . . .they don't know all of the things that happened. [U1A1E1_8_15_08]

Based on these reported concerns and incidents, different categories of misuse are emerging: misappropriation, misinterpretation and disregard of good faith practices. The geologist's latest experience is an example of misappropriation; the agronomist's concerns with understanding and application are misinterpretation problems that she would manage with a strategy of strict control of any data provided to anyone beyond her immediate collaborators. The institution-industry situation is an example of disregard for normative scientific practices. Improper citation is another kind of misuse that falls in this category. An anthropologist documented two such experiences. The first was particularly unsettling for researchers, and the second demonstrates problems with proper citations for promoting data access:

> A couple months ago I was at a national meeting and looking at a poster, which I didn't understand at all, I was trying to understand it, and some strange stuff about age estimation, and it came out that it was quote, 'my data'. It just didn't, in that case the attribution was so slim that I couldn't even tell, you know, what the data were. [U2A1S1_5_22_2008]

In the second incident,

> [S]omebody used it in a journal article, used the [part of the] Korean War data. Now those are paper records that I only found out about because a former colleague of mine stumbled upon them, basically. So I got those all into a database so [that] people could have them, and I told

> somebody about it, they then downloaded it, used it, and, cited the source of the original [location] of the data, which is fine, but you can't get it there. I mean, there's a summary publication which everybody cites, but it doesn't give you the data. [U2A1S1_5_22_2008]

The experiences of the scientists studied here substantiate some of the reported concerns about misuse of data (Edwards *et al.* 2007; RIN 2008); more importantly, they illustrate that the concept is complex and requires further articulation of levels of risk and how they can be managed. For example, as seen above in the misinterpretation case, the scientist suggested that her reputation was at risk 'because the data still kind of reflects back on who collected it' [U1A1E1_8_15_08]. These are valid concerns that have been previously recognized in other fields such as cognitive neuroscience, in which it has been suggested that such problems could be resolved at the stage of publication through the peer-review system (Postle *et al.* 2003). There is still much to learn about the nuances of conditions for sharing and the effects on practice of actual incidents of 'misuse'; additional research to produce a taxonomy of misuse would facilitate decision making for institutional data-policy development and scholarly communication services.

5. Implications for institutional repositories and university data services

Data curation services will need to accommodate a wide range of subdisciplinary data characteristics and sharing practices. Although it is possible that a given research area might require specialized service strategies, our research indicates that there will be particular types of data that can be managed well in an IR context when characteristics and practices are well understood. For example, disciplines for which IP might be complicated by cross-disciplinary data 'ownership' will benefit from planning and negotiation services for data acquisition and deposition. This was exemplified by one anthropologist whose research was funded by a National Science Foundation programme requiring public access, and the data were generated from a bone collection owned by a museum that wanted control over those images.

Scientists rarely have the skills or resources needed to prepare all their data for public sharing, as noted in the RIN (2008) report, and it is clear that leaving the effort to the end of a project is inefficient and much more costly. Sharable forms ought to be acquired at the research stage when it is produced, and this would facilitate other data curation processes by providing a temporal guide for resource allocation. For example, there was an obvious disconnect between participants' perceptions of the value of metadata and knowledge of metadata: while they see the application of metadata standards as highly important, many are unaware of existing standards. Working with scientists early in the research cycle to identify appropriate metadata standards and to support application facilitates deposition before the end of a project and potentially reduces cost at the point of ingest.

Several baseline IR services are suggested by the prominent data-management needs or system requirements seen in our results. First, it was evident that embargo services are an essential component of a repository system and that embargo periods need to be flexible and controlled by the scientists. Secondly, support for data exchange among members of small project teams or close

colleagues will be of particular value for many scientists who are not involved with large-scale projects or collaboratories that have information and communication systems at hand. DataStar is an example of a 'staging repository' that may prove optimal for this kind of function, as it supports deposit, use, exchange and storage of data until it is ready to be moved to a long-term repository or discarded (Steinhart 2007). Thirdly, explicit data-citation information must be included in IR records; these references may be different from those listed to citing the related publication, and users will benefit from actual references posted with unambiguous directions on when to use them.

IRs are part of a larger set of strategies emerging across academic institutions, and nations, to provide for the stewardship of scientific research data and mobilize data for 'e-Research' and 'e-Learning'. Within universities, IRs are also part of a larger technology and service structure that can contribute to data curation service development. Various offices, including the Offices of Research, University Archives and Campus IT organizations, are potentially strong partners for building on existing infrastructure and services. At the institutional level, for example, services to support data-management planning will be critical, as it is anticipated that federal funding agencies will soon require these as part of that grant proposals. At the individual level, many scientists in this study reported that they would welcome data-management consultation during the research process. The implication for information professionals across the academic institution is for increased engagement with scientists during research-production cycles and interacting with them in new and systematic ways about the data they produce (M. H. Cragin 2009, unpublished data).

6. Conclusions

In the small-science research areas represented by our participants, there were no field-wide norms for sharing, and none of the scientists routinely deposited data into any shared repositories, except for five scientists who could contribute genomics data to Genbank. Although sharing with close, trusted collaborators happened regularly, sharing with anyone outside this inner circle, sometimes including other members of a project team, took place through 'just in time' negotiations. Views on public sharing of data in repositories were primarily speculative, as most respondents had only shared data within collaborations or by request. Scientists could readily identify the data they considered the most sharable, and this tended to be data that were the most 'presentable' or easy to share, although not necessarily the most valuable for preservation over the long term, especially for reuse by researchers in other disciplines.

The high level of variation and complexity in data forms and sharing practices, even among the selected research areas represented in this study, indicate that resource demands for curation services for small science will be high, particularly at acquisition and ingest stages (Beagrie *et al.* 2008). Moreover, IRs will be underutilized as a resource for data if they do not support existing data practices and protect against misuse.

Although this study was intended to inform institutional repository development, the findings highlight the need for additional investigation of use and non-use of other types of repositories. In addition, patterns and outcomes

of use, and misuse, of shared data need ongoing study and evaluation; more research is needed on reuse risks (such as with dataset selection) with various data forms. Collection development decisions could be guided by a fuller understanding of reuse value, and local studies on expectations for durations of data use are needed to support budget projections for preservation services. There is also a need to untangle the concept of data 'sharing', recently labelled a 'misnomer' in Nature Genetics (Integrating with integrity 2010). This work will become easier as repository activity increases, which is more likely to happen when curation service policies are in place, which mitigate risks of sharing data (RIN 2008; Pryor 2009).

This research is supported by the Institute of Museum and Library Services grant no. LG-06-070032-07, D. Scott Brandt, PI. We thank our Research Assistants, Deborah Leiter and Marina Kogan for their assistance in data collection and processing. We are also grateful to the anonymous reviewers, whose comments served to make this a stronger contribution.

References

Beagrie, N., Chruszcz, J. & Lavoie, B. 2008 Keeping research data safe: a cost model and guidance for UK universities. Final Report to JISC. See http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf.

Bertzky, B. & Stoll-Kleemann, S. 2009 Multi-level discrepancies with sharing data on protected areas: what we have and what we need for the global village. *J. Environ. Manage.* **90**, 8–24. (doi:10.1016/j.jenvman.2007.11.001)

Borgman, C. L., Wallis, J. C. & Enyedy, N. 2007 Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *Int. J. Dig. Libr.* **7**, 17–30. (doi:10.1007/s00799-007-0022-9)

Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E. & Olson, G. 2007 From shared databases to communities of practice: a taxonomy of collaboratories. *J. Comp.-Mediat. Commun.* **12**, article 16. See http://jcmc.indiana.edu/vol12/issue2/bos.html.

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A. & Blumenthal, D. 2002 Data withholding in academic genetics: evidence from a national survey. *J. Am. Med. Assoc.* **287**, 473–480. (doi:10.1001/jama.287.4.473)

Carlson, S. 2006 Lost in a sea of science data. *The Chronicle of Higher Education*, 23 June 2006.

Chompalov, I., Genuth, J. & Shrum, W. 2002 The organisation of scientific collaborations. *Res. Policy* **31**, 749–767. (doi:10.1016/S0048-7333(01)00145-7)

Choudhury, G. S. 2008 Case study in data curation at Johns Hopkins University. *Libr. Trends* **57**, 211–220. (doi:10.1353/lib.0.0028)

Crompton, S., Aziz, B. & Wilson, M. 2009 Sharing scientific data: scenarios and challenges. In *W3C Workshop on Access Control Application Scenarios, Luxembourg, 17–18 November 2009.* See http://epubs.stfc.ac.uk/bitstream/4463/w3c1_syc.pdf.

De Solla Price, D. J. 1963 *Little science, big science.* New York, NY: Columbia University Press.

Edwards, P. N., Jackson, S. J., Bowker, G. C. & Knobel, C. P. 2007 Understanding infrastructure: dynamics, tensions, and design. Final Report of the Workshop on History and theory of infrastructure: lessons for new scientific cyberinfrastructures. See http://hdl.handle.net/2027.42/49353.

Foster, M. W. & Sharp, R. R. 2007 Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nat. Rev. Genet.* **8**, 633–638. (doi:10.1038/nrg2124)

Gardner, D. *et al.* 2003 Towards effective and rewarding data sharing. *Neuroinformatics* **1**, 289–295. (doi:10.1385/NI:1:3:289)

Hall, S. R., Allen, F. H. & Brown, I. D. 1991 The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Cryst.* **A47**, 655–685. (doi:10.1107/S010876739101067X)

Heidorn, P. B. 2008 Shedding light on the dark data in the long tail of science. *Libr. Trends* **57**, 280–299. (doi:10.1353/lib.0.0036)

Integrating with integrity. 2010 Editorial. *Nat. Genet.* **42**, 1. (doi:10.1038/ng0110-1)

Karasti, H., Baker, K. S. & Halkola, E. 2006 Enriching the notion of data curation in e-Science: data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. *Comp. Support. Cooperative Work* **15**, 321–358. (doi:10.1007/s10606-006-9023-2)

Kling, R. & McKim, G. 2000 Not just a matter of time: field differences and the shaping of electronic media in supporting scholarly communication. *J. Am. Soc. Inf. Sci.* **51**, 1306–1320. (doi:10.1002/1097-4571(2000)9999:9999%3C::AID-ASI1047%3E3.0.CO;2-T)

Murray-Rust, P. 2007 Data-driven science—a scientist's view. In *NSF/JISC Repositories Workshop, Phoenix, AZ, 17–19 April 2007*. See http://www.sis.pitt.edu/~repwkshop/papers/murray.html.

National Science Board (NSB). 2005 NSB-05-40, long-lived digital data collections: enabling research and education in the 21st century. See http://www.nsf.gov/pubs/2005/nsb0540/.

Pachura, C. M. & Martin, J. B. (eds) 1991 *Mapping the brain and its functions: integrating enabling technologies into neuroscience research*. Washington, DC: National Academy Press.

Palmer, C. L. & Cragin, M. H. 2008 Scholarly and disciplinary practices. *Annu. Rev. Inf. Sci.* **42**, 165–212.

Peterson, B. J. 1993 The costs and benefits of collaborative research. *Estuaries* **16**, 913–918. (doi:10.2307/1352449)

Postle, B. R., Shapiro, L. A. & Biesanz, J. C. 2003 On having one's data shared. *J. Cogn. Neurosci.* **14**, 838–840. (doi:10.1162/089892902760191063)

Pritchard, S. M., Anand, S. & Carver, L. 2005 Informatics and knowledge management for faculty research data. *EDUCAUSE Res. Bull.* **2005**. See http://net.educause.edu/ir/library/pdf/ERB0502.pdf.

Pryor, G. 2009 Multi-scale data sharing in the life sciences: some lessons for policy makers. *Int. J. Dig. Curat.* **4**, 71–82.

Research Information Network (RIN). 2008 To share or not to share: publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. See http://www.rin.ac.uk/data-publication.

Rice, R. 2009 DISC-UK DataShare Project: Final Report. See http://ie-repository.jisc.ac.uk/336/1/DataSharefinalreport.pdf.

Sieber, J. E. 1989 Sharing scientific data I: new problems for IRBs. *IRB: Ethics Hum. Res.* **11**, 4–7. (doi:10.2307/3564184) See http://www.jstor.org/stable/3564184.

Steinhart, G. 2007 DataStaR: an institutional approach to research data curation. *IASSIST Quart.* **31**, 34–39.

Sterling, T. D. & Weinkam, J. J. 1990 Sharing scientific data. *Commun. ACM* **33**, 112–119. (doi:10.1145/79173.79182)

Van House, N. A., Butler, M. & Schiff, L. 1998 Cooperative knowledge work and practices of trust: sharing environmental planning data sets. In *CSCW '98: Proc. ACM Conf. on Computer Supported Cooperative Work, Seattle, WA, 14–18 November 1998*, pp. 335–343. New York, NY: ACM.

Witt, M. 2008 Institutional repositories and research data curation in a distributed environment. *Libr. Trends* **57**, 191–201. (doi:10.1353/lib.0.0029)

Witt, M., Carlson, J., Brandt, D. S. & Cragin, M. H. 2009 Constructing data curation profiles. *Int. J. Dig. Curat.* **4**, 93–103.

Wong, G. K. W. 2009 Exploring research data hosting at the HKUST institutional repository. *Ser. Rev.* **35**, 125–133. (doi:10.1016/j.serrev.2009.04.003)