Government Information Quarterly 30 (2013) S19–S31

Contents lists available at SciVerse ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data

Djoko Sigit Sayogo [a,*], Theresa A. Pardo [b,1]

[a] Rockefeller College of Public Administration and Policy, University at Albany, Center for Technology in Government, University at Albany, 187 Wolf Road, Suite 301, Albany, NY12205, USA
[b] Center for Technology in Government, University at Albany, 187 Wolf Road, Suite 301, Albany, NY12205, USA

ARTICLE INFO

Available online 6 December 2012

Keywords:
Open data initiative
Data sharing
Data management
Research datasets

ABSTRACT

The research community is working to create new capabilities to share data and to deal with issues of data quality, standards, and protection, and ethical and responsible use of shared data. These issues have been found to influence the willingness of researchers to publish data created during the course of their research. We use the results of a survey conducted by the working groups of the DataONE project to present a new understanding of challenges to the development of global data collections and preservation by systematically examining the determinants of the researchers' likelihood to openly publish research data. This study found two key determinants affecting researchers' willingness to publish their data. First is data management in terms of data management skills and organization support. Second is the acknowledgement of the data set's originator in terms of appreciation and legal and policy requirements. This study also found that the impact of the significant determinants is contingent on the amount of data to be published. Finally, this study calls for further investigation to ascertain the relationship of data management and data quality, and systematic investigation on the roles and responsibility of government within these global data preservations.

1. Introduction

Advances in computing, information and communication technologies produce dramatic and significant impacts on scientific research, making them increasingly data intensive and collaborative (Hey, Tansley, & Tolle, 2009; Tenopir et al., 2011). The rapid advances in computing capabilities also provide useful tools in manipulating and exploring massive data sets (Hey et al., 2009; Savage & Vickers, 2009). Recognizing the magnitude and significance of digitalization and data-intensity in scientific research, in 2007, NSF solicited a proposal entitled *Sustainable Digital Data Preservation and Access Network Partner* (*DataNet*)[2] to trigger development of the necessary systems of data preservation by engaging the research community and other interested stakeholders at the frontiers of computer and information science. Two projects were selected to create a set of exemplar national and global data research infrastructure organizations (called the DataNet Partners) that would provide unique opportunities to communities of researchers to advance science and engineering research and learning. DataONE and The Data Conservancy were established to build a robust national and global digital data framework. DataONE, short for "DataNet Observation Network for Earth", is a virtual federated database to support universal access to earth and environmental data (www.dataone.org). The Data Conservancy (DC) is intended to collect, organize, validate, and preserve data for future reuse (www.dataconservancy.org).

Open data initiatives for preservation of research data such as DataONE and Data Conservancy could encourage a wealth of scientific opportunities with less effort and fewer resources. Having access to such data, data in some cases collected over a lifetime, researchers could creatively innovate from archival data sets, promote new discoveries from old data sets, and connecting new meaning from existing research data sets (Nelson, 2009). Researchers could efficiently create more opportunities without the burden of data collection and repetition of efforts. As such, with an increase in the importance of the open data initiative, the role of data sharing becomes more important (Tenopir et al., 2011). Historically, access to and sharing of research data sets was part of collegiate tradition (Stanley & Stanley, 1988), operationalized through one-to-one personal means. The act of sharing data sets was regarded as a privilege among trusted colleagues based on mutual interest and respect (Kaye, Heeney, Hawkins, de Vries, & Boddington, 2009). A researcher seeking access to a data set would begin by locating the data and the owner, initiate a relationship, build trust, respect and mutual interest, and create a collaborative enterprise in the form of shared data sets. On the other hand, the proliferation of efforts to create global data preservation, to enable data sharing and reuse, challenges the generally accepted data sharing practice and raises new uncertainties and

* Corresponding author. Fax: +1 518 442 3886.
E-mail addresses: dsayogo@albany.edu (D.S. Sayogo), tpardo@ctg.albany.edu (T.A. Pardo).
[1] Fax: +1 518 442 3886.
[2] NSF Cyberinfrastructure Vision for 21st Century Discovery—January 20, 2006 (http://www.nsf.gov/od/oci/ci_v5.pdf)—p.19.

0740-624X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.giq.2012.06.011

concerns for researchers regarding methods of sharing research data sets with the public.

This paper uses the survey response conducted by the Usability and Assessment and Sociocultural Working Group of DataONE project.[3] The objective of the survey is to understand and assess the current data sharing practices (Tenopir et al., 2011). The survey results provide an assessment of the perceptions of the barriers and enablers of data sharing that a federated data repository such as DataONE needs to consider in building the system. Through the understanding of the barriers and concerns inhibiting the willingness of researchers to publish their data, DataONE could design their project to provide secure but flexible infrastructure, policies and best practices that would help to build researchers' confidence in data sharing ("DataONE," n.d.; Tenopir et al., 2011).

In the natural sciences, a number of researchers have found that the existence of basic setups for scientific data sharing, such as technical, organizational and legal conditions, are necessary but do not automatically convince researchers to engage in data sharing practices. Extant literature asserts that data sharing practices at present are minimal, with researchers more likely to withhold their data than to share it publicly (Rodriguez, 2009; Tenopir et al., 2011). Building from this assertion, this research focuses its analysis on the supply part of the data sharing process, attempting to understand the determinants of individual researchers' motivation, factors that may convince individual researchers to publish their research data, particularly in earth and environmental science. In this regard, this paper does not consider the challenges facing the users in accessing, using, and extracting data from particular open data initiatives.

Existing literature has discussed at length the challenges of data publication in open data initiatives, for example, Reichman, Jones, and Schildhauer (2011), Tenopir et al. (2011), Zimmerman (2007, 2008), Nelson (2009), Piwowar and Chapman (2010), and others. Furthermore, a limited number of studies have focused on the role of the researchers' motivation and intentions for data publication using bibliometric measure (Piwowar & Chapman, 2010). On the other hand, the matter of how challenges affect the researchers' motivation to publish their data has received little systematic attention. This research was designed to contribute greater understanding of the behavior in publishing research data by correlating the challenges to the propensity of researchers to openly share their data. Using the survey response from DataONE, this paper will address two main research questions: 1) what are the critical challenges facing individual researchers in publishing their research data openly to the public and 2) to what extent do these challenges influence the propensity of researchers to openly share their data sets?

In accordance with the research objectives and questions, the rest of the paper is organized as follows. Section 2 will outline the theoretical background, focusing on the challenges for researchers to publish research data sets, and subsequently propose the model of the determinants in sharing research data. Section 3 briefly explains the research design and methodology used in this study. In accordance to the objective, we equate data owner as the researchers/initiator who initiate and conduct the research the first time. Limitation of such assumption is discussed in the implication section. Section 4 presents the findings and Section 5 provides discussion highlighting the findings' implications on policy for open data initiatives and, finally, Section 6 provides concluding remarks.

2. Theoretical background

The theoretical background consists of two parts. The first part summarizes the challenges and barriers for data sharing and the second part outlines the research model and theoretical justification. In this paper, we follow the formal definition given by the U.S. OMB (Office of Management and Budget)[4] to define research data as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues." We review literature on data sharing in ecological and biological domains by focusing on the two leading journals in natural science, namely: Nature and PLoS ONE. Efforts to create a global data repository to encourage data sharing in ecology and biology have a long history. For instance, NSF sponsored workshops in 1977 as a starting point on the establishment of the Long Term Ecological Research (LTER) network. Nevertheless, a current study found that data sharing practices in those domains are minimal (Rodriguez, 2009; Tenopir et al., 2011). Thus, focus on ecology and biology domains will not only allow for deriving insights from abundant studies addressing the researchers' motivation to share data in the global data repository, but will also allow identification of the elements that inhibit the sharing. In addition, we use the snowball approach reviewing citations of the identified articles in the first step and web searching using Google scholar to search the articles using keywords such as "data sharing", "scientific data sharing" and "data sharing by scientists." Where necessary, we supplement the review with literature from inter-agency information sharing particularly in the legal and policy discussion and, when doing so, we provide plausible justifications for the appropriateness of the literature.

2.1. Barriers to data sharing

Research by Savage and Vickers (2009), which explores the willingness of researchers to share their data sets to independent investigators following the publication of the result in the scientific journal, found that only one out of ten researchers agreed to do so. Thus, arguably, sharing research data sets is mostly driven by personal decision (Savage & Vickers, 2009; Vickers, 2006) propelled in part by social influence (Tucker, 2009). The researchers have specific reasons, ranging from technological aspect, organizational aspect including financial and budgetary elements, legal and policy aspect, and behavioral aspect (Arzberger et al., 2004). The first part of this literature review outlines these specific reasons from four perspectives, namely: technology, organizational, legal and policy, and data complexity due to local context and specificity.

2.1.1. Technological barriers

Technology infrastructure to ensure open access is reasonably established (Parr & Cummings, 2005), but technology to ensure data protection and data quality is still open for discussion. This section will discuss the technology-related issue in the perspective of data protection and data quality. Numbers of researchers in the natural sciences express their concerns over the issue of ineffective sharing infrastructures, more specifically related to the data architecture and data protection (Nelson, 2009; Schofield et al., 2009). In terms of data protection, researchers often question the existence of a mechanism that would guarantee that data will not be scooped, poached, or misused (Nelson, 2009; Van House, Butler, & Schiff, 1998), and that freely shared data will indeed be used ethically and responsibly (Schofield et al., 2009). Similarly, the PARSE Insight survey suggested that misuse of data becomes the major concern for scientists publishing their research data to the public (Kuipers & van der Hoeven, 2009). The technology to support the protection of their data

[3] For further description of the survey instruments and descriptive interpretation of the survey result, refer to Tenopir et al. (2011).

[4] The Office of Management and Budget's (OMB) Circular A-110, Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (http://www.whitehouse.gov/omb/circulars_a110).

once it is made available to the public significantly influences their willingness to share the data openly (Nelson, 2009).

The second major issue is related to data architecture, in particular the issue of data quality and standardization. Assurance of data quality and compatibility are pertinent not only to the scientific data sharing domain but also to interagency information sharing. There are two aspects which influence effective sharing, namely: sufficiency of data or information quality and compatibility/standardization. Inconsistencies of data definition and content could challenge the integration of a large volume of different forms of data from organizationally and geographically dispersed origins (Pardo, Cresswell, Dawes, & Burke, 2004). The expected result from data reuse and sharing can be distorted when the data sources have different formats, definitions and contexts resulting in disjointed information. This is especially critical considering the heterogeneity and complexity of data in earth and environmental science (Reichman et al., 2011). Thus, a degree of standardization of the procedure for data management becomes a prerequisite for data sharing (Postle, Shapiro, & Biesanz, 2002). The standardization and increased quality of data is achievable by creating more attention to the usability and user-friendliness of metadata management and increasing skills for data management (Michener, 2006).

Similarly, the heterogeneity and complexity of earth and environment data influences the concern for preserving data quality. The literature pointed at researchers' perception that no one other than themselves could understand the data because their primary data is far too complex for others to understand (Koslow, 2000, 2002). And vice versa, the researchers also expressed their distrust or limited understanding of the data produced by others (Koslow, 2002). The issue of preserving data quality to avoid misinterpretation once again points at the significance of data management skills. The concern over misinterpretation of the data can be mitigated if the description of relevant experimental conditions and variables are clearly specified and explained in the data sets (Koslow, 2002).

2.1.2. Social, organizational, and economical barriers

Four aspects of organizational and economic barriers have been found to discourage sharing: 1) cost of sharing, 2) incentives and merit system, 3) the culture of open sharing, and 4) structural conflicts and managerial practices in organization.

Researchers argue for the logistic barriers to data sharing (Parr & Cummings, 2005). Cost and incentives for sharing the research data become one of the major barriers to sharing research data. Based on a national survey of 3000 scientists, Campbell et al. (2002) found that the majority of scientists were withholding their data and refusing to share because of lack of resources (Campbell et al., 2002). Scientists tend to deny requests to share data due to the costs of complying with such requests in terms of time and effort (Postle et al., 2002; Vision, 2010). Data documentation is a labor-intensive process (Borgman, 2010) and researchers need to prepare data sets before they are readily available for sharing (Nelson, 2009), not to mention that some data and materials are costly (Blumenthal, Campbell, Anderson, Causino, & Louis, 1997), and some of these data are collected over the lifetime of the researchers.

The reluctance to share is also due to the current lack of sufficient rewards and incentives for researchers to do so (Koslow, 2000, 2002). The composition of incentives and rewards in the scientific world range greatly from financial incentives to reputation and public recognition. Birnholtz and Bietz (2003) argue that research data sets accrue "monopoly rents" for scientists. Monopoly rents refer to the creation of a monopoly situation, that is, exclusive control over resources, which typically generate higher real income (Krueger, 1974). Thus, there is a possibility that a scientist's ability to maintain exclusive control over his or her data will significantly correlate to the financial interest of the researchers or their home organization (Birnholtz & Bietz, 2003; Blumenthal et al., 1997). The researchers might be concerned about losing patent rights or securing future grants/funding (Ceci, 1988; Rodriguez, 2009).

The other types of incentives relate to reputation and merit. Competition for merit and reputation is more aggressive in the scientific field, which contributes to stronger resistance to data sharing (Birnholtz & Bietz, 2003). Researchers and scientists in the academic world rely on their reputations to signify prominence, and to act as a warranty of the quality of their work (Ding, Levin, Stephan, & Winkler, 2010). The researcher's reputation depends on his or her ability to understand data, and the consequent generation of publications based on the research data (Tucker, 2009). As such, Ardichvili, Page, and Wentling (2003) found, in the virtual community of practice, "fear of losing face" to be the most important barrier in sharing data/information (Ardichvili et al., 2003, p. 70). The fear that someone else might find different results, or something that the researcher missed which would undermine the previous publication and pre-empt subsequent publications, significantly affects their propensity to share (Koslow, 2000, 2002). The researchers also have concerns that procedural and computational errors will be discovered (Ceci & Walker, 1983).

Similarly, personal recognition and public appreciation significantly affects the desire to make the data publicly available. Promotion for academe is tied to publication and not data, which makes publishing data sets less advantageous for scientists (Borgman, 2010; Rodriguez, 2009). In addition, data is viewed as private intellectual property regardless of the funding sources (Nelkin, 1982) and is tied to personal glory and recognition. For instance, Barnes (1987) found that the motivation for personal recognition significantly affects researchers' willingness to share AIDS research data (Barnes, 1987).

Finally, data sets might be stored in the researchers' home institution. As a result, there exist administrative reasons for refusal to share data, such as security reasons for non-release (Ceci & Walker, 1983) or protection of institutional financial interests (Blumenthal et al., 1997). The individual researcher might be constrained by their institutional policies in sharing the data (Tenopir et al., 2011). On the other hand, this issue might not hold in the future. With the contention that research data funded by public funding is regarded as a public good and should be made openly available, the availability of data should only be restricted by legitimate considerations such as protection of national security (Arzberger et al., 2004). For instance, starting from January 18, 2011, all NSF funded research should include supplementary documents detailing the data management plan.

2.1.3. Legal and policy barriers

Legal and policy can both enhance and inhibit data sharing. Researchers could use the rigidity of policies and regulations as legitimate reasons for not sharing their data under the pretext of confidentiality or legal rights (Savage & Vickers, 2009). The liabilities associated with the breach of privacy deter researchers from sharing their data sets publicly (Tucker, 2009) and "de-identifying" sensitive data could take a lot of time (Vickers, 2006). Thus, there is a trade-off between the vagueness of sharing policies and the flexibility of sharing (Nelson, 2009). On the other hand, learning from the lesson from inter-agency information sharing, legal regulation and policy is necessary to enhance data sharing by ensuring proper and accountable use of data and information, considering that lack of policies does not guarantee a neutral sharing environment (Zhang & Dawes, 2006). In fact, unresolved legal issues can deter or restrain the development of collaboration, even if scientists are prepared to proceed (Bos et al., 2007). Thus, the legal and policy requirements could actually mitigate the primary concerns of data being scooped, misused or misinterpreted. Policies for data management and regulation to protect confidentiality and privacy have a crucial role in supporting data accessibility (Arzberger et al., 2004).

2.1.4. Local contexts and specificity

Local contexts and specificity are regarded as the distinctive challenges to data sharing in the natural sciences. By nature, data in earth

and environment domains are heterogeneous and complex (Reichman et al., 2011). The data collection process in earth and environment domains are extremely complex and significantly influenced by the local context where the data is collected (Zimmerman, 2007, 2008). Borgman (2010) points to four dimensions pertinent to the difficulties in sharing due to the complexity of data, namely: specificity of purpose, specificity of events, specificity of methodology, and the duration of research. Natural scientists usually pursue a specific question at a specific site about a specific phenomenon; therefore, each subject might have different characteristics and require a different methodology (Borgman, 2010).

There are primary concerns expressed by researchers who pointed out that details and local contexts underlying the data are often unknown (Smith, Seligman, & Swarup, 2008) and not captured in the secondary data (Borgman, 2010; Zimmerman, 2007, 2008). This condition results in accentuating the fear of data misinterpretation and misuse because the data will not be easily understood by others (Koslow, 2000, 2002). According to the intrinsic data quality pattern formulated by Strong, Lee, and Wang (1997) local context and judgment will eventually lead to data not being used, due to little added value or poor reputation (Strong et al., 1997). Strong et al. (1997) argued that local contexts from multiple sources of data will lead to questionable *believability* of data, resulting in poor application and little added value. Similarly, the involvement of judgment in data production will result in questionable *objectivity* and *believability* of data that lead to poor application, little added value, and eventually, data uselessness (Strong et al., 1997).

2.2. Research model

This section summarizes the above reviews of challenges to data sharing as well as providing theoretical justification for the tested model. Motivation to act is constrained by certain facilitating conditions, and their absence represents a barrier to perform particular actions (Ajzen, 1991). Building from this assertion, this paper argues that the challenges confronting the researchers in publishing their research data sets openly to the public affect their likelihood to share their data sets. Particularly with the tendency that researchers will withhold their data due to the assumption that research data is private intellectual property. Drawing from a rich, cumulative body of empirical research, this study identifies the following factors which influence the willingness of researchers in sharing research data sets openly: support from the organization to manage the data, technology to manage data, legal and policies in place to protect the data, and assurance from data misinterpretation and misuse.

Data in earth and environment science is heterogeneous and very complex (Reichman et al., 2011), thus complicating the open sharing of research data sets. It is not easy for others to understand the data owing to the complexity of research design and methodology (Koslow, 2000, 2002). This complexity and heterogeneity provokes the fear of data misuse and misinterpretation. The data need to be understood within the environment in which the data is originally collected to preserve the important details and contexts during the data collection (Smith et al., 2008). As a result, extant literature argues that the inclusion of relevant contexts and conditions underlying the data in the data sets could overcome and mitigate the fear of data misuse and misinterpretation. Hence, researchers have the responsibility to ensure that their data is clean, accurate, well-annotated, and that they protect confidentiality and privacy in the data (Vickers, 2006).

Undertaking this responsibility requires the acquisition of data management skills, and appropriate legal and policy requirements. Data management skills are needed to ensure the quality of data. In addition, legal and policies are needed to support data protection and prevention from misuse. Legal and policy will ensure proper and accountable use of data and information (Zhang & Dawes, 2006). Hence, adequate data management skills and the existence of legal and policies protecting data sharing become necessary elements to promote data sharing.

On the other hand, preparing data for publication requires a lot of effort and can be costly. Researchers need to invest adequate resources and time to prepare the data for publication and sharing (Borgman, 2010; Nelson, 2009; Postle et al., 2002; Vision, 2010). Unfortunately, the resources are usually lacking for researchers, and data management is a labor intensive process (Borgman, 2010). Consequently, promoting data sharing necessitates the availability of two elements, 1) creating more usability and user-friendliness of technology supporting metadata and data management, and 2) support from organization to manage data. Resource lacking researchers need support from their organization for technical training, managing, and storing the data during and beyond the research project. Previous studies in knowledge sharing pointed out that organizational supports are significant in influencing sharing behavior, either directly (Bock, Zmud, Kim, & Lee, 2005), or indirectly, mediated by subjective predictors such as perceptions and trust (Lin, 2006).

Another important aspect to promote data sharing is sufficient rewards and incentives for researchers sharing their data (Blumenthal et al., 1997; Koslow, 2000, 2002; Postle et al., 2002). Research data have various meanings for researchers; it entails financial, reputational, and promotional functions. Thus, adequate forms of rewards and incentives need to be in place to stimulate researchers to share. In light of the above theoretical justification, there are seven factors that the literature identified as potential predictors which could affect the motivation of researchers in sharing their data, namely: data management, organizational support, institutional barriers, legal and policy barriers, incentives (in terms of economic and acknowledgement), and misinterpretation of data reuse.

The extant literature also pointed out the influence of age, gender, and time to encourage sharing. Tenopir et al. (2011) found that younger researchers are prone to associate a barrier to access data with an impediment to scientific progress and demonstrate a different behavior and place different conditions on the fair exchange of using other people's data. Time is also regarded as an important constraint for data sharing, considering that preparing data for open publication requires a lot of effort (Postle et al., 2002; Tenopir et al., 2011; Vision, 2010). As a result, this study generates a model illustrated in Fig. 1.

3. Methods

3.1. Sample description

Data used in this paper is based on the online survey administered by two DataONE working groups—1) the Usability and Assessment and 2) Sociocultural working group.[5] The link for this baseline assessment survey was open from October 07, 2009 to July, 2010. The survey's objective is to understand the current practices in data sharing (Tenopir et al., 2011). The descriptive interpretation of researcher perceptions in regard to the barriers and enablers of data sharing, including the description of survey instruments, was published in Tenopir et al. (2011). The research sample is a random selection of individuals identified by the Usability and Assessment and Sociocultural working groups as stakeholders: scientists, librarians, computer scientists, decision makers, citizen scientists, students, and teachers. This paper analyzes the data without differentiating the stakeholders and assumes that all of the above mentioned stakeholders are researchers. After data cleaning the final observations of 555 respondents were used in this paper.

The respondents were mostly mature adults (an average of 44 years old) and well-educated (49% were employed as professors or lecturers; 19% were graduate or post-doctoral students). The

[5] The survey data was extracted from the DataONE working group platform and access to the data required authorization from DataONE.

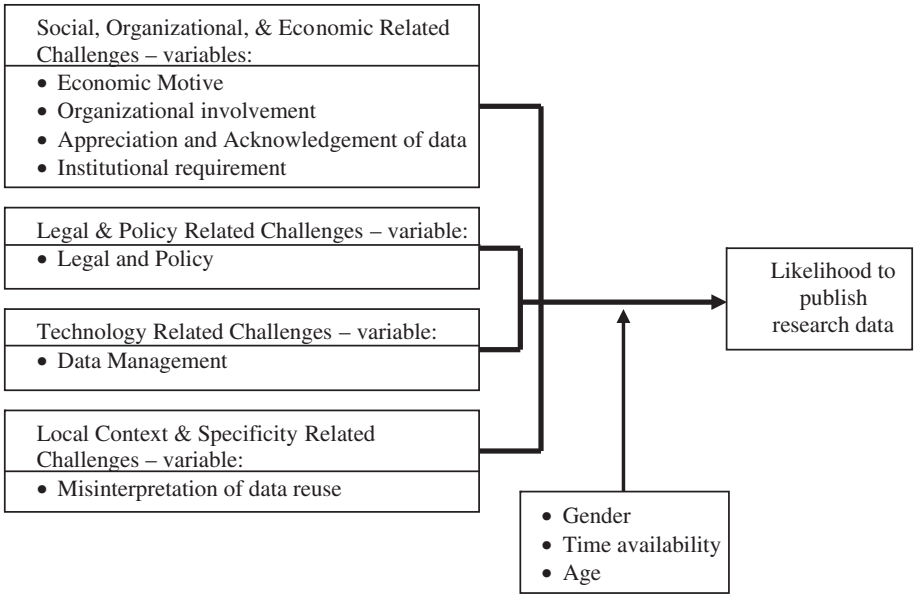


Fig. 1. The tested model.

respondents' distribution was comprised of a North American majority with 73% of the sample, 15% of participants from European regions, and the remaining participants from various other regions (see Table 1.).

3.2. Description of the variables

3.2.1. The dependent variable

The main dependent variable for this research study is the propensity of researchers to publish their data sets. This variable measured the extent to which researchers publish their data sets online or in a research network, and the amount of data that they are willing to publish. This variable consists of four sub-variables, namely: a) publish on PI (principal investigator) website, b) publish on organizational website, c) publish in regional research network, and d) publish in national research network. The measurement of this variable is based on a 4-point ordinal scale with a value of 1 indicating "does not publish data," value of 2 indicating "publishes some data," value of 3 indicating "publishes most data," and value of 4 indicating "publishes all of their data." The ordinal scale measurement of the dependent variable affects the choice of statistical analysis in this paper. In order to avoid biases, this paper uses ordered logistic regression as the main technique to examine the determinants of data publication.

3.2.2. The independent variables

There are seven independent variables of interest in this research study. These seven independent variables are: organizational involvement, data management skills, misinterpretation of data, legal regulations and conditions of use, appreciation and acknowledgement of data reuse, economic motives, and institutional requirements. Several of these variables are composite variables created from the survey questions.[6] These independent variables were measured using different scales.[7] The instruments to measure organizational involvement, data management skills, and researcher perspective of sharing data are measured using a 5-point Likert scale (1: strongly agree, 2: agree, 3: neither, 4: somewhat disagree, and 5: strongly disagree). The instruments to measure conditions for fair exchange of data and institutional requirements to publish the data are measured using dichotomous coding. As specified in the Introduction, we assume that the researcher/initiator of data who conducted the research is the owner of the data. The operationalization of these independent variables is as follows:

a. *Organizational involvement* refers to the level of organizational support provided to the scientist during and after the research project is completed. Organization support is a significant predictor of sharing behavior (see Bock et al., 2005; Lin, 2006). This variable is a composite variable derived from seven questions in the baseline survey pertaining to organizational support in terms of funding, training, technical support, storing data, and data management.
b. *Data management skill* refers to the researchers' level of skills in managing their data sets before openly publishing them online or in a network database. This variable is a composite derived from seven questions in the survey regarding skills in storing, cataloging, documenting, searching, and collecting data sets. These skills conform to the main activities in data management lifecycle such as discovering, collecting, describing, preserving, and analyzing (Ball, 2012; Hook, Santhana-Vannan, Beaty, Cook, & Wilson, 2010).
c. *Appreciation and acknowledgement of data reuse* refers to the form of acknowledgement given to the data owner for publishing his/her data sets online or in a network database. Incentives and rewards are important aspects to promote data sharing (Blumenthal et al.,

Table 1
Sample characteristics.

Characteristics	Category	%
Regional Distr. (region)	North America	73
	Europe	15
	Other regions	12
Age distr. (age)	20 to 39 years	37
	40 to 50 years	30
	Over 50 years	33
Status and position (status)	Professor or lecturers	49
	Grad./post doc students	19
	Researchers	21
	Other occupation	11

[6] The theme of each survey question is depicted in Tables 2 to 6 in Sections 4.2 to 4.5.
[7] The survey instrument was developed by the Usability and Assessment and the Socio-cultural working groups of DataONE and not by the authors.

1997; Koslow, 2000, 2002; Postle et al., 2002). This variable is a composite variable consisting of three questions from the survey pertaining to the form of acknowledgment for sharing data, namely: co-authorship on publications, formal acknowledgement, and collaboration opportunity.

d. *Legal and policy* refers to the legal requirements and policy in reusing the data sets placed by data set owners. A number of studies pointed at the significance of protecting privacy and security of data (see Tucker, 2009). This variable is a composite variable resulting from six questions on the survey to specify respondents' agreement for data reuse, namely: by placing conditions on access, acquiring approval from the data provider, asking for review from the data provider, giving a copy of product to the data provider, obtaining legal requirement for reuse, and acquiring a statement of use from the data provider.
e. *Misinterpretation of data reuse* refers to the concern of data owners that their data might be misinterpreted or misused. Researchers might fear that data is misinterpreted or misused owing to the complexity of research design and methodology (see Koslow, 2000, 2002). This variable is a composite variable resulting from three questions from the survey. These questions ask for the respondent's agreement on the statement about the possibility of data misinterpretation due to complexity of data, data misinterpretation due to poor quality of data, and data being misused.
f. *Economic motive* refers to the extent to which the data owner requires recovery of part of the cost to generate the data sets. This variable is based on the assertion that to some researchers, research data entails financial interests (Birnholtz & Bietz, 2003), be they the cost of collecting and preparing the data (Blumenthal et al., 1997) or potential income accrued from the research data (Rodriguez, 2009). This variable is not a composite variable.
g. *Institutional requirement* refers to the degree to which the primary funding agency requires the data owner to publish the data. Institutional factors could constrain individual researchers' ability to share data (Tenopir et al., 2011). This variable is a dummy variable and not a composite variable.

In addition to the seven independent variables of interest, this research study also includes three control variables, namely: age of researchers, gender of researchers, and time availability of the researcher.

3.3. Analysis tools

This paper uses four different techniques to construct and analyze the data. One technique – factor analysis – is used to construct the data sets. The other three techniques – network visualization, descriptive statistics and ordered logistic regression – are used to analyze the data. The main analysis technique in this paper is using the ordered logistic regression to correlate the determinants with the propensity of researchers to publish their research data. The choice of these four techniques is motivated by the two underlying reasons, 1) to meet the research objective of analyzing the determinants affecting researchers' motivation to share their data and 2) to conform to the construct of the data. The construct of the data will influence the appropriateness of the analysis technique. For instance, a social network is used because this technique is regarded as the most appropriate tool to study a direct understanding of social structures (Wellman & Berkowitz, 1988, p.3) and "offers a more powerful way of describing social interactions" (Emirbayer & Goodwin, 1994, p. 1413). Similarly, the ordered logistic regression is used, due to the categorical characteristics of the dependent variable. The use of ordinary linear regression, could be problematic, as it tends to induce bias for dependent variables with categorical properties (Scott Long & Freese, 2006). A brief description of each technique is presented below.

3.3.1. Factor analysis

Factor analysis is used to create the composite score for each of the seven independent variables. The factor analysis is run for each set of questions which pertain to a particular variable. The composite score, predicted from each factor with the highest eigenvalue, is used as the score of the independent variable. A composite can be expected to have higher reliability. Each of the dependent variables in this research is generated from the factor with the highest eigenvalue. The analysis found that each set of questions only loaded to a single factor. Collapsing the sub-variables into a single composite variable is necessary to avoid multicollinearity issues and freeing the degree of freedom (Thompson, 2004).

3.3.2. Network visualization

The purpose of this analysis is to provide a visualization of respondents' data access patterns. This is a supporting analysis to depict the current practices in accessing data across different data repositories. The use of a 2-mode network is due to the fact that the connection among actors is based on their use of a particular data repository. The visualization of the 2-mode network of data access on a particular network or databases refers to the degree of relationship between researchers, mediated through their access to particular networks or databases. The visualization is generated using NetDraw.

3.3.3. Descriptive statistics

Descriptive statistics of survey responses are provided for each of the seven independent variables. The descriptive statistics are used to characterize the challenges identified from the survey results on an aggregate level.

3.3.4. Ordered logistic regression

The ordered logistic regression is the main analysis technique in this paper. The ordered logistic regression is used to predict the likelihood of researchers to publish their research data sets online or in a network database, given the set of determinants (the independent variables). This paper uses ordered logistic regression because the dependent variable is measured on an ordinal scale. When the dependent variable is measured by categorical value, the use of an ordinary regression could induce bias (Scott Long & Freese, 2006).

4. Result

4.1. The disconnected clusters of access pattern

Efforts to create a network to promote national and international collaborative research are not new. The U.S. LTERN (Long Term Ecological Research Network), for instance, was founded by the NSF in 1980 to serve this purpose. Yet, the challenge of creating global collections of research data is compounded by the diversity, size and complexity of those data sets. This diversity and complexity is arguably influencing the sharing of research data sets. This section presents a mapping of patterns of access to research network databases to provide a visualization of the behaviors in sharing research data sets.

The network visualization is generated from the 2-mode network of data access patterns (see Fig. 2). This network is created from the affiliations of researchers with particular network databases, such as LTERN, NEON, or GBIF. The findings from this network visualization show the existence of many disconnected clustered relationships. The majority of researchers (the circles) used only one network database (the square) with only a few of the researchers connecting or accessing more than one database. The disconnected cluster suggests that sharing is actually somewhat limited and that a number of specific factors that limit data sharing may be operating in the network. This suggests that there exist certain challenges that restrict the propensity of researchers to share. In this regard, the next section

outlines the challenges, as identified in survey results, of sharing research data sets openly with the public.

4.2. The organizational challenges to data sharing

This section outlines the organizational related challenges identified through the survey results. Respondents were asked to express to what degree they agreed with the importance of the following items: 1) organizational involvement in supporting data sharing; 2) the importance of appreciation and acknowledgement for the data sets owner; and 3) the importance of having part of their research and collections costs recovered.

In terms of organizational involvement, the survey results indicated that 57% of the respondents noted the importance of receiving technical support for data management during the period of the project, while only 43% noted the importance of technical support for data management beyond the life of the project. Similarly, 51% of the respondents noted the significance of having organizational support in terms of data management. In addition, 46% of the respondents found organizational support for data storage during the life of the project to be important. On the other hand, organizational funding and training support for data management were regarded as less important determinants. By averaging the respondents' responses on the two questions related to funding, between 60% and 70% of the responses showed organizational support for funding and training for data management to be of lesser importance than the other determinants.

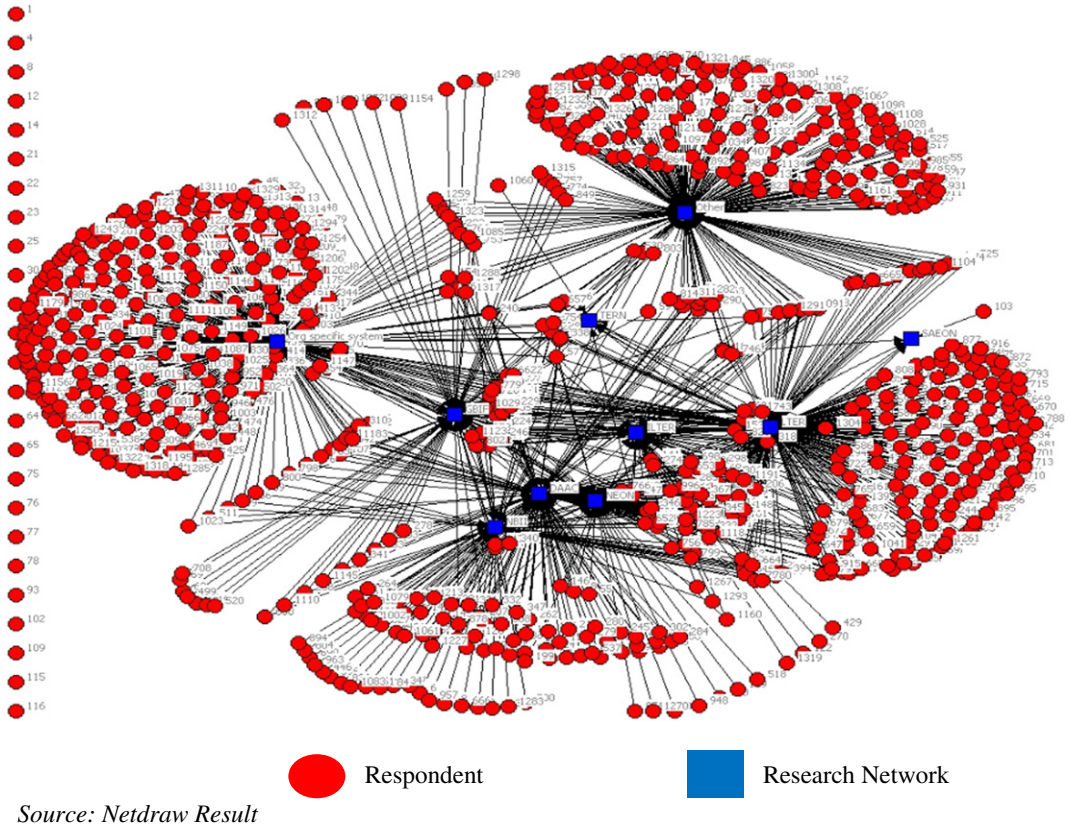
The majority of respondents expressed the importance of appreciation and acknowledgement of data owners in the process of reusing data; 93% indicated the significance of a formal acknowledgement to the data owner. The respondents also indicated the importance of collaboration in terms of co-authorship on publications (60%). The majority of respondents (70%) noted that acknowledgement in terms of cost recovery or other economic purposes is not important.

4.3. Technology related challenges

In this study the technological barriers were measured in terms of data management skills and data management tools. An average of the responses from three questions on searching, collecting, and cataloging data, shows that 85% of the respondents seem to agree on the importance of data management skills in the form of searching, collecting, and cataloging the data. 85% of the respondents also expressed the significance of skills in storing research data sets during the life of the project, but the significance of this skill becomes less important (56%) after the project is over. Interestingly, 82% of respondents pointed out that the lack of data standards is not an important determinant in the propensity to publish research data sets online (Table 4).

4.4. Legal and policy-related challenges

In terms of legal and policy issues related to publishing research data sets online, the majority recognize the significance of citation, legal agreement, statement of use, conditions of use, and approval for reuse. 98% of the respondents indicated the importance of citations when using the data sets from another researcher. In addition, 62% indicated that a review from data set owners should be required before results of data sets' reuse are disseminated, and that reprints of products from data sets' reuse must be provided to data set owners (70%). The respondents also indicated that statements of use and agreement on statements of use should be required before reusing data sets (67%). In addition, the respondents agreed that legal agreements should be obtained for data sets' reuse (45%) and approval



Source: Netdraw Result

Fig. 2. Data access pattern.

from the owner of data is a must (48%). If the respondents are required to publish their data sets, 82% of respondents asserted that conditions should be placed on access when data sets are made available (Table 5).

4.5. Local context and specificity challenges

The majority of respondents acknowledged the presence of a fear of data misinterpretation and misuse. 91% of respondents indicated concerns that data may be used in ways that were not initially intended. Additionally, the majority of respondents noted that their fear that data may be misinterpreted came from either the poor quality of the data (87%) or because of the complexity of the data itself (90%) (Table 6).

The descriptive statistics show that the majority of respondents found several potential challenges to be significant determinants of publishing research data sets openly to the public. These descriptive results support the initial visualization of the 2-mode network which highlights the potential of several determinants to be challenges and inhibitors of data sharing.

The next step employs stochastic-based statistics to ascertain the suggested findings from the descriptive statistics and visualization. This involves an examination of the relationship of selected variables from the four categories of challenges to the likelihood of researchers to publish their data sets online or in a research network identified previously.

4.6. The determinants and likelihood of sharing data

4.6.1. The significant determinants of publishing research data sets

Table 7 presents the ordered logistic regression used to test the relationship of selected challenges to the likelihood of researchers to publish their data sets on the public web or in research networks. There are four models in Table 7 pertaining to four different publication outlets, namely: organization website, principal investigator website, national research network, and regional research network. The number in parentheses represents the standard error of coefficient and the number above it represents the coefficient estimate. The coefficient of estimate of each variable is in log-odds ratio. In logistic regression, the coefficient estimate does not provide a meaningful interpretation without transforming the result. For instance, the estimate of organizational involvement in the organizational website model is −0.3763 (Table 7). Direct and brute interpretation of the estimate will be that a one-unit increase in organizational involvement is associated with a 0.3763 decrease in the ordered log-odds of publishing data sets on an organizational website. For this reason, and considering the objective of this study which is to identify significant determinants, we focus on the significance test of each variable. Subsequently, we will transform the estimates to measure the probability of publishing data sets and the influence of significant determinants on this probability.

The results in Table 7 show that organizational involvement consistently appears as a significant determinant of publishing data sets on an organizational website with a z value of −6.71, a principal investigator website with a z value of −2.56, a national level research network with a z value of −7.36, and a regional level research network with a z value of −5.26. Acknowledgement and appreciation for data reuse is not significant for all four publication outlets. Arguably, researchers do not consider acknowledgement and appreciation as an important determinant for publishing their research data online. On the other hand, economic motive emerges as a significant predictor for publishing data on an organizational website ($z = 1.70$) and national network ($z = 1.82$) at 0.10 significant levels and on the regional network ($z = 2.79$) at 0.05 levels.

Legal and policy requirements also emerge as a significant determinant of publishing data sets on an organizational website ($z = -2.99$) and a national level research network (−3.26) at 0.05 significant levels and on a principal investigator website ($z = -1.85$) at 0.10 levels. But, legal and policy requirements do not have a significant impact on publishing data sets in the regional network. Data management skills also appear to be a significant determinant for an organizational website ($z = 2.31$), national network ($z = 1.65$) and regional network ($z = 2.25$) while the result shows insignificant influence of data management skills for the principal investigator website.

It appears that the outlet for publishing is relevant in terms of the significance of only some determinants. Data misinterpretation is statistically significant in affecting the publication of research data sets (at a level of 10%), only if the data set is published in a research network ($z = 1.90$). Logically, publishing research data sets in a research network will have wider distribution and therefore, the researcher assumes that the data will have a greater chance of being misinterpreted. Coercive requirements from funding agencies are statistically significant at a level of 5% only for publication of data sets on a principal investigator website. Arguably, the principal investigator is the person responsible for complying with the funding agency requirement.

In terms of the control variables, the significant impact of the control variables varies especially for age and gender, while the availability of research time is shown to have a non-significant influence on publishing research data sets. Age has a significant influence for publishing research data sets on an organizational website ($z = 1.87$) and at the national network ($z = 1.79$). On the other hand, gender only has influence for publishing research data sets in an organizational website with a z value of 2.81 and in a principal investigator website with a z value of 1.70.

4.6.2. The different likelihood of publishing research data sets

Considering that the dependent variable consists of four sets of ordinal values representing the number of data sets respondents are

Table 2
Form of organizational support.

Form of organizational support	Support	
	Yes	No
Support on data management	51%	49%
Support on data storing during the project	46%	54%
Tech. support for data management during the proj.	57%	43%
Tech. support for data management beyond the proj.	43%	57%
Training for data management	27%	73%
Fund for data management during the project	39%	61%
Fund for data management beyond the project	27%	73%

Table 3
Form of acknowledgement.

Form of acknowledgement	Agreement	
	Yes	No
Co-authorship on publications	60%	40%
Formal acknowledgement of the data providers	93%	7%
Opportunity to collaborate with others	81%	19%
Part of cost of data must be recovered	30%	70%

Table 4
Technological challenges.

Technological challenges—data management	Satisfaction	
	Yes	No
Lack of data standards	18%	82%
Research data collection process	89%	11%
Research data searching process	86%	14%
Research data cataloging process	76%	24%
Research data storing during the project	85%	15%
Research data storing beyond the project	56%	44%

Table 5
Legal context and policy challenges.

Legal context and policy challenges	Agreement	
	Yes	No
Place a condition on access to make data available	82%	18%
Citation is required to use the data	98%	2%
Approval from data providers is a must for data reuse	48%	52%
Review from data providers is need for data reuse	62%	38%
Reprint copy of products must be given to data owner	70%	30%
Legal agreement should be obtained for data reuse	45%	55%
The data provider is agrees to a statement of uses	67%	33%
The funding agency requires data management plan	34%	66%

willing to publish, either on a website or in a research network, it is necessary to measure the probability of publishing data sets, contingent to the amount of data to be published. Table 8 represents the probability to publish data sets on a website or research network based on the mean values of the independent variables. Hence, if all independent variables are set at their mean value (and gender is set to 1, to represent male), the probability of not publishing data sets is higher than the probability of publishing some, most, or all of the data sets for male respondents. However, among the probability to publish some level of data, there is a higher probability to publish just some of the data sets, rather than publishing most or all of the data sets.

The findings in Table 8 suggest that the probability of publishing data sets either on a website or in a research network depends on the amount of data a researcher had to publish. There is an inverse relation between probability to publish and the amount of data to be published. The probability to publish a data set will decrease with an increase in the amount of data. Findings from Table 7 point out three variables that consistently emerge as significant predictors at 0.05 levels, namely: organizational involvement, data management skills, and legal and policy requirements. Considering both findings, a simulation is used to test the influence of the significant predictors (Table 2) to predict the change in probability to publish research data sets.

4.6.3. The amount of data and the likelihood to publish research data sets

Fig. 3 represents the change in probability to publish data sets on an organizational website depending on the three significant predictors, namely: organizational involvement, data management skills, and legal and policy requirements.

In general, there is a high probability of researchers not publishing data sets on an organizational website (see Table 8). However, the data shows that changes in the significant predictors will subsequently change the probability to publish. As shown in Fig. 3, an increased level of data management skills by one unit will move the likelihood from "not to publish" into "publish some." Furthermore, increased requirements to publish higher numbers of data sets (most or all data) is also expected to decrease the probability of publishing data sets on an organizational website. A change from "publish some" to "publish all" decreases the probability from 0.1 to approximately 0.04 with the inclusion of data management. In terms of legal and policy requirements, there is a high probability of not publishing data sets, but the data also shows that increased legal and policy requirements will change the probability from "publish some" into "publish all" of the data sets; from −0.1 for "publish some" to approximately −0.02 for "publish all". A higher level of legal and policy requirements correlates positively with the amount of data to be published. Additionally, organizational involvement will increase the likelihood that researchers will publish data if they are required to publish all of their data; increasing from approximately −.028 for "publish most" to −0.11 for "publish all" (see Fig. 3).

Table 6
Local context and specificity challenges.

Local context and specificity challenges	Agreement	
	Yes	No
Misinterpretation of data due to complexity of data	90%	10%
Misinterpretation of data due to poor quality of data	87%	13%
Data may be used in other ways than intended	91%	9%

5. Discussion

Connecting back to the larger picture of NSF's global vision of a Cyberinfrastructure for 21st Century Discovery, the analysis in this paper presents a new understanding of challenges to the development of global data collections and the preservation of those collections. Creating the open system is arguably influenced not only by new methods, management structures and technologies but by the human aspect as well. The willingness of researchers and data owners to participate in the creation of this global network through the sharing of their data affects the achievement of the above-mentioned vision. This paper contributes to understanding the propensity of researchers to share their data. For its preliminary assessment, this paper visualized the affiliated network based on the survey results (Fig. 2). The visualization shows a disconnected access pattern where only very few researchers are connecting to more than one database. This result might raise questions of specific factors limiting data sharing that may be operating in the network. The survey results, analyzed using descriptive and inferential statistics, identify several key determinants affecting researchers' motivation to share their research data sets openly, namely: data management, organizational involvement, legal and policy requirements, and acknowledgement to the data set's owner.

5.1. Data management as a crucial element of publication of research data sets

The significance of data management is captured in two determinants, namely: 1) data management skills and 2) organizational involvement. The survey questions for organizational involvement pertain to the organizational support for data management (see Table 2). The survey results indicate that 80% of respondents emphasized the importance of data management skills and 51% noted the importance of organizational involvement to support data management. The ordered logistic regression results also support this finding; data management skills and organizational involvement emerge as significant determinants of the likelihood of researchers to publish research data sets on a website or research network. This study offers three plausible explanations in regard to the significance of data management, namely: 1) achieving certain levels of data quality, 2) maintaining reputation and merit and 3) resource and time constraints.

The literature points out the inherent complexity and heterogeneity of the data, especially ecology and earth and environmental data (Reichman et al., 2011; Tenopir et al., 2011; Zimmerman, 2007). With this inherent complexity and heterogeneity, achieving certain levels of data quality before publishing data openly becomes necessary to avoid data being misinterpreted. Achieving the benefits of an open data initiative is subject to the quality and format of the data (Vogel, 2011), it must be credible, usable and interpretable (Hinrichs & Aden, 2001). The researchers need to take precautionary actions to make sure that their data will not be misinterpreted. As a result, the description of relevant information needs to be clearly specified and explained in the data sets (Koslow, 2000, 2002). Thus, the researchers need to acquire certain levels of data management skills to help them prepare the data for open publication. The result indicates that data management skills are a significant predictor influencing researchers' willingness to publish their data on the

Table 7
The odds of publishing data on the website (in odd ratio).

Variable	Website		Research network	
	Organizational	Principal invest.	Regional	National
Organizational involvement	−0.3763 (0.0560)*	−0.1411 (0.0550)*	−0.3177 (0.0604)*	−0.4354 (0.0591)*
Appreciation for data re-use	0.0329 (0.0884)	0.0159 (0.0854)	0.1248 (0.0919)	0.0767 (0.0855)
Economic motive	0.1487 (0.0876)**	0.0671 (0.0923)	0.2493 (0.0895)*	0.1461 (0.0804)**
Data management skills	0.1281 (0.0555)*	−0.0221 (0.0547)	0.1310 (0.0581)*	0.0889 (0.0538)**
Legal and policy requirements	−0.2419 (0.0807)*	−0.1474 (0.0797)**	−0.1189 (0.0732)	−0.2435 (0.0748)*
Institutional requirements	0.0507 (0.0886)	0.1845 (0.0858)*	−0.1000 (0.0873)	−0.0014 (0.0793)
Misinterpretation of data	0.0174 (0.0701)	0.0905 (0.0690)	0.1326 (0.0697)**	0.0962 (0.0812)
Availability of research time	0.4307 (0.3712)	−0.2062 (0.3479)	−0.1442 (0.3982)	0.1214 (0.3692)
Age	0.0155 (0.0083)**	0.0016 (0.0079)	0.0036 (0.0085)	0.0145 (0.0081)*
Gender	0.5817 (0.2069)*	0.3372 (0.1986)*	0.2683 (0.2138)	0.3138 (0.2029)
Wald Chi^2	66.44	26.36	41.64	73.84

Standard errors in parentheses.
* Significant at 0.05 level.
** Significant at 0.1 level.

organizational website, national network and regional network. This result is interesting since data management skills were found to be not significant for open publication in the principal investigator (personal) website (Table 7). Anecdotally, this might relate to the magnitude of impact for a different outlet in the sense that data publication in organizational and national or global networks has a greater impact than on a personal website. Hence, researchers are more concerned with preparing their data for outlets with larger impact than those with less impact. Yet, this assertion warrants further investigation to understand more about the value of data quality as a predictor of open publication.

Likewise, this finding supports the argument on the significance of merit and reputation for researchers' willingness to publish their data (Ceci & Walker, 1983; Koslow, 2000, 2002). Developing and preparing quality data for publication is closely related to researchers' incentives to maintain their reputation and merit. Birnholtz and Bietz (2003) argued that competition for merit and reputation is more aggressive in academe. One of the indications of researchers' prominence and quality is their ability to understand the data and generate publications based on the data (Tucker, 2009). Thus, when publishing their data openly to the public, the researchers were faced with several concerns, such as 1) someone else could find different results and something that the researcher missed in the previous publication that would undermine their work (Koslow, 2000, 2002) and 2) someone else could find procedural and computational errors from their previous publications (Ceci & Walker, 1983). Arguably, ensuring data quality before publishing it openly becomes an important factor to save researchers from future humiliation regarding errors in the data. Thus, researchers value data management skills and organizational support for data management as significant predictors of their action to publish data.

On the other hand, researchers are always faced with resources and time constraints. Data, materials, and storage facilities are costly. Active researchers have many activities to do that restrict them in preparing their data by themselves. Researchers need assistance in preparing their data for open publication. Such supports could be in two forms: technology and/or organizational support. Michener (2006) argues for an increase in usability and user-friendliness of metadata management to help alleviate the burden of preparing data. On the other hand, an organization could play a significant role in helping researchers to prepare their data for publication. The survey results indicate the need to have support for data storing (47%), technical support (57%) and support for data management (51%). Organizational supports for data management could help researchers preserve the quality of their data and avoid data misinterpretation.

Table 8
Probability to publish data sets.

Probability	Website		Research network	
	Org.	PI	Regional	National
Pr(not_publish)	55%	62.6%	68%	52.6%
Pr(publish_some)	34.5%	28.9%	26%	34.5%
Pr(publish_most)	7.5%	5.8%	4.6%	10%
Pr(publish_all)	3%	2.6%	1.4%	2.9%

5.2. The significance of acknowledgement

The significance of acknowledgement is captured in two variables, namely: forms of acknowledgement and legal and policy requirements. Although the logistic regression result shows that the forms of acknowledgement are not statistically significant (Table 7), the survey results pointed at the high percentage of respondents agreeing on the importance of certain forms of acknowledgement. Due to this, we argue that there is suggestive evidence for the importance of acknowledgement. The survey results indicate that, on average, 90% of respondents expressed the importance of acknowledging data providers. The survey results also indicate a high percentage of agreement on the various forms of acknowledgement, including opportunities for collaboration, co-authorship, formal recognition, and proper citation (Table 3). This result suggestively supports the assertion of extant literature arguing for the importance of appreciation and acknowledgement of the data owner (Birnholtz & Bietz, 2003; Ceci & Walker, 1983; Koslow, 2000, 2002). The fact that the logistic regression result does not conform with the survey indicates a venue for future research to understand better the impact of reward and acknowledgement. The compositions of acknowledgement in the survey instruments only focus on four aspects, namely: co-authorship, formal acknowledgement, cost recovery and legal permission to use. The survey instrument did not consider the

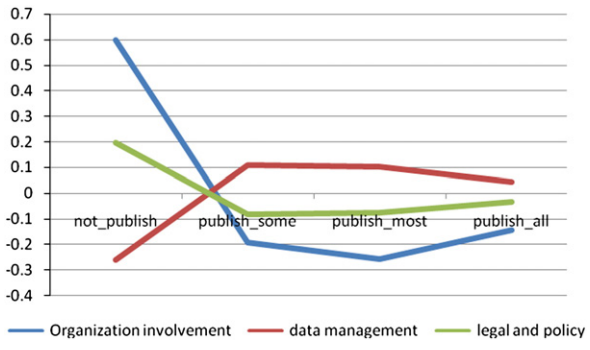


Fig. 3. Change in probability to publish in organizational website.

intrinsic reward as another form of acknowledgement, while extant literature argued for the significance of reputation and merit for researchers.

As mentioned above, the majority of respondents indicated the significance of appreciation and acknowledgement by emphasizing the importance of legal and policy requirements to support data re-use and attribution. On average, 70% of the respondents asserted the need to have in place a proper legal and policy framework before making their data sets open to the public. The ordered logistic regression result also shows that the existence of a legal and policy framework is a significant predictor of a researcher's likelihood to publish his or her data sets on a website or in a research network.

Regulation and policy requirements could enhance data sharing by ensuring proper and accountable use of data and information (Zimmerman, 2007). Similarly, the results show that respondents are emphasizing legal and policy requirements to ensure that data are not poached and owners are properly acknowledged. Knowing that a legal framework exists to support and guide data sharing might provide assurances for data set owners that their data sets will not be misused, and attribution upon re-use can be ascertained which will result in a higher propensity to share. The literature also indicates the importance of an appropriate balance in legal and policy requirements. Rigidity of these requirements could deter sharing; yet lenient legal policies also can lead to not sharing due to the fear of liabilities (Savage & Vickers, 2009; Tucker, 2009). In this regard, much can be learned from the interagency information sharing literature in the public administration domain. For instance, Dawes (2010a, 2010b) proposes a useful framework as a guide for information policies in information sharing across different agencies. Similarly, in the public sector, the needs to share information across different agencies are hindered by the need to protect sensitive data and privacy rights. In this regard, the need to use shared information conflicts with the need to guarantee that the same information is not being misused (Dawes, 1996). Dawes proposes the balancing of two complimentary requirements in the framework, namely: information stewardship and information usefulness (Dawes, 2010a, 2010b). Information stewardship conveys the idea of safe handling of the information while information usefulness refers to the principle that more accessible information provides a wide variety of public and private uses.

5.3. Implications

5.3.1. Policy and management implications

The findings show that data management emerges as one of the key determinants of open data publication. In addition, the study also found an inverse relationship between the probability to publish online and the amount of data to be published. The previous Section 5.1 extensively argues for the connection between data management, data quality and time and resources constraint. These findings present implications for policies and management supporting open data initiatives.

a. Systematic investigation on the roles and responsibility of government within these global scientific data preservations. Extant literature pointed out that researchers have the upper hand in deciding to release or withhold the data due to various reasons, such as: cost, lack of resources, reputation, or financial aspect (Blumenthal et al., 1997; Campbell et al., 2002; Vision, 2010). On the other hand, more efforts are on the way to govern the data produced by publicly funded research. The movement toward "open initiatives" could plausibly change the culture of sharing among researchers. In time, publicly funded data will no longer be regarded as private intellectual property. Thus, more research is needed to understand the implications of change in this culture. In addition, with an increased effort on global data repositories and access, it will be necessary to further examine the roles and responsibilities of governments as they relate to the governance of this new practice. For instance, despite the growing efforts to position publicly funded research data as public good, policy makers need to clearly specify the conceptualization of data ownership within a global data repository.
b. The result shows that the amount of data matters. The level of data to be published significantly correlates to the likelihood to publish the data. Linking this issue to the time and resource limitations and the time needed to prepare the data, publishing large amounts of data required large chunks of time commitments. This condition led to two implications:

1) Policy makers and managers of federated data repositories such as DataONE need to take into account the amount of data researchers are required to publish. For instance, the development of policies in setting a requirement on the amount of data could be taken into consideration. Policymakers and managers of federated data initiatives need to be able to balance their policies with the amount of data researchers are required to publish.
2) Another policy implication related to the difficulties in preparing the data for publication is specifying which types of metadata principles the repositories used. Different researchers might use different metadata principles when preparing their data. As a result, policymakers of federated data repositories should consider specifying policies to either accommodate all different metadata principles or provide tools to transform different metadata principles.
3) Considering the difficulties related to preparing data sets for publication, the managerial implication for managers of federated data repositories is to consider the level of support provided to the researchers when publishing in certain repositories. For instance, virtual data management trainings and supports could be created and applied to guide researchers in publishing their data.

c. For researchers, appreciation and acknowledgement are significant factors affecting their sharing behavior. In addition, policy makers also need to consider that reward and appreciation have various forms in academe including recognition for merit and reputation. Policies are needed to ensure appropriate recognition of the originator of the data and of those who are responsible for the re-use of the data and to include factors such as reputation as consideration.
d. There exist large numbers of dialogues, discussions and literature in the fields of public administration and information science offering rich insights on the information/data policies and securities. This literature could provide a useful framework to understand the trade-off between policies and flexibility of sharing for earth and environmental science. This paper cited one of them, a framework proposed by Dawes (2010a, 2010b).

5.4. Limitations and future research

This study attempts to provide a preliminary assessment on the determinants of a researcher's motivation of sharing data sets from the perspective of a personal decision. As a preliminary assessment, this study identifies and acknowledges various limitations which warrant future research.

a. This study is based on the assumption that the researchers have a free role in deciding what and when to publish their raw data sets. On the other hand, individual researchers might be constrained by their institutional policies in sharing the data (Tenopir et al., 2011). This study does not differentiate between institutional and individual constraints due to the limitation in the data. Still, this issue warrants further research, adding new institutional variables as mediating variables to understand the impact of institutional restriction.
b. This study found that data management skills is significant for open publication of data except on the principal investigator website. An anecdotal assertion is that researchers are more concerned with

data quality when publishing in outlets with larger impact, thus requiring data management skills. Contrastingly, the researchers assume that publishing on a personal website requires less effort to ensure data quality. This assertion demands further research to better understand the value of data quality for publication for the researchers.

c. Extant literature shows the importance of reputation and merit as one form of appreciation and acknowledgement. Yet, the survey instruments do not account for this form of acknowledgement. Considering the possible importance of intrinsic rewards such as reputation and merit, future surveys could consider including this factor in their instrument.
d. This study only focuses on the propensity of researchers to supply data to an open data initiative such as DataONE. The success of sharing depends on both the initiator/owner and the requester (Rodriguez, 2009). Future research could consider integrating both the supply and the demand aspect of data sharing.
e. We recommend future research to consider using the adoption theory as the theoretical basis, such as: the Theory of Planned Behavior, the Technology Adoption Model, the United Theory of Acceptance and Use of Technology, or other adoption theories. For instance, using the Theory of Planned Behavior by Ajzen (1991), we can examine the determinants which could be influencing researchers' intentions to share data, as well as assessing to what extent the researchers' intentions lead them to an action to share data.
f. Different domains might yield different behavior in regard to sharing research data sets openly. This study does not account for the impact of different domains into the equation. Future research should consider integrating the different behavioral impacts across domain.
g. The authors acknowledge that almost 75% of the distribution of survey respondents was from the North American region. Thus, the findings of this research study are best interpreted as evidence of North American researchers' motivation to share data sets. Further research is needed to establish the generalizability of these findings in other regions of the world.

6. Conclusion

NSF's vision of global scientific discovery through global data preservation and access entails major challenges. These challenges rest on human aspects as well as the new methods, management structures, and technologies. Achieving the above vision rests in large part on the researchers' willingness to contribute to this vision by sharing their data. This research study provides a preliminary analysis of the determinants of the likelihood of researchers to publish their research data sets online. Survey results were analyzed using descriptive and inferential statistics. The analysis identified two key determinants for sharing research data sets. Each of the determinants is supported by two challenges, namely: 1) data management in terms of skill and organizational involvement, and 2) acknowledgement in terms of legal and policy requirements and acknowledgement to the data set's owner.

The importance of data management to ensure open publication of data manifests not only in the significance of data management skills, but also in terms of organizational involvement to provide support for data management. Data management skills and support for data management are necessary elements for ensuring data quality. The second key determinant found is the significance of proper attribution to data set owners. The importance of attribution in this study manifests in various forms of articulated acknowledgement of the data owners, such as opportunities for collaboration, co-authorships, formal recognition, and proper citation. Equally important are legal regulations and policies to support data re-use and attribution.

This finding highlights the significance of creating new capabilities to share data and to deal with the issues of sharing data. The capability of researchers to manage their data sets and deal with the issues of data quality, standards, and protection, in addition to ethical and responsible use of shared data will significantly influence their propensity to share their data.

Acknowledgments

The authors would like to acknowledge the Sociocultural working group and the Usability and Assessment working group of the DataONE project for their permission to use the project's data and information to enhance the presented arguments. The authors would also like to thank the anonymous reviewers, for their insights and comments in giving a clearer focus to this paper.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211, http://dx.doi.org/10.1016/0749-5978(91)90020-T.
Ardichvili, A., Page, V., & Wentling, T. (2003). Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Journal of Knowledge Management*, *7*(1), 64–77.
Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., et al. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, *3*, 135–152.
Ball, A. (2012). *Review of data management lifecycle models.* Research Report, Bath, UK.
Barnes, D. M. (1987). Meeting on AIDS drugs turns into open forum. *Science*, *237*(4820), 1287.
Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: Supporting sharing in science and engineering. *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 339–348).
Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N., & Louis, K. S. (1997). Withholding research results in academic life science. *JAMA:The Journal of the American Medical Association*, *277*(15), 1224–1228
Bock, G., Zmud, R. W., Kim, Y., & Lee, J. (2005). Behavioral intention formation in knowledge sharing: examining the roles of extrinsic motivators, social–psychological forces, and organizational climate. *MIS Quarterly — Special Issue on Information Technologies and Knowledge Management*, *29*, 87–111.
Borgman, C. L. (2010, September). *Research data: Who will share what, with whom, when, and why?* Paper presented at the 5th China—North America Library Conference, Beijing. Retrieved from http://works.bepress.com/borgman/238.
Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., et al. (2007). From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, *12*(2), 652–672.
Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., et al. (2002). Data withholding in academic genetics. *JAMA:The Journal of the American Medical Association*, *287*(4), 473–480.
Ceci, S. J. (1988). Scientists' attitudes toward data sharing. *Science, Technology & Human Values*, *13*(1/2), 45–52.
Ceci, S. J., & Walker, E. (1983). Private archives and public needs. *American Psychologist*, *38*(4), 414.
DataONE (n.d.). Retrieved March 27, 2012, from http://www.dataone.org/.
Dawes, S. S. (1996). Interagency information sharing: Expected benefits, manageable risks. *Journal of Policy Analysis and Management*, *15*(3), 377–394.
Dawes, S. S. (2010a). Stewardship and usefulness: Policy principles for information-based transparency. *Government Information Quarterly*, *27*(4), 377–383.
Dawes, S. S. (2010b). Information policy meta-principles: Stewardship and usefulness. *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* (pp. 1–10).
Ding, W. W., Levin, S. G., Stephan, P. E., & Winkler, A. E. (2010). The impact of information technology on academic scientists' productivity and collaboration patterns. *Management Science*, *56*(9), 1439.
Emirbayer, M., & Goodwin, J. (1994). Network analysis, culture, and the problem of agency. *The American Journal of Sociology*, *99*(6), 1411–1454.
Hey, A. J. G., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: Data-intensive scientific discovery.* WA: Microsoft Research Redmond.
Hinrichs, H., & Aden, T. (2001). *An ISO 9001: 2000 compliant quality management system for data integration in data warehouse systems. International Workshop on Design and Management of Data Warehouses.*
Hook, L., Santhana-Vannan, S. K., Beaty, T. W., Cook, R. B., & Wilson, B. E. (2010). *Best practices for preparing environmental data sets to share and archive.* Manuscript. Oak Ridge National Laboratory.
Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics—Re-shaping scientific practice. *Nature Reviews Genetics*, *10*(5), 331–335.
Koslow, S. H. (2000). Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neuroscience*, *3*(9), 863.
Koslow, S. H. (2002). Sharing primary data: a threat or asset to discovery? *Nature Reviews Neuroscience*, *3*(4), 311–313.
Krueger, A. O. (1974). The political economy of the rent-seeking society. *The American Economic Review*, *64*(3), 291–303.
Kuipers, T., & van der Hoeven, J. (2009). *Insight into digital preservation of research output in Europe (Survey Report No. D3.4). PARSE Insight: PARSE Insight.* (Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf)
Lin, H. (2006). Impact of organizational support on organizational intention to facilitate knowledge sharing. *Knowledge Management Research and Practice*, *4*, 26–35.

Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, *1*(1), 3–7.
Nelkin, D. (1982). Intellectual property: The control of scientific information. *Science*, *216*(4547), 704–708.
Nelson, B. (2009). Data sharing: Empty archives. *Nature*, *461*.
Pardo, T. A., Cresswell, A. M., Dawes, S. S., & Burke, G. B. (2004). *Modeling the social and technical processes of interorganizational information integration. Proceedings of the Hawaiian International Conference on System Sciences.* Hawaii: IEEE Computer Society.
Parr, C. S., & Cummings, M. P. (2005). Data sharing in ecology and evolution. *Trends in Ecology & Evolution*, *20*(7), 362–363.
Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research data sets: A pilot study of associations. *Journal of informetrics*, *4*(2), 148–156.
Postle, B. R., Shapiro, L. A., & Biesanz, J. C. (2002). On having one's data shared. *Journal of Cognitive Neuroscience*, *14*(6), 838–840.
Reichman, O., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703.
Rodriguez, V. (2009). Access to data and material for research: Putting empirical evidence into perspective. *New Genetics and Society*, *28*(1), 67–86.
Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS One*, *4*(9), e7078.
Schofield, P. N., Bubela, T., Weaver, T., Portilla, L., Brown, S. D., Hancock, J. M., et al. (2009). Post-publication sharing of data and tools. *Nature*, *461*(7261), 171–173.
Scott Long, J., & Freese, J. (2006). *Regression models for categorical dependent variables using STATA.* College Station, TX: Stata Press.
Smith, K., Seligman, L., & Swarup, V. (2008). Everybody share: The challenge of data-sharing systems. *Computer*, *41*(9), 54–61.
Stanley, B., & Stanley, M. (1988). Data sharing. *Law and Human Behavior*, *12*(2), 173–180.
Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.
Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., et al. (2011). Data sharing by scientists: Practices and perceptions. *PLoS One*, *6*(6), e21101.
Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.
Tucker, J. (2009). *Motivating subjects: Data sharing in cancer research.* Virginia Polytechnic Institute and State University.
Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative knowledge work and practices of trust: Sharing environmental planning data sets. *Proceedings of the 1998 ACM conference on Computer Supported Cooperative Work* (pp. 335–343).
Vickers, A. (2006). Whose data set is it anyway? Sharing raw data from randomized trials. *Trials*, *7*(1), 15.
Vision, T. J. (2010). Open data and the social contract of scientific publishing. *BioScience*, *60*(5), 330–331.
Vogel, L. (2011). Opening the gates on US government data. *Canadian Medical Association Journal*, *183*(7), E377–E378.
Wellman, B., & Berkowitz, S. D. (1988). *Social structures: A network approach*, *Vol. 2*, Cambridge Univ. Press.
Zhang, J., & Dawes, S. S. (2006). Expectations and perceptions of benefits, barriers, and success in public sector knowledge networks. *Public Performance & Management Review*, *29*(4), 433–466.
Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1), 5–16.
Zimmerman, A. (2008). New knowledge from old data. *Science, Technology & Human Values*, *33*(5), 631–652.

Djoko Sigit Sayogo is a PhD candidate in the Rockefeller College of Public Administration and Policy, at the University at Albany, SUNY. His research interests include e-government, collaborative network, and data sharing.

Theresa A. Pardo is director of the Center for Technology in Government and is a faculty member of the Rockefeller College of Public Administration and Policy and is affiliate faculty member in the Informatics Program of the College of Computing and Information at the University at Albany, State University of New York.