Institutional and Individual Influences on Scientists' Data Sharing Practices

Youngseek Kim
Syracuse University
221 Hinds Hall
Syracuse, NY 13244
+1-315-443-4508
ykim58@syr.edu

Jeffrey M. Stanton
Syracuse University
206 Hinds Hall
Syracuse, NY 13244
+1-315-443-2879
jmstanto@syr.edu

ABSTRACT
Many contemporary scientific endeavors now rely on the collaborative efforts of researchers across multiple institutions. As a result of this increase in the scale of scientific collaboration, sharing and reuse of data using private and public repositories has increased. At the same time, data sharing practices and capabilities appear to vary widely across disciplines and even within some disciplines. This research sought to develop an understanding of this variation through the lens of theories that account for individual choices within institutional contexts. We conducted a total of 25 individual semi-structured interviews to understand researchers' current data sharing practices. The main focus of our interviews was: (1) to explore domain specific data sharing practices in diverse disciplines, and (2) to investigate the factors motivating and preventing the researchers' current data sharing practices. Results showed support for an institutional perspective on data sharing as well as a need for better understanding of scientists' altruistic motives for participating in data sharing and reuse.

Keywords
Data Sharing, Data Reuse, Data Repository, Institutional Theory, Theory of Planned Behavior, IT Capability, Altruism

1. INTRODUCTION
As the scope and scale of science has increased, sharing and reuse of data have become essential to many scientific and engineering activities. In the 2003 report entitled, "Revolutionizing Science and Engineering through Cyberinfrastructure," members of a blue ribbon National Science Foundation (NSF) panel wrote, "We envision an environment in which raw data and recent results are easily shared, not just within a research group of institution but also between scientific disciplines and locations" [2]. Years later researchers are realizing this vision in some disciplines and sub-disciplinary areas such as high energy physics [6], climate change [19], and proteomics [22]. In other fields, however, and particularly the social sciences [34], progress has been very slow. Although scientists in these areas generate considerable amounts of valuable data every year, disciplinary traditions, institutional barriers, intellectual property concerns, and other factors appear to impede the sharing and reuse of data.

Send correspondence concerning this article to: Youngseek Kim (ykim58@syr.edu) at Syracuse University's School of Information Studies.

For example, even though the American Psychological Association (APA) mandates data sharing for researchers who publish articles in their flagship journals, Wicherts et al. [35] found it difficult to convince 103 out of 141 research teams who had published with APA to fulfill this responsibility, despite repeated attempts and extensive assurances that the requested data would not be publicly released or reused. While it is tempting to attribute this failure to particular characteristics or situations in that discipline (e.g., long publication lags), Savage and Vickers [25] experienced an even worse failure rate when requesting data from researchers who had published in two PLoS (Public Library of Science) journals – PLoS Medicine and PLoS Clinical Trials. Note that the PLoS journals reflect the new trend of "open access" in journal publishing and have explicit requirements in their editorial policies that require researchers who publish there to share their data freely with the research community. It seems evident from these examples that the idea of data sharing and reuse as a strategy to accelerate scientific discovery is appealing, but the impediments to doing so across a range of disciplines are still substantial.

Several prior studies, such as Wicherts et al. [35] and Savage and Vickers [25] have sought to document the extent of the problem in the context of different disciplines. For this paper, we take as a given that data sharing and reuse is highly variable across disciplines, and we sought to explore why this was the case. We also began with the assumption that data sharing and reuse practices were not a matter of whimsy for individual researchers, but rather that the decisions whether or not to share data for reuse (outside of one's own research group) reflected choices among communities of colleagues embedded within their universities and disciplines. More explicitly, we asked what combination of individual and contextual factors influenced scientists' decisions to share data for reuse. Because relatively little is known about this question, we elected to use the rich, qualitative data collection method of one-on-one semi-structured interviews to explore the landscape. We were guided in this exploration by a few promising theoretical perspectives that consider individual decision makers in their institutional contexts in order to understand their decisions. In the next section, we provide a brief overview of these perspectives prior to a presentation of our interview data.

2. BACKGROUND
Contemporary collaboration in nearly all of the Science, Technology, Engineering, and Mathematics (STEM) fields requires a three way combination of technological infrastructure,

institutional support, and interpersonal interactions. John Taylor, the former Director General at the Office of Science and Technology in Great Britain, focused the attention of that office on the development of sensors, networks, and computing infrastructure: "e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it" [16]. Using examples from the World Health Organization and the Large Hadron Collider, Sonnenwald [29] highlighted the powerful social and interpersonal aspects of scientific collaboration. Avery [3] reported the history and challenges of creating the multi-institutional Open Science Grid and drew particular attention to the institutional context that allowed this large scale cyberinfrastructure collaboration to emerge.

Although data sharing and reuse is only one facet of collaboration in the STEM fields, it represents a microcosm of these same three areas: institutions, infrastructure, and people. For example, to have a well functioning data repository with lots of raw data going in and lots of other users tapping into that data, one or more institutions must have the financial wherewithal to establish the infrastructure, publicize its existence within the community, work with the community to enhance and support the systems, and maintain the infrastructure over time. Meanwhile, individual contributors to that data repository must see personal and/or professional advantage – again in part set by the context of their home institutions, disciplinary training, and professional organizations – to contributing data into that repository. Individual contributors must also have a certain degree of mastery of the tools involved in preparing and submitting the data. Training and personnel support provided by their host organizations can lower the barriers to using these tools and preparing the data for reuse. As this scenario suggests, however, the institutions, infrastructure, and people are intimately connected in ways that are not easy to subdivide.

One perspective from sociology and organizational studies that may help to weave together the intertwined forces of institutions, infrastructure, and people arises in an area called institutional theory. While the traditional center of attention in institutional theory has been on the organizational level of analysis, neo-institutional theories add the proviso that macro-level influences affect micro-level behaviors [14]. Contemporary perspectives on institutional theory consider individual beliefs concerning proper social behavior, specifically when those beliefs arise from organizational rules, structures, and practices [5, 7, 10]. This idea meshes nicely with individual-level motivational theories (e.g., the Theory of Reasoned Action) that describe behavior as jointly influenced by attitudes, norms, and intentions.

In fact, institutional theory posits three kinds of institutional influences on behavior: coercive, normative, and mimetic pressures [8, 9, 27]. Coercive influence arises from the rules that the organization and its leaders set for desirable behavior of organizational members. Normative pressures refer to those typical behavioral patterns that are established historically either by organizations or by members of relevant professions. Newcomers to an organization or the profession must follow these patterns to succeed within the organization (or more broadly, within the industry, sector, or profession). Finally, mimetic pressures result from the observation of how other comparable organizations accomplish key tasks. Generally speaking, the leaders of one organization will observe the activities of another organization that is performing well and will seek to adopt those activities or methods for use within their own organization. Such imitation is often cast as a form of risk reduction: by following the lead of an apparently successful peer, one avoids the risks involved in alternative, novel activities that may be untested or may have unforeseen consequences.

These three forces map plausibly onto the data sharing role of the individual STEM researcher in the context of his or her organization and profession. First, institutions may have coercive pressures that they apply to foster the desired behavior by the individual. For example, funding organizations that support a researcher's work may stipulate that funding is conditional on the researcher's agreement to participate in data sharing. Savage and Vickers [25] documented such a condition in the editorial policies of the PLoS journals. Second, research disciplines (professions) may have historically-rooted practices that encourage or discourage data sharing. Particle physicists, for example, with their expensive, large scale experiments, pioneered the practice of large-scale scientific collaboration in the 1950s and 1960s [18], and thus were arguably the first discipline to actively exploit the Internet for collaborative scholarship (e.g., with arXiv, which was deployed prior to the availability of the World Wide Web). Finally, with respect to mimetic pressures, the National Science Board documented an explosion of genomic repositories – each focused on a different organism – following the success in the late 1980s of the European consortium that sequenced and published the genome of *Saccharomyces cerevisiae* (budding yeast [21]).

In classic institutional theory, coercive, normative, and mimetic pressures legitimize certain organizational structures and practices in a given sector. In turn, this legitimacy tends to foster isomorphism across many organizations within the sector. In other words, laws and regulations, acceptable practices, and copying of methods diffuse through the community of organizations, causing them to become more alike. In this paper, we are less concerned with the effects of coercive, normative, and mimetic pressures on *organizational* isomorphism and more concerned with how these pressures trickle down to influence the behavior of *individual* STEM researchers. Legitimacy and isomorphism arguably map on to the individual decision making level through their influence on individual researchers' motivations to share data. Prior research has frequently made such a linkage: For example, Shi, Shambare, and Wang [28] connected institutional theory and the Theory of Reasoned Action [1, 12, 13] to examine the adoption of Internet banking. Teo, Wei, and Benbasat [30] took a similar strategy in predicting the development of inter-organizational electronic data interchanges.

The Theory of Reasoned Action and its successor, the Theory of Planned Behavior (TPB), are well-established perspectives from social psychology that describe how salient beliefs influence behavioral intentions and subsequent behavior [1, 13]. TPB explains an individual's behavior based on his or her behavioral intention, which is influenced by his/her attitude toward the behavior, perception of the subjective norms regarding the behavior, and perceived behavioral control to conduct the behavior. Behavioral intention refers to a person's aim to perform a particular behavior. An attitude is a cognitive and emotional evaluation of an object. A subjective norm is a person's belief that people who are important to her expect that she should or should not perform a particular behavior. Perceived behavioral control is an individual's perceptions of his/her ability to perform a given behavior easily [1]. Each of the determinants of behavioral intention is in turn influenced by underlying belief structures (including behavioral, normative, and control beliefs [1, 13]).

Using a perspective such as the Theory of Planned Behavior, we suggest that individual researchers' data sharing intentions and behaviors emerge from their formation of attitudes about: 1) their beliefs about the "outcomes" of data sharing, and 2) their understanding of the normative behavior of other scientists in their field [20, 27]. In effect, as institutional and disciplinary pressures on data sharing increase due to increased data sharing among colleagues within a scientific community, individual researchers will respond to these pressures with some consideration of the merits of participating in the trend [27, 31].

Data sharing behavior may also be influenced by the controllability of the behavior. Perceived controllability is similar to a construct proposed by Bandura [4] – self-efficacy – that reflects judgments of one's own capabilities to enact a behavior successfully. With respect to data sharing behavior, a sense of perceived behavioral control may arise from their expertise (or lack thereof) in using the tools and technologies that facilitate data sharing. Likewise, a researcher's judgments about the availability of IT support within a team or organization, and the existence of data sharing standards, procedures, and data repositories may influence how likely they are to engage in data sharing [17].

In summary, this study adopts a neo-institutional theory perspective that incorporates the influence of individual motivations (e.g., as considered in the Theory of Planned Behavior) to examine how institutional factors impact individual researchers' attitudes and decisions about data sharing and reuse. Schein [26] argued that the institutional environment shapes participants' shared beliefs and, eventually, their attitudes towards certain behaviors in the environment. By focusing on STEM researchers' perceptions of the benefits and costs of data sharing, this study seeks to expose what combination of individual and contextual factors influences scientists' decisions to share data for reuse.

3. METHOD

3.1 Overview

We conducted a total of 25 individual interviews to understand STEM researchers' current data sharing practices. The main focus of our interviews was two-fold: (1) to explore domain specific data sharing practices in diverse disciplines, and (2) to investigate the factors motivating and impeding STEM researchers' current data sharing. Our Institutional Review Board (IRB) provided approval of a plan to conduct the individual interviews within three research universities in the eastern U.S. We sent a recruiting email message directly to the STEM researchers, and we also contacted department chairs to distribute the recruiting email message to their STEM researchers. We received 28 responses in total from STEM researchers in three research universities, and we ultimately interviewed 25 interviewees. The remaining three respondents could not be scheduled in time to complete data collection. In order to understand the domain specific data sharing practices in diverse disciplines, we tried to include at least one or two researchers in each research discipline (see Table 1).

All the interview sessions were audio-recorded and subsequently transcribed. All the interviews were conducted in English except one interview, which was conducted in Korean for the convenience of the interviewee. The first author transcribed the interview in Korean and then translated into English for the data analysis. Each interview took 25-35 minutes. We used an open-ended semi-structured interview method by asking similar structured interview questions to all the interviewees including STEM researchers' current data sharing methods, types of data generated and shared, their motivations and barriers of data sharing, and lastly interviewees' demographic information and work environments. An example of our interview questions was: "What motivates researchers (including you) in your field to share their data?" During the interviews, the participants were asked to answer the questions based on not only their own experience but also their observations in their research disciplines in general.

The 25 participants for the interviews include 11 tenured (full and associate) professors, eight assistant professors, one emeritus professor, one professor of practice, two post-doctoral research associates, and two doctoral candidates from three major research universities in the eastern U.S. (17 men and eight women). Given the goals of this research, we mainly interviewed professors rather than graduate students, but the two post-docs and two senior doctoral students provided perspectives that seemed complementary to the other data, so we retained them in the corpus. Table 1 shows the research disciplines of the 25 interview participants. There were a few minor differences between the names of the departments the interviewees belonged to versus their disciplinary affiliations.

Table 1. Research Disciplines of Interviewees

Discipline	Number of Interviewees
Biology	2
Chemistry	3
Computer Science	2
Ecology	5
Electrical Engineering	1
Environmental Engineering	4
Mathematics	1
Mechanical Engineering	2
Physics	3
Radiation Oncology	1
Science Education	1
Total	*25*

3.2 Data Analysis

The transcribed interviews were imported into "QDA Miner," a qualitative data analysis tool optimized for coding, annotating, and analyzing textual information. QDA Miner is designed to analyze interview or focus-group transcripts, documents (e.g. journal articles), and even images, and it also provide statistical data analysis along with content analysis and text-mining. The coding scheme was developed by using both deductive and inductive approaches. We started with ideas arising from neo-institutional theory and individual motivation perspectives to create our coding scheme. As we processed the data, we also used an inductive approach to create more specific codes. The basic coding scheme included institutional theory based constructs (coercive, normative, mimetic pressures), individual motivation based constructs (benefits and costs), and perceived controllability constructs (internal and external capabilities). The interview corpus contained 837 utterances overall; we applied codes to 276 of these utterances regarding the factors both motivating and

preventing researchers' data sharing (Table 2 reports the number of respondents out of 25 interviewees whose interview contained one or more instance of each code.)

4. RESULTS

The codes and verbatims revealed STEM researchers' work environments, the types of data they commonly generated in their work, current data sharing methods (if any), and their motivations for and barriers to data sharing. In the following sections, we report on each of these topics by providing a holistic overview of what the codes and their underlying utterances revealed. Table 2 shows the coding scheme we used for the motivating and impeding factors of data sharing, provides a brief explanation of each code, and the numbers of respondents out of 25 interview participants in each code.

Table 2: Content Code Explanations and Counts

Cate-gory	Code Name	Brief Explanation	# of Responses
Coercive Pressures	Funding agency push	Funding agencies (e.g. NSF and NIH) require researchers to share their data	16
	Journal's requirement	Journal publishers require researchers to publish their data before their articles are published	9
	Special funding restrictions	Sharing private companies' and military data is restricted	6
Normative Pressures	Professionalism in the fields	Data sharing is a part of their professional mission to develop science	13
	Colleagues' expectations	Feel social pressures by colleagues (being expected to share their data)	7
Mimetic Pressures	Colleagues' performance	Observed other colleagues who use shared data improve their research performance	3
Perceived Benefits	Demonstration of quality work	Shared data indicates the quality of your work; improve the overall research quality	6
	Credits and reputation	Expect credits (e.g. authorship, citations, acknowledgements), reputation, and recognition	15
	Research performance	Conduct a comparative study or large-scale study (novel scientific finding); save time and effort in replicating and collecting data	14
Perceived Costs	Data annotation	Need to annotate data with their own metadata schemes (no standardized metadata scheme)	10
	Data organization	Takes time to organize data for more understandable, compatible, interoperable formats	11
	Data set location and interpretation	Takes time to find appropriate data sets and understand the data exactly	4
	Technical problems	Being involved with compatibility and interoperability issues with data	9
Perceived Risks	Losing publication opportunities	Have less opportunities for future publications; make more exclusive publications if data is not shared	15
	Getting Scooped	Worried about data theft; cannot trust others	8
	Misinterpretation and scrutiny	Worried about having different results by not being analyzed properly or being criticized by others because data is not reliable or low quality	13
Internal IT Capability	IM/IT expertise	Have technology expertise to manage data	5
	IM/IT support	Have internal IT/IM supports from their organizations	11
External IT Capability	Data repository	Have data repositories or enough space to share data	9
	Data standard	Have data sharing standards (metadata schemes) and systematic procedures	13
Altruism	Altruistic motivation	Allow other researchers to find something interesting that the first people missed; contribute to scientific developments; help others to save time and effort	12

4.1 Research Environment and Data Generated

Most of our interview participants worked in team-based research environments or a mixture of team-based and individual work; only two scholars, a mathematician and theoretical physician, mainly worked solo. The research teams usually included a lead professor, one or two post-doctoral research associates, and a few doctoral and masters' students.

The researchers reported that they generated a large amount of domain-specific original data including experimental data (e.g. genome sequencing data, compound data), field data (e.g. soil measurement, animal behavior, tree counts), and computational data (e.g. software code, computer simulation data). Most of the interviewees felt that they have limited *individual authority* to share their data by acknowledging that sometimes they need to seek permission from others for any collaboratively collected data. Only two interviewees (one post-doc and one doctoral candidate) felt they had *no authority* over sharing the data they collected.

Researchers reported different perceptions of the *importance* of data sharing in their fields. The researchers in biology, chemistry, and ecology agreed that data sharing is critical for novel scientific findings, but the researchers in computer science, electrical engineering, mechanical engineering, mathematics, and radiation oncology disagreed with this belief. Researchers in environmental engineering and physics reported a mixture of both perspectives.

4.2 Data Sharing Methods

Researchers in different disciplines reported different data sharing methods. Most researchers reported *internal* data sharing within their research teams or among collaborators; they usually used email, FTP servers, and website as the major internal data sharing methods. We assumed from the start that this type of internal sharing was occurring, and did not further investigate beliefs or motivations in this area.

Researchers also reported diverse forms of *external* data sharing with the researchers outside their research team or collaborators. First, researchers asserted that they share their data upon request; they use email or website upload as method of fulfilling such requests. Researchers also reported contacting other researchers individually to gain access to their data sets from published articles. Across different disciplines, this data sharing method was common, and it was the only data sharing method in the disciplines that do not have any informal or formal data repositories.

Second, some researchers who do not have any formal data repositories in their disciplines used a personal website to share their data with other researchers. A group of scholars in a similar research subject develop an informal or ad hoc data repository and share data with other researchers in the research subject area.

Third, some disciplines, including biology, chemistry, and ecology, use a range of external repositories (e.g. Dryad), and domain-specific data repositories (e.g. GenBank, Protein Data Bank, Computational Chemistry Database, Crystallography Open Database, Long Term Ecological Research Data Repository). These researchers reported well-developed data sharing protocols including data repository and data standard. In these same disciplines, most of the journals require researchers to publish their data in data repositories.

Finally, researchers in certain disciplines such as chemistry – where there are small, but highly structured data sets – share their data as an electronic supplement through the journals' websites. For example, some scholars in chemistry share their compound data through their journals' online supplements.

Some researchers reported an explicit expectation of various types of professional credits for data sharing including co-authorship, citation, and acknowledgement when their data are used by other researchers. There was insufficient information to judge the differences for these expectations among different disciplines, but we noted that the researchers whose disciplines have well established data sharing practices expected less credit than the researchers who do not have any formal way of data sharing. Additionally, we noted that junior researchers had higher expectations for credit (i.e., by means of co-authorship) than senior researchers; they mentioned strengthening the tenure case as the primary motivation for this. Senior researchers seemed to have less desire for credit, as well as more altruistic motivation for other researchers.

Roughly one third of our interviewees reported that the researchers in their field generally share their data after publication. The researchers in the disciplines that do not have formal data sharing mechanisms almost always share their data only after publication. For example, researchers in the engineering fields reported sharing their data only after publication. Another third of our interviewees reported that the researchers in their disciplines shared their data right after their data collection or after a fixed embargo period, regardless of publication status. For example, the researchers in molecular biology and genetics shared their data to a data repository right after data collection. These particular researchers reported a strong sense of trust that their colleagues would not "scoop" them using the shared data.

Lastly, where data sharing was a journal requirement, researchers in chemistry and biology and some researchers in ecology shared their data along with their publications. As noted above, these were cases where journals support a simultaneous publication of relatively small, structured data sets as supplements.

In terms of types of data shared, the researchers in some disciplines (e.g., biology, ecology, environmental engineering) shared raw data, but the researchers in other disciplines (e.g., chemistry, physics) share more refined or processed data. Additionally, the researchers in computer science, computational chemistry, and physics were likely to share both software and simulation results.

4.3 Factors Influencing Data Sharing

The primary focus of this research was on the factors influencing researchers' current data sharing practice. Based on the coding we did, we confirmed specific factors both motivating and preventing researchers' data sharing. In the material below, we explain these factors in three separate groups including institutional influences, individual influences, and IT capabilities.

4.3.1 Institutional Factors

Pressures by funding agencies, journal publishers, and private funding organizations influenced researchers' data sharing practice. First, the single most significant motivation for scientists' data sharing (giving) is a push by funding agencies to make data from funded projects available. Scientific funding agencies in the U.S. including NSF and National Institutes of Health (NIH) require their awardees to share the research data from projects they fund. Second, journals' requirement of data sharing is another factor. The journals in biology, chemistry, and some in ecology require their researchers to publish their data in any types of data repositories. Third, private and certain government funding agencies restrict researchers' data sharing. For example, some pharmaceutical companies and military agencies typically do not allow their awardees to share their data.

Disciplinary influences also affected researchers' data sharing. In many disciplines, data sharing is considered part of the professional responsibility; researchers believe that data sharing is one of their missions, and that it will help the development of their research disciplines. In these same disciplines, researchers reported that they are *expected* to share their data; they feel pressure from their colleagues to do so. Researchers reported observing what other researchers do, and they indicated that they tried to follow colleagues' practices that they saw as useful. A few researchers reported a belief that the research performance of other researchers who use the shared data would improve.

4.3.2 Individual Motivation Factors

Researchers also gave evidence that they carefully examined pros and cons of data sharing before they committed to sharing data. First of all, some researchers reported a belief that data sharing could highlight the quality of their work in research. For some, data sharing provided professional "credit" including co-authorship, citation, and acknowledgement, and reputation. In terms of using the shared data, researchers also believed that data sharing would improve their research (e.g. time saving in collecting the same data, replicating data for another research, conducting diverse comparison studies and large scale research).

In contrast, researchers also believed that data sharing imposes costs for them. In some scientific disciplines (e.g. ecology and environmental engineering) researchers saw the importance of data sharing, but they saw data sharing as very costly in time and effort. Due to a lack of established metadata standards and data preparation procedures, they saw the processes of organizing and annotating their data as very expensive. These same researchers also reported technical problems in the data sharing such as data compatibility and interoperability issues. This was a similar finding across each discipline that did not have well-established data sharing standards (metadata), procedures, and repositories. Researchers in those disciplines also reported that it took substantial time to locate and understand other researchers' data since the data do not have any established data repositories and standardized metadata.

Certain perceived risks by researchers also prevented them from sharing their data with other researchers. Many researchers worried about losing publication opportunities by sharing their data. It took a lot of time and effort to collect data, and they desired having as many publications as possible from their data. These researchers also worried about getting scooped on innovative findings when they shared their data with other researchers. Two scholars in environmental engineering mentioned that "data sharing is a little bit of a threat to our science because it is less incentive (sic) to collect your own data when all data is freely shared." Additionally, several researchers considered that misinterpretation and heightened scrutiny of their data would be possible risks if they shared their data.

4.3.3 Perceived Controllability: IT Capability Factors

IT capabilities were found to be important factors influencing researchers' data sharing practice. We focused our questioning on two distinct areas: an individual's self perceived capability to work with the relevant IT tools, including local support (internal capability), and the availability of appropriate community tools and infrastructure (external capability). Internal capability included researchers' own expertise in information and technology management in sharing their data, and also included any information management and/or IT support from within their own research team or host organization. Researchers with strong expertise and internal support in these areas also reported more extensive data sharing and reuse.

External IT capability referred to supports for researchers to share their data provided by the research community at large. In this area, researchers reported data repositories, data standards (i.e., metadata standards), and established data sharing procedures as key features. Biologists and chemists reported that they could easily share their data because they have well-developed data repositories, standards, and procedures to share their data with other researchers. Researchers in engineering fields generally did not report any central or domain data repositories. These engineers also reported needing to spend a lot of time to annotate, organize, upload, and manage their data on subject-specific or ad hoc data repositories. Researchers in ecology reported that they are aware of the importance of data repositories and standards and they have developed domain specific repositories and subject specific repositories. Since their data were unstructured, however, they reported that they still needed to develop better metadata standards and data sharing procedures.

4.3.4 Altruism

Unexpectedly, altruism emerged in about half of the interviews as a factor influencing researchers' data sharing. Some researchers reported a strong desire to help their colleagues to save time in collecting data and to avoid replicating experiments unnecessarily. Additionally, these researchers believed that their colleagues could exploit the data in ways that would extend the original findings and thereby benefit the scientific area where they collectively worked. These researchers reported a sense of personal satisfaction coming from sharing their data. A couple of our interviewees mentioned the importance of data sharing across disciplines not only within a discipline. A biologist mentioned that "it is also critical to improve [data] sharing across disciplines because a lot of research nowadays is becoming more multi-disciplinary so for example you have engineers working with biologists or physicists working with engineers and especially in my field in tissue engineering its very multidisciplinary field… If scholars in different disciplines could share that information, then the field of tissue engineering would progress a lot faster."

4.4 Changes in Data Sharing

Our interviewees reported that during recent years they had observed changes in their data sharing practices. Many of our interviewees reported that researchers' awareness, funding agencies' push, journals' requirements, technological improvements, and increased availability of data repository as changes they had experienced within recent memory. Just a few mentioned the emergence of data sharing standards as another recent change.

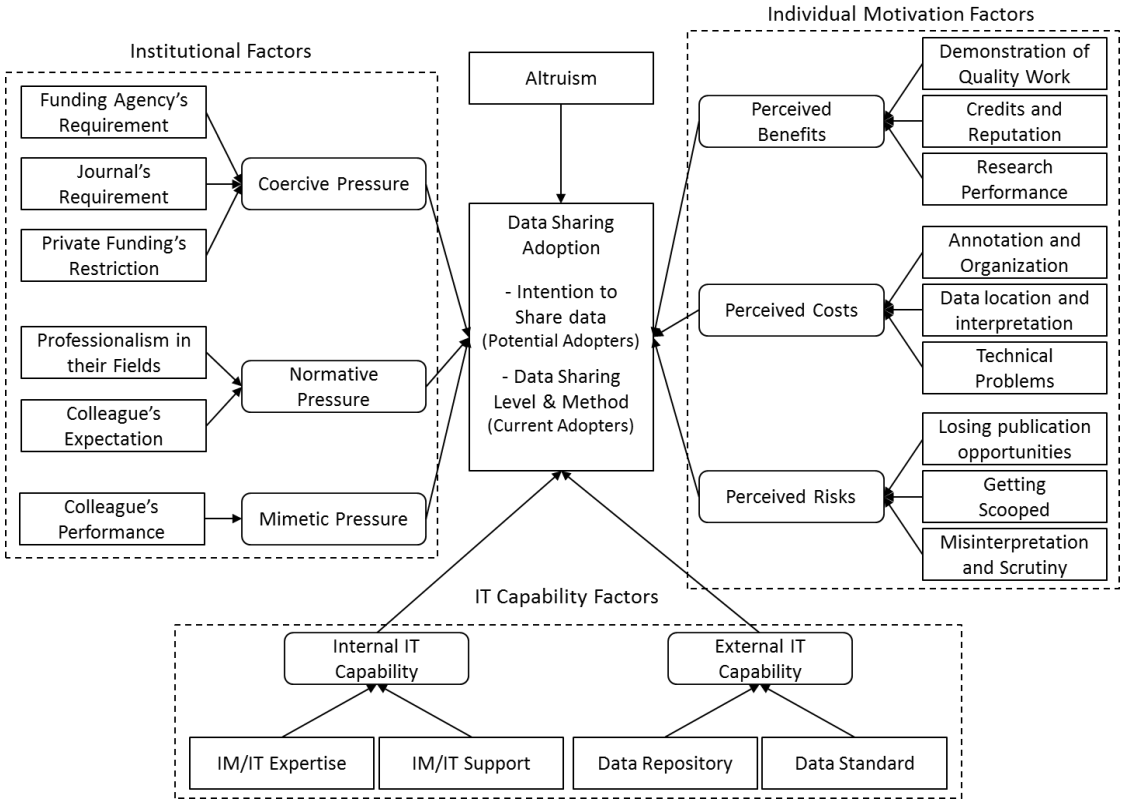
4.5 Supports Needed for Data Sharing

We asked our interviewees what kinds of additional supports they needed to facilitate data sharing. Ten of our 25 interviewees mentioned they do not need any supports since they are satisfied with their current data sharing practices. One biologist and one chemist said that they can easily share their data because they have well-established metadata standards, data sharing procedures, and data repositories. However, the remainder of our interviewees mentioned that metadata standards and data repositories are the main concerns of their current data sharing practice. Additionally, two researchers mentioned that they desired a data portal site where they could search available data sets. Several interviewees indicated that they needed better technology support. In particular, they reported that they needed professionals who could manage data sets, databases, storage, and other IT infrastructure.

5. DISCUSSION

Neo-institutional theory provided a productive lens for reviewing our interview data. Recall that some newer forms of institutional theory incorporate a cross-level perspective by linking institutional forces together with the motivations and behaviors of individual actors. We began this paper by framing the situation of the researcher as an individual actor embedded within his or her discipline as well as within the host institution and a variety of external institutions (e.g., funding agencies). Coercive, normative, and mimetic forces acting on institutions may trickle down to influence the decisions and behaviors of individuals who work within those institutions. Figure 1 provides an overview of our findings.

Figure 1. Factors Influencing STEM Researchers' Data Sharing Practices.



To have well-established data sharing practices, researchers need to have supportive institutional environments (e.g. data sharing structures, norms, policies), sufficient IT capability (e.g. data standards and repositories), and positive attitudes toward data sharing (e.g., perceived benefits, costs, risks). The combination of these can lead to more proactive data sharing practices among researchers.

One surprising finding arose from the spontaneous reports of altruistic motivations for sharing data. Typical formulations of institutional theory do not explicitly account for altruism among individual actors or groups embedded within institutions; when altruism appears, researchers use other theories to account for it [33]. Yet one essential and, arguably, widely shared value in contemporary science lies in the sharing of scientific resources for the common good [24]. Unlike commercial organizations, which generally use competition in an attempt to succeed in the

marketplace, or public agencies, that ostensibly serve the common good as their central mission, STEM researchers (at least the ones working in universities) work within a middle ground that is marked by both competition and service to the common good – what some researchers [32] have termed "coopetition" and others call "competitive altruism" [15].

While institutional theory often focuses on risk reduction (institutional isomorphism is typically a strategy for avoiding risks by conducting activities in the "generally accepted" manner), perspectives on both coopetition and competitive altruism tend to focus on performance, both at the individual and the collective levels [23]. From either an evolutionary or a game theoretic perspective, behaviors that help others and thereby increase the overall performance or fitness of a group can also have benefits to individual performance and fitness. Interestingly, in situations where an individual's reputation is important, competitive altruism appears to be a powerful strategy [11]. This idea seems to map quite neatly onto the typical contemporary situation of a STEM researcher who seeks to enhance his or her reputation through publications, presentations, and other acts of sharing with the community. Possibly, future analyses of data sharing behavior among STEM researchers should incorporate some of the theoretical elements emerging from the altruism literature.

6. LIMITATIONS

Our sample included only a subset of the range of STEM disciplines, only one or two researchers from each of these disciplines, and only researchers from eastern U.S. research universities. Each interviewee reported observations and own experiences from their own research careers, so it is likely that the results are idiosyncratic for certain disciplines – and particularly those where there is substantial variation in sub-disciplinary practices. In future research, we need to include a more representative range of scholars and a more deliberate effort to obtain participants from a representative set of sub-disciplinary areas. Although the interview method provides rich data, future research should also include mixed methods (e.g., surveys) in order to triangulate on the findings offered here. In addition, an objective snapshot of available repositories and data standards for presentation to informants could elicit more specific responses to why a researcher uses or does not use a particular data sharing resource. In addition, we focused in this study primarily on the motivations and challenges to *sharing data* rather than those associated with using deposited data. Although certain questions assessed both sides of the data sharing equation, we found that using other researchers' data is still new to many researchers.

7. CONCLUSIONS

Under the assumption that data sharing and reuse can help in the overall advancement of the scientific endeavor, we sought to understand STEM researchers' data sharing. The institutional perspective seems helpful in this regard. In the disciplines of biology and chemistry as well as within some areas of physics, researchers seem to have well-established data sharing methods covering the data lifecycle. These methods are supported by many of the institutions in which they are embedded, mainly through the availability of data sharing standards and repositories.

Contrasting biology or chemistry with the discipline of ecology, many ecologists realize that data sharing is critical for their research, but they have difficulties in data sharing because they have few well-established metadata standards and domain-specific data repositories. For those who do share data, this means spending more time and effort to annotate and organize their data with their own metadata and format. Relatedly, because they do not have well-established central or domain specific data repositories, they share their data through ad hoc mechanisms such as Web servers and email exchanges among their collaborative group members. One ecologist mentioned that "[they] should have the official protocol for [data they collected] … those should be peer reviewed and approved and archived just like our data documentation … [they need to] share the procedures, not the data only." Researchers also mentioned the importance of having access to information professionals who can support their data sharing in terms of information and technology management. The information professional can help not only share their data, but also use other researchers' data by locating and interpreting the data.

In addition, it seems important to have a central data search mechanism so that researchers can find appropriate data sets for their research. Some researchers mentioned that they have difficulties in locating and interpreting other researchers' data, and they mentioned the necessity of a central data search mechanism. Even in areas where researchers are very good at sharing their data with other researchers, many researchers still do not actively seek other researchers' data sets. Data sharing is a two-way process of providing their own data and using other researchers' data. In order to achieve the promise of data sharing, researchers need to not only provide their data, but also use other researchers' data more actively.

Finally, and perhaps most importantly, our data indicated the importance of aligning institutional pressures with individual motivations for professional achievement. The most frequently mentioned driver of data sharing behavior was the "push" by the funding agencies that support research to ensure that data from the projects they support are made available to other researchers. This force, together with pressure exerted from scholarly journals, can have a strong influence over time on the choices and activities of individual researchers. Ultimately, the advocacy of funders and journals will also need to reflect on universities' policies and mechanisms for promotion and tenure in order to have a more direct influence on the data sharing activities of researchers. When sharing (and reuse) of data leads directly to an improvement of professional reputation and resulting career rewards, researchers will have strong individual motivations to participate in data sharing and reuse.

Taken together, our results support the idea that when institutional forces, infrastructure, and individual motives converge, the behavior of individual researchers will change in response. Many of the researchers we interviewed reported having seen this convergence and these changes during the course of their own careers. Further research efforts are needed to examine the role that altruistic motivations may play in establishing a virtuous cycle of data sharing and reuse that can increase the collective benefits obtained from societal investment in science and engineering.

For future research, it would be valuable to conduct a study to understand how the factors depicted in Figure 1 may influence scientists' data sharing and reuse in different science communities. A multi-level model including individual and institutional variables may serve as a useful research strategy to understand the dynamics of different factors and their cross-

level relationships. A broad-based survey study that incorporates a representative sample of scientists from several different disciplines may help us to compare STEM researchers' data sharing and reuse by validating and confirming the multi-level research model.

8. ACKNOWLEDGMENTS

This research was supported in part by an award from the National Science Foundation (OCI-0753372). The National Science Foundation does not necessarily endorse any of the findings or conclusions of this report.

9. REFERENCES

[1] Ajzen, I. 1991. The Theory of Planned Behavior. *Organizational Behavior and Human Decision Process.* 52, 2, 179-211.

[2] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., et al. 2003. Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure.

[3] Avery, P. 2007. Open science grid: Building and sustaining general cyberinfrastructure using a collaborative approach. *First Monday.* 12, 6.

[4] Bandura, A. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory.* Englewood Cliffs, NJ: Prentice-Hall.

[5] Barley, S. R., & Tolbert, P. S. 1997. Institutionalization and Structuration: Studying the Links between Action and Institution. *Organization Studies.* 18, 1, 93-117.

[6] Becla, J., & Lim, K. T. 2008. Report from the first workshop on extremely large databases. *Data Science Journal.* 7, 1-13.

[7] Daniels, K., Johnson, G., & de Chernatony, L. 2002. Task and Institutional Influences on Managers' Mental Models of Competition. *Organization Studies.* 23, 1, 31-62.

[8] DiMaggio, P. J., & Powell, W. W. 1983. The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review.* 48, 2, 147-160.

[9] DiMaggio, P. J., & Powell, W. W. 1991. Introduction. In W. W. Powell & P. J. DiMaggio (Eds.), *The New Institutionalism in Organizational Analysis (*pp. 1-38). Chicago: The University of Chicago Press.

[10] Duxbury, L., & Haines, G. 1991. Predicting alternative work arrangements from salient attributes: A study of decision makers in the public sector. *Journal of Business Research.* 23, 1, 83-97.

[11] Fehr, E., & Fischbacher, U. 2003. The nature of human altruism. *Nature.* 425, 6960, 785-791.

[12] Fishbein, M. 1979. A theory of reasoned action: Some applications and implications. Nebraska Symposium on Motivation. 27, 65-116.

[13] Fishbein, M., & Ajzen, I. 1975. *Belief, Attitude, Intention, and Behavior.* Reading, MA: Addison-Wesley.

[14] George, E., Chattopadhyay, P., Sitkin, S. B., & Barden, J. 2006. Cognitive understandings of institutional persistence and change: A framing perspective. *Academy of Management Review.* 31, 2, 347-365.

[15] Hardy, C. L., & Van Vugt, M. 2006. Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin.* 32, 10, 1402.

[16] Hey, T., & Trefethen, A. 2003. e-Science and its implications. *Philosophical Transactions of the Royal Society A.* 361, 1809, 1809-1825.

[17] Hsu, M.-H., & Chiu, C.-M. 2004. Predicting electronic service continuance with a decomposed theory of planned behaviour. *Behaviour & Information Technology.* 23, 5, 359-373.

[18] Kevles, D. J. 1995. *The physicists: The history of a scientific community in modern America*: Harvard Univ Pr.

[19] Meehl, G., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J., et al. 2007. *The WCRP CMIP3 multimodel dataset. Bull. Am. Meteorol. Soc*, 88, 1383–1394.

[20] Meyer, J. W., & Rowan, B. 1977. Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology.* 83, 2, 340-363.

[21] National Science Board. 2005. NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. http://www.nsf.gov/pubs/2005/nsb0540/

[22] Nesvizhskii, A., & Aebersold, R. 2004. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug discovery today.* 9, 4, 173-181.

[23] Padula, G., & Dagnino, G. B. 2007. Untangling the rise of coopetition: The intrusion of competition in a cooperative game structure. *International Studies of Management and Organization.* 37, 2, 32-52.

[24] Resnik, D. B. 1998. *The ethics of science: an introduction*: Psychology Press.

[25] Savage, C. J., & Vickers, A. J. 2009. Empirical study of data sharing by authors publishing in PLoS journals. *PloS one*. 4, 9, e7078.

[26] Schein, E. H. 1996. Culture: The missing concept in organization studies. *Administrative Science Quarterly.* 41, 2, 229-240.

[27] Scott, R. W. 2001. *Institutions and Organizations*, 2nd Edition. Thousand Oaks, CA: Sage Publications.

[28] Shi, W., Shambare, N., & Wang, J. 2008. The adoption of internet banking: An institutional theory perspective. *Journal of Financial Services Marketing.* 12, 4, 272-286.

[29] Sonnenwald, D. H. 2007. Scientific collaboration: a synthesis of challenges and strategies. *Annual review of information Science and Technology.* 41.

[30] Teo, H. H., Wei, K. K., & Benbasat, I. 2003. Predicting intention to adopt interorganizational linkages: An institutional perspective. *Mis Quarterly.* 19-49.

[31] Tolbert, P. S., & Zucker, L. G. 1983. Institutional Sources of Change in the Formal Structure of Organizations: The Diffusion of Civil Service Reform, 1880-1935. *Administrative Science Quarterly.* 28, 1, 22-39.

[32] Tsai, W. 2002. Social structure of "coopetition" within a multiunit organization: Coordination, competition, and

intraorganizational knowledge sharing. *Organization science.* 13, 2, 179-190.

[33] Vandenabeele, W. 2007. Toward a public administration theory of public service motivation. *Public Management Review.* 9, 4, 545-556.

[34] Wicherts, J., & Bakker, M. 2009. Sharing: guidelines go one step forwards, two steps back. *Nature.* 461, 7267, 1053-1053.

[35] Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. 2006. The poor availability of psychological research data for reanalysis. *American Psychologist.* 61, 7, 726.