CHAPTER 8
MOVING FROM SMALL SCIENCE TO BIG SCIENCE: SOCIAL AND ORGANIZATIONAL IMPEDIMENTS TO LARGE SCALE DATA SHARING

Eric T. Meyer

University of Oxford
Oxford Internet Institute
1 St Giles
Oxford
OX1 3JS
United Kingdom
tel. +44 (0) 1865 287218
email: eric.meyer@oii.ox.ac.uk
Word count: 5994

Cite As:
Meyer, E.T. (2009). Moving from small science to big science: Social and organizational impediments to large scale data sharing. In Jankowski, N. (Ed.), *e-Research: Transformation in Scholarly Practice* (Routledge Advances in Research Methods series). New York: Routledge, pp. 147-159.

INTRODUCTION

One of the challenges of building collaborative information systems for scientific and social scientific data is that many new projects are actually extensions of existing projects, often going back decades, which have embedded logic and work practices that are highly resistant to change. This resistance to change cannot, however, simply be attributed to conservatism on the part of individual scientists. On the contrary, many of the scientists that are discussed in this chapter are enthusiastic about the idea of contributing data to larger collaborations in exchange for the additional data that they will, in turn, have available to them. In practice, however, protocols that are the result of years of cumulative decisions at the local level have resulted in information storage systems that are highly idiosyncratic and often resistant to federation. To demonstrate this point, I report on two projects in very different domains which nevertheless share similar barriers to building a collaborative infrastructure.

The issues surrounding moving from small science to big science are not new. Derek de Solla Price (1963) identified some of these issues four decades ago in his work that helped to develop the field of scientometrics. More recently, scholars in computer science have addressed issues of scalability (Simmhan, Plale, & Gannon, 2005; Zheng, Venters, & Cornford, 2007), and any number of papers discussing the implementation of grid-enabled projects have identified scalability as one of the key issues developers have had to deal with (Pakhira, Fowler, Sastry, & Perring, 2005; Shimojo, Kalia, Nakano, & Vashishta, 2001). Many of these discussions of scalability, however, are focused on large projects such as physics and astronomy Grid-based projects. Smaller e-Science and e-Social Science projects, however, also face issues of scale as they attempt to share data more widely and contribute to larger datasets. One issue to be raised in this chapter is how legacy data can cause significant problems during efforts to standardize and federate datasets. This is not a new issue, as countless scientists are dealing with these issues.[1] Only recently, however, have researchers begun to pay attention to how small scientific projects negotiate the changes required as they move towards becoming large, collaborative scientific projects (Carlson & Anderson, 2006; Walsh & Maloney, 2007) and attempt to sustain these collaborations over time (Bos et al., 2007).

In this chapter, I discuss two case studies that serve to illustrate some of the issues faced when small scientific projects move to large-scale data sharing and collaboration. This text is not intended to be an exhaustive treatment of this topic since it relies on two possibly idiosyncratic cases, but is meant to stimulate discussion on these issues among researchers studying e-Science, e-Social Science and, more generally formulated, e-Research projects.

This chapter discusses some of the issues that arise when small scientific projects make the transition to becoming part of larger scientific collaborations, as seen from a social informatics perspective. The data for the paper is drawn from two cases: a systematic study of a humpback whale research project involving federating data about the population and movements of humpbacks in the Pacific Ocean, and observations based on the author's personal experiences as part of a psychiatric genetics collaboration that has recently become involved in contributing data to a large, shared data repository. While these two projects are in very different scientific domains, they share a number of characteristics including decentralized decision-making, limited data management expertise, and long-term collections of legacy data that have contributed to the difficulties the projects have faced in moving from small science to big science. One of the important issues raised by this paper is the tension between the desire for flexibility and innovation in scientific practice as weighed against the

[1] See for instance the ESML Earth Science Markup Language (Ramachandran, Graves, Conover, & Moe, 2004) and the Kepler system for dealing with legacy data in scientific workflows (Altintas et al., 2004).

need for compatible data standards in large-scale scientific data infrastructures. This tension must be resolved if e-Science and e-Social Science projects are to succeed in the long term.

MARINE MAMMAL SCIENCE

Scientists who study whales, dolphins and other marine mammals use a variety of scientific techniques, including acoustics, genetics, and photo-identification to gather data pertaining to marine mammal population characteristics and behavior. In 2006-2007, I studied marine mammal scientists who use photo-identification as a main data collection tool (Meyer, 2007a, 2007b). The study was designed to understand the ways in which the scientists' work had changed when they switched from film-based to digital photography. The research involved 41 interviews with principal investigators, junior researchers, and technicians working at 13 different laboratories in the United States and Europe. While the marine mammal scientists are primarily engaged in e-Science as opposed to e-Social Science, their experience trying to build collaborative scientific infrastructures for studying the social behavior of whales and dolphins shares much in common with scientists who study the social behavior of humans. The characteristics of the animal populations under study influences any given marine mammal scientist's desire to share data collaboratively and to spend time, resources and effort building infrastructure for the ongoing sharing of data. For instance, some of the dolphin projects in this research focused on relatively small populations of animals (200-500 dolphins) that did not travel widely. Since the animals were located in a small geographic area and were often only studied by a single group of scientists, there was little incentive to share the data. Quite the reverse was actually true, as some of the scientists studying these small populations had concerns about others using their data without having gone through the trouble of collecting it, as reflected in the following quote:

> Leah Tull:[2] Well, honestly, I'm very protective about it… I guess it rather bugs me that I have to do the work and everyone always asks me for a CD…it's our scientific study.

Compare this to a large group of scientists studying humpback whales:

> Jacob Tipton: We knew the success of our project we had done in [location], but also its limitations because it largely funded the contribution and analysis of photographs, but not the dedicated gathering of data. So it very much relied on who was already doing work in certain places. So there were these huge gaps and we knew to really to answer the questions about population size, trends, human impacts, stock structure we had to cover some of these new areas.

Humpbacks and other whales can travel thousands of miles during annual migrations. The total population size of humpback whales in the Pacific Ocean is in the range of 15,000-20,000 animals. Scientists hoping to learn more about humpbacks, then, have an incentive to collaborate with other scientists studying humpback whales throughout the Pacific Ocean. By sharing data, they open up the possibility of being able to track individual animals' movements from place to place, rather than just recording their repeated visits to a single location over time. Until quite recently, however, relatively little formal collaboration occurred and most collaborations were formed based on informal relationships. These informal relationships were often based on common attendance at a university or through shared contacts built during professional conferences and meetings. Recently, however, there have been efforts to build much larger databases, including a project called SPLASH (Structure of Populations, Levels of Abundance, and Status of Humpbacks). SPLASH involves over 300 scientists working in 50 research groups working in various areas in the

[2] All names are pseudonyms.

Pacific Ocean (Calambokidis et al., 2007). Several contributing groups of scientists were included in the research study reported here.

In SPLASH, as in other marine mammal photo-identification projects, researchers use photographs taken in the field to identify individual animals and to track the sightings of the animals geographically and over time. The initial efforts to develop the technique of photo-identification go back nearly 30 years. Prior to the early 1970s, much of the dolphin and whale research involved techniques that either disrupted the animals' behavior (such as freeze-branding) or used dead animals (including necropsies on carcasses of animals that had either died naturally or were killed for research purposes). Increasing public interest influenced by the nascent environmental movement in the 1960s and the "Save the Whales" campaign of the 1970s helped draw attention to the need to develop less invasive techniques. More importantly, the passage of the U.S. Marine Mammal Protection Act of 1974 banned most harassment of marine mammals; current research requires special federal permits to be allowed even to approach the animals closely with research vessels to take photographs. Even though there was some initial skepticism about the ability to unambiguously identify individual animals using photographs, the technique is now widely accepted and is practiced by many of the scientists who study these species:

> Dr. Gerald Lemoine: The original seed of the idea…came from talking around the campfire… It was one of these fun things where ideas come to fruition independently due to synergism and the overall status of the sciences. In the '50s, I don't think anyone would have really come up with that idea... I remember telling [a prominent scientist in 1971] about this idea of photo-identifying, and he said, 'Don't do it. It is not worthwhile. You're barking up the wrong tree. You can't do it, you'll be disappointed. The only way to do it is to catch them and brand them.' But, of course, they use photo-identification now very successfully.

Using photo-identification techniques, scientists have amassed large amounts of data on thousands of whales collected in individual catalogs of animal images and databases of related information. The problem as relates to e-Science and e-Social Science, however, is that the data have been collected by dozens of individual scientists who maintain catalogs of humpback whale images, each with their own cataloging schemes, numbering protocols, and databases of associated information. Most catalogs prior to 2003 consist of film photographs in the form of slides, black and white negatives, or black and white prints. Since 2003, many scientists in the field have switched to digital photography, and have designed additional idiosyncratic systems to deal with the digital catalogs which may or may not be consistent even with their own prior catalogs.

> Dr. Marcia Parrett: It's just too complicated – so, right now I have two data bases; one on my older data from 2003 back, which was all of the data collected on film, and now I have a new…database that's all the data collected on digital... So, this spring, I'm actually going to [location]…and we have a collaborative agreement where we share data back and forth and we'd kept it pretty much in the same format except we need to get more on the same page and we're going to work with their computer guy up there at the end of May and really get our databases uniform. Maybe then, it won't be all…the data won't be the same but they'll be the same format.

While there are many in the field who believe strongly in the desirability of sharing these data, considerably fewer currently see a need to make local data collection and storage procedures more consistent within the field. One respondent in this study who had spent time considering these issues was also one of the scientists who, for the time being at least, continued to use film-based photography rather than switch to digital photography:

> Robert Newton: And if you don't have a really good filing system standardized, that doesn't change every time someone thinks it might be better done a different way. So

> I'm kind of waiting, I guess, to see if it really stabilizes with a naming protocol and a filing protocol that is not going to wander every time someone comes up with a new software for digital pictures. That happens frequently and you'll get, people send us pictures off a camera and they'll be in files maybe a Canon software, or a Nikon one. And you can convert them all to jpegs and fart around with them but, basically, I don't want to be a film processor.

Even among the SPLASH collaborators scientists often continued to work primarily using their established practices, leaving it up to the five SPLASH area coordinators to reformat and rename their contributions to conform to the project's standards.

> Jacob Tipton: …We were not as dictatorial [as we might have been]. Because we were working with established researchers in the area and kind of seeking broad collaboration, many of the researchers maybe started incorrectly with the assumption that people had their own ways to do things that worked and weren't necessarily trying to force them to do it one way. But partly because of the rapid start of SPLASH, we weren't fully thought out ourselves…. I haven't fully thought out why some of it was as screwy as it was in terms of experienced researchers and the one thing I do think about is that they were dealing with the transition to digital as well and so they had their own system that worked.

This illustrates a key point that is discussed again below: when small scientific projects are faced with sudden and rapid growth by adding numerous non-collocated collaborators, issues of data management and organization often fall to the wayside until problems later surface. This will appear again in the discussion of the GAIN psychiatric genetics project below. For the SPLASH collaboration, the scientists thought first and foremost about getting out into the field, finding humpback whales, recording their identifying features with a digital camera, and recording data such as GPS information and environmental data. The timing of the beginning of the SPLASH collaboration also contributed to the confusion. The first year of the SPLASH collection was 2004, and many of the contributors had either switched from film-based to digital photography either in 2003 or 2004 and were still working through how to adapt their methods and organizational practices to the new technology.

The SPLASH collaboration is just one example of scientists who have struggled as their small disconnected projects are faced with relatively sudden increase in collaboration and in scale. In the case of SPLASH, two forces pushed this change. The first is scientific: the desire of the scientists to better understand the population structure and long-range behaviors of humpback whales. The second force, however, is economic: SPLASH was a funded project, and thus offered scientists the first real chance to begin to respond to the first scientific force. Science costs money, and new funds attract new scientific projects (Carlson & Anderson, 2007). This monetary inducement can make the scientific desire to collaborate come into sharper focus for busy scientists with many demands on their time and attention.

This brief overview of some of the issues facing the SPLASH collaborative should serve to give some impression of how a small scientific project can struggle with information issues as it tries to contribute to larger scientific data infrastructures. Next, I will turn to a scientific project in a completely different domain to understand how some of the specifics of SPLASH are not unique.

PSYCHIATRIC GENETICS

The second project illustrating the ways in which small scientific projects can struggle when faced with contributing to larger scientific infrastructures involves genomic research into the basis of certain psychiatric disorders, specifically bipolar disorder. Like the humpback research, this is also a long-standing project; over a period of 20 years, researchers have collected blood, genetic data, and phenotypic data on thousands of subjects. Again, while this study at first blush may appear to be primarily an e-Science project, the portion which I discuss here is primarily a social science project: collecting interview data on the behaviors and social interactions of individuals with certain mental health disorders, and interviewing their family members. My data for this section of the paper was not collected systematically, but is the result of my personal involvement of this as a central player in the collection and management of phenotypic data for the project over a period of ten years from 1997-2007.[3] During this time, the bipolar (BP) collaboration grew somewhat (expanding from four collaborating institutions to 11), but still was primarily an example of small science. Each contributing university had a small number of staff working on the project, usually from one to five staff members, and the entire collaboration involved fewer than 50 people.

In 2006, the BP project was one of six long-term studies in the U.S. selected to be a contributor to GAIN, the Genetic Association Identification Network. GAIN is a public-private partnership project between the U.S. National Institutes of Health and a number of private sector firms, including Pfizer, Affymetrix, Perlegen, and Broad. There was no funding offered to studies selected to participate in the GAIN project. Instead, the approach of GAIN was to use a carrot to attract scientists to contribute their data in exchange for getting access to extensive genotyping information on their research subjects in the form of genotyping using one million SNP (single nucleotide polymorphism) microarrays. These 1M SNP chips are an order of magnitude larger than many of the previous genotyping projects available to the scientists.

The contribution of the data, however, also has a price: both the phenotypic information contributed by the scientists and the genotypic information generated as part of GAIN were to be made immediately available to researchers worldwide, including to pharmaceutical companies hoping to use the information to help develop new (and presumably potentially profitable) drugs. One major change for the scientists contributing to GAIN has to do with the embargo period. In the past, data collected by the scientists was generally released to other researchers one year after the final collection of data ended and the data had been cleaned for use. This meant that the scientists had exclusive use of the data in its raw format throughout the data collection period, and for at least a year in analyzable, final format. In the case of GAIN, however, the genotypic information is being released to all parties at exactly the same time. The contributing scientists have a 9-month period during which they have exclusive publication rights, but after the 9-month period is up, anyone may publish findings from the data. While the difference between 9 months and a year may seem minor, recall that the previous embargo of a year was for access to the data, which would then require time for analysis. In the case of GAIN, however, there is no embargo at all for access to the data, only for the ability to publish. As a result, the scientists are faced with working on a much tighter schedule and, at the same time, are analyzing datasets that are an order of magnitude larger than those to which they are accustomed. This project is still ongoing at the time of writing and plans are in place to continue to monitor how the scientists deal with this increased pressure to quickly analyze and publish their data.

One hurdle that the BP project had to overcome after being selected as one of the initial six GAIN studies was that the subjects being included in the genotyping had been collected over a period of 20 years using three different versions of the interview instrument, which in turn were encoded into three different phenotypic databases with incompatible variable names and formats. The largest set of items in these phenotypic databases are the

[3] No data was collected as part of an IRB-approved study, but the author has written permission from the collaboration's lead investigator to discuss the workings of the projects and his own role therein.

answers to over 100 pages of questions, administered as semi-structured interviews performed by trained clinical researchers. Interviews take from 4-6 hours on average to administer, and result in recorded values for approximately 2600 variables. In addition to these data, there are also tables that record each research participant's "final best estimate" which is the clinical diagnosis assigned to that person based on a trained clinician's analysis of his or her interview, family history, medical records, and other information. Each subject has multiple best estimates in the database because at least two clinicians plus the interviewer and an editor each assign a diagnosis, but each has only one final best estimate. These final best estimates are in the form of a hierarchical diagnosis using diagnostic systems that have changed over the years of the study. The earliest diagnoses use a combined DSM-IIIR/RDC system, while the latest subjects are diagnosed with DSM-IV.[4]

Because this interview schedule has gone through several iterations over the 20 years of the project, there are three main versions of the phenotypic database. The first includes data that were collected via paper interviews and entered into an Oracle database designed and maintained by a federal contractor. The second set of interviews were also completed on paper, and then entered into a Paradox database designed by a database designer located at one of the project sites. The third set of interviews were initially done on paper but were then transitioned to direct entry interviewing using laptops and tablet PCs using a proprietary database designed for the study by an external company. All three versions are converted from their native storage formats into SAS files for use and analysis, but their variable names are not consistent. For instance, a similar variable in the first set variables might be "I1120", in the second set "Number_of_manic_episodes", and in the third set "V756". While this seems confusing to outsiders, those familiar with the data have found that having very different names serves as a quick shorthand for being able to see at a glance the source of a variable or set of variables.

The differences in variable names illustrate the difficulty in combining data from several iterations of the same project, let alone trying to combine that data with data from other projects. Because the decisions regarding things like variable naming conventions was left to database designers rather than done in a systematic fashion, trying to later combine these data requires a fairly high degree of understanding of the research project. One of the contributing sites has had several staff members working on a combined dataset that converts variables from all three versions to a standard naming system; this project has taken over two and a half years. Also, because the interview schedule changed between iterations, questions have been rewritten, added and deleted from version to version, so there is no clear mapping from one to another in the majority of cases. When the data were primarily used internally by people very familiar with the research, the analysts were able to informally share knowledge about how best to use the data. When such data needs to be shared more widely, however, these idiosyncrasies can be very confusing. In addition, the group in charge of GAIN data distribution also required well documented data dictionaries for the databases, which had not been kept in a format compatible with the GAIN requirements.

Although considerably more detail could be shared about this study, this short description should illustrate how decisions made by a wide variety of people over a period of

[4] DSM is the Diagnostic and Statistical Manual of Mental Disorders that lists different categories of mental disorders and gives specific criteria required for a set of symptoms to "meet criteria". These criteria generally include a list of potential symptoms and number of symptoms required, plus the number of days the episode must have lasted to meet criteria. The DSM-IIIR was published in 1987 as a revised version of the 1980 DSM-III, and DSM-IV was published in 1994. The RDC (Research Diagnostic Criteria) is a similar, older system developed in the 1970s.

many years can have major implications should the scientific data later be re-used in ways that the original designers were unable to foresee. This accumulation of many small decisions, most of which were perfectly sensible decisions at the time, can subsequently result in considerable work trying to reconcile the many differences that are the result of those decisions.

DISCUSSION

As we saw in both the SPLASH and BP collaborations described here, the shift from small to big science can often be fairly rapid and tumultuous. Most of the personnel working on both of these projects are trained in scientific methods and theory; there are few participants with any systematic background in data management and organization. In the case of SPLASH, all of the personnel responsible for designing the database systems and methods of information organization were trained in biology, and none had any formal training in database design or information management. The decision regarding which personnel to assign to these duties relied primarily on identifying staff members with an affinity for and skill with computers. While the databases that were designed as a result were perfectly useable, they did not incorporate fully normalized designs or other features that more trained designers may have included. More importantly, because they were designed by a single user or small group of users, if SPLASH wishes in the future to federate its data even further (possibly by expanding to other regions, or incorporating additional species of whales) it will be faced with trying to integrate incompatible designs.

In the case of the BP project, a small number of people with considerable expertise in data management were part of the decision-making process, but even in this case many of the decisions were not made systematically. For instance, the format for variable names for the third iteration of the phenotypic interview was decided by the company that programmed the database, and had much more to do with the structure of their particular implementation of an EAV (entity-attribute-value) database design than with the needs of the scientific analysts. Even the second iteration of the interview, which was stored in a database designed by a skilled analyst, experienced unexpected variable naming confusion when a number of variable names including the ampersand (&) sign were altered when the data were imported into SAS; SAS does not support the ampersand, so converted both those characters and any spaces to an underscore. Thus, "Total Manic & Depressive Episodes" in Paradox became "total_manic___depressive episodes" in SAS, with three consecutive underscores in the center.

Since much of the analysis in the past had relied on *ad hoc* requests made by investigators to the small number of data experts working in the collaboration for subsets of the data, much of the knowledge about the idiosyncrasies of the datasets was never written down in a systematic fashion. When GAIN required that this data be shared and documented, considerable effort had to be made to translate this knowledge into a written format. Also, the data sent to GAIN was cleaned, but the three versions of the dataset were still separate. The internal effort by one research group to construct a unified dataset was not yet finished at the time the data needed to be provided to GAIN, nor was the group who had spent such time and effort on combining the data willing to release the combined dataset publicly until they had gotten use out of it. The plan was to release the combined dataset to the scientific community after a year of internal use.

Among the striking similarities between these very different scientific domains is the extent to which existing practices are the result of many small decisions made in many small research projects by many individual researchers and technicians over a long period of time. These decisions were often made with little or no discussion of the impact beyond the particulars of the specific local study, often because scientists at the time did not anticipate the future need to share the data with other scientists. As a result, both projects had, over a period of years, adopted highly idiosyncratic methods that hindered an easy transition to sharing data. Because of this, scientists find they have had to spend a great deal of time, effort and money to transform their data into forms that can be used in a larger data-sharing project. In some cases, these barriers may be high enough to dissuade scientists from contributing at all. In other cases, unless the idiosyncratic nature of the data is diminished, users of the collaborative data may find it confusing or misleading. These social realities of scientific practice must be addressed if e-Science and s-Social Science projects are to be successful, particularly when applied to existing scientific protocols.

One of the reasons both the cases described here developed their levels of idiosyncrasy was that both had historically engaged in smaller scale collaboration, but had done so in a very non-centralized fashion. The quote from Jacob Tipton cited earlier in the chapter mentioned that SPLASH didn't "try to force them to do it one way". Likewise, the BP project was always very decentralized, to the extent that individual contributing sites were able to choose to skip portions of the interview schedule and choose to use alternative systems of organization and management locally as long as their final data was provided in the agreed upon formats. Both the SPLASH and BP projects have non-dogmatic leaders who have been flexible in their approach to managing the collaborations. They have, by and large, not attempted to impose decisions, but instead have sought consensus and have allowed considerable individual latitude to their colleagues and contributors. This flexible and decentralized form of leadership is common among scientific and creative teams (Mumford, Scott, Gaddis, & Strange, 2002), and is not inherently problematic. Science relies on the freedom of scientists to innovate (Bush, 1945; Gordon, Marquis, & Anderson, 1962), although some recent work suggests that these patterns are changing in the face of calls for measures of increased accountability and relevance for scientific work (Demeritt, 2000; Harman, 2003). Having collaborative science designed with wide latitude for individual contributors to pursue unique contributions can arguably lead to better, more innovative science. From a data management perspective, however, lack of reliance on standard data structures, naming styles and metadata makes federating data either difficult or impossible.

In addition, data management often is not considered a top priority during the startup phase of scientific research projects. In the cases described here, few of the scientific decision makers had detailed knowledge of the demands of data management, and as a result treated data management as ancillary to the main scientific research design. Also, in both cases the studies were rather hastily implemented, and saw a number of operational changes during the early phases. Encoding these changes into data systems became a case of trying to hit a moving target until the scientific protocol had stabilized. Based on personal observation of these and a number of other scientific projects, however, this initial uncertainty is not uncommon with grant funded research. In the United States, grants are written, submitted, and revised over a period of years in many cases, and by the time funds are secured there have often been local changes in personnel and wider changes in the state of current scientific knowledge. The grants are also written with some flexibility in terms of specific activities, and the decisions about how to actually concretize the research are often left until the funds have been secured.

The question, of course, is to what extent data management demands should dictate scientific decisions, and conversely to what extent should individual scientists be allowed to ignore issues of compatibility and data availability. This is an important and enormous ongoing issue for scholars interested in working with large, federated datasets. This tension between conformity to standards and freedom to innovate is not resolved, and will arise time and again as e-Science and e-Social Science projects continue to develop and to attract new contributors. It is unlikely that there can be a single answer that will lessen this tension.

Collaborative scientific projects will continue to balance the needs of individual scientists for flexibility in their data collection protocols with the demands of federated databases for data to be organized in a consistent and structured manner. However, if e-Science, e-Social Science and, more generally, e-Research projects are to succeed in the long term, this tension must be successfully resolved.

ACKNOWLEDGMENTS

An earlier version of this chapter was presented at the 2007 NCeSS conference in Ann Arbor, Michigan, United States. Portions of this work have been supported by the Rob Kling Center for Social Informatics, Indiana University, Bloomington, Indiana, U.S.; and by the School of Library and Information Science, Indiana University, Bloomington, Indiana, U.S.; and by the U.K. National Centre for e-Social Science through a grant from the U.K. Economic and Social Research Council (RES-149-25-1022).

REFERENCES

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., & Mock, S. (2004). *Kepler: An extensible system for design and execution of scientific workflows.* Paper presented at the 16th International Conference on Scientific and Statistical Database Management, Santorini Island, Greece, June 21-23.

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. *Journal of Computer-Mediated Communication, 12*(2), article 16. Retrieved 8 August 2008 from: http://jcmc.indiana.edu/vol12/issue2/bos.html

Bush, V. (1945). Science: The endless frontier. *Transactions of the Kansas Academy of Science, 48*(3): 231-264.

Calambokidis, J., Barlow, J., Burdin, A. M., Clapham, P., Ford, J. K. B., Gabriele, C. M., et al. (2007). *New insights on migrations and movements of North Pacific humpback whales from the SPLASH project.* Paper presented at the Biennial Conference on the Biology of Marine Mammals, Cape Town, South Africa.

Carlson, S., & Anderson, B. (2006). *e-Nabling Data: Potential impacts on data, methods and expertise.* Paper presented at the Second International Conference on e-Social Science, Manchester, UK, June 28-30.

Carlson, S., & Anderson, B. (2007). What a*re* data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication, 12*(2), Article 15. Retrieved 8 August 2008 from: http://jcmc.indiana.edu/vol12/issue2/carlson.html

de Solla Price, D. J. (1963). *Little science, big science*. New York: Columbia University Press.

Demeritt, D. (2000). The new social contract for science: Accountability, relevance, and value in US and UK science and research policy. *Antipode, 32*(3): 308-329.

Gordon, G., Marquis, S., & Anderson, O. W. (1962). Freedom and Control in Four Types of Scientific Settings. *American Behavioral Scientist, 6*(4): 39-43.

Harman, J. R. (2003). Whither Geography? *The Professional Geographer, 55*(4): 415-421.

Meyer, E. T. (2007a). *Socio-technical perspectives on digital photography: Scientific digital photography use by marine mammal researchers.* Ph.D. dissertation, Indiana University, Bloomington, IN. ProQuest Digital Dissertations database Publication No. AAT 3278467.

Meyer, E. T. (2007b). Technological change and the form of science research teams: Dealing with the digitals. *Prometheus, 25*(4): 345 - 361.

Mumford, M. D., Scott, G. M., Gaddis, B., & Strange, J. M. (2002). Leading creative people: Orchestrating expertise and relationships. *The Leadership Quarterly, 13*(6): 705-750.

Pakhira, A., Fowler, R., Sastry, L., & Perring, T. (2005). *Grid enabling legacy applications for scalability – Experiences of a production application on the UK NGS.* Paper presented at the UK e-science All Hands Meeting (AHM'05), Nottingham, UK.

Ramachandran, R., Graves, S., Conover, H., & Moe, K. (2004). Earth Science Markup Language (ESML): A solution for scientific data-application interoperability problem. *Computers & Geosciences, 30*(1): 117-124.

Shimojo, F., Kalia, R. K., Nakano, A., & Vashishta, P. (2001). Linear-scaling density-functional-theory calculations of electronic structure based on real-space grids: Design, analysis, and scalability test of parallel algorithms. *Computer Physics Communications, 140*(3): 303-314.

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Record, 34*(5): 31-36.

Walsh, J. P., & Maloney, N. G. (2007). Collaboration structure, communication media, and problems in scientific work teams. *Journal of Computer-Mediated Communication, 12*(2), Article 19. Retrieved 8 August 2008 from: http://jcmc.indiana.edu/vol12/issue2/walsh.html

Zheng, Y., Venters, W., & Cornford, T. (2007, June). Distributed development and scaled agility: Improvising a grid for particle physics. *Working Paper Series #163.* Retrieved August 30, 2007, from: http://is2.lse.ac.uk/wp/pdf/wp163.pdf

AUTHOR BIOGRAPHICAL SKETCH

Eric T. Meyer is a Research Fellow at the Oxford Internet Institute (OII), a department of the University of Oxford. He is one of the researchers on the Oxford e-Social Science (OeSS) project, which is a node of the U.K. National Centre for e-Social Science (NCeSS). OeSS studies the social, legal, institutional, and ethical issues related to e-Research. He has written on a variety of topics related to science and technology from a social informatics perspective; his PhD in social informatics at Indiana University was one of the first doctoral degrees awarded in this field. He also has extensive experience working with large data sets and scientific collaborations from his 10-year stint as a national data manager in the U.S. Web site: http://people.oii.ox.ac.uk/meyer; email: eric.meyer@oii.ox.ac.uk