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).