Information Processing and Management 53 (2017) 871–885

Contents lists available at ScienceDirect

Information Processing and Management

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

Fostering scientists' data sharing behaviors via data repositories, journal supplements, and personal communication methods

Youngseek Kim

University of Kentucky, School of Information Science, 331 Little Library Building, Lexington, KY 40506, United States

ARTICLE INFO

Article history:
Received 2 November 2016
Revised 15 March 2017
Accepted 21 March 2017

Keywords:
Data sharing
Data reuse
Cyberinfrastructure
e-Science
Data repository
Institutional theory

ABSTRACT

The purpose of this study is to examine how institutional pressures, individual motivations, and resources all affect scientists' diverse data sharing behaviors, including (a) making data accessible through data repositories, (b) submitting data as journal supplements, and (c) providing data via personal communication methods upon request. A combined theoretical framework integrating institutional theory and theory of planned behavior was used to create a research model which presents how scientists make the decision to share data in diverse ways, and how the data sharing factors differ across diverse data sharing behaviors. A survey method was employed to evaluate the research model by using multivariate regression analysis technique with a total of 2172 survey responses in the U.S. The results of this research show the dynamic relationships between diverse data sharing factors and different forms of data sharing behaviors. For data sharing via data repository, journal pressure, perceived effort, and availability of data repositories are significant factors; for data sharing through journal supplement, journal pressure, perceived career benefit, perceived effort, and availability of data repository are significant factors; for personal data sharing, funding agency pressure, normative pressure, perceived career risk, perceived effort, and availability of data repositories are significant factors. This research suggests that funding agencies, journal publishers, and scientific communities that different strategies need to be employed to promote different forms of data sharing behaviors.

1. Introduction

Data sharing is essential to contemporary scientific research from the perspective of e-Science and open science. The term e-Science is defined as "networked and data-driven science," (Hey & Hey, 2006) and a critical aspect of it centers on global collaboration in key areas of science being enabled by data-centric scientific research based on shared data sets (Hey & Trefethen, 2002). Open science refers to conducting research in a collaborative manner by sharing and reusing research data and relevant materials (FOSTER, 2016). Both e-Science and open science promise to reshape and enhance the way science is done by empowering data-driven scientific research and improving the synthesis and analysis of scientific data in a collaborative fashion (Atkins, 2006; Molloy, 2011). The recent advancement in information and communication technologies such as data repositories and personal communication methods has enabled scientists to share their research data along

E-mail address: youngseek.kim@uky.edu

http://dx.doi.org/10.1016/j.ipm.2017.03.003
0306-4573/

with their research publications, thus achieving the core vision of e-Science and open science, which is data-driven science based on shared data sets.

Traditionally, formal scholarly communication is based on journal articles, conference proceedings, and sometimes article preprints. Recently, however, original research data has taken its place in formal scholarly communication. Scientific research now requires original data sets for diverse purposes such as large-scale computation, comparative research, or replication of previous works for further research. As primary data becomes important in terms of e-Science and open science, data sharing becomes critical to scientific research (Borgman, 2012; Tenopir et al., 2015). Scientific communities have developed diverse data repositories, and scientists have become more aware of the importance of data sharing (Borgman, 2010; Gewin, 2016; Tenopir et al., 2011, 2015). Furthermore, national funding agencies in the U.S. such as the National Science Foundation (NSF), National Institutes of Health (NIH), and National Cancer Institute (NCI) have mandated data sharing in many disciplines as a part of their grant requirements (NCI, 2006; NIH, 2007; NSF, 2012); funding agencies in European Union (EU) and United Kingdom (UK) such as European Research Council (ERC), Research Councils in UK (RCUK), and Wellcome Trust (WT) have also promoted open science research based on shared data sets and research articles (ERC, 2012; RCUK, 2016; WT, 2015). In addition, a number of journals have implemented data sharing policies (Piwowar & Chapman, 2008; Savage & Vickers, 2009). Despite this, data sharing is still not well-deployed throughout diverse disciplines as a common research practice (Tenopir et al., 2011; Tenopir et al., 2015; Wallis, Rolando, & Borgman, 2013).

Technological infrastructures, institutional set-ups, and individual motivations often contribute to scientists' data sharing behaviors. Contemporary collaboration in science and engineering fields requires the orchestration of technological infrastructure, institutional support, and interpersonal interactions (Kim & Stanton, 2012). Similarly, scientists' data sharing as the microcosm of contemporary collaboration involves the same three areas of infrastructure, institutions, and people. Individual scientists are embedded in institutional contexts, including belonging to universities and academic disciplines, and drawing support from organizational and disciplinary technological infrastructure. This research considers the combination of infrastructure, institution, and people as important components influencing scientists' data sharing, and examines how those factors influence diverse forms of data sharing behaviors.

The data sharing behaviors of scientists occur in diverse forms, including uploading data in data repositories, submitting data as journal supplements, and providing data via personal communication methods upon request. Since each discipline has its own data sharing practices, scientists' data sharing differs across disciplines. Furthermore, even in the same discipline, scientists' data sharing behaviors can vary because of their technological infrastructures, institutional set-ups, and individual expectations. Therefore, it is very important to understand how diverse data sharing factors, including institutional pressures, individual motivations, and technological resources, all affect scientists' data sharing behaviors. In this research, the diverse forms of data sharing behaviors are categorized into three different actions: (a) making data accessible through data repositories, (b) submitting data as journal supplements, and (c) providing data via personal communication methods upon request.

This research investigates how institutional, individual, and resource factors all map to scientists' data sharing behaviors. This research focuses on the scientists in STEM (Science, Technology, Engineering, and Mathematics) disciplines, as their data sharing and reuse become more important by institutional policies, technological infrastructure, and their scientists' awareness (Kim & Stanton, 2016; Kim & Zhang, 2015). This research assumes that the data sharing behaviors of scientists are not a matter of an individual scientist's arbitrary choice, but rather, that decisions as to whether to share data with researchers outside of their research group reflect the choices of communities of colleagues embedded within their disciplines. Therefore, it is necessary to investigate how the combinations of institutional, individual, and resource factors influence scientists' diverse forms of data sharing behaviors. This investigation provides a holistic view of the institutional, individual, and resource factors influencing scientists' diverse forms of data sharing in different institutional settings.

2. Literature review

Research data (data in general) refer to the extensive range of relevant information about research processes and results. Individual researchers or groups of researchers collect data using diverse collection methods including observations, experiments, and simulations. In this research, "data sharing" is defined as scientists' providing the research data behind their published article(s) to other researchers in "diverse forms," including (a) making data accessible through data repositories, (b) submitting data as journal supplements, or (c) providing data via personal communication methods upon request. Data are considered to be a fundamental infrastructural component of scientific research (Uhlir, 2010), especially because, when considering data-intensive research, data are not the end products of research, but need to be considered as part of an evolving stream in a scientific field (Faniel, Kriesberg, & Yakel, 2016). If we understand data as fundamental infrastructure and an evolving stream in scientific research, sharing and reusing data becomes critical to modern scientific research (Piwowar & Vision, 2013; Wallis et al., 2013).

Previous research in data sharing studied the benefits of data sharing in terms of validating prior studies, advancing scientific research, and educating science trainees. First, scientists validate previous studies by peer review of the original data (Borgman, 2012). By reanalyzing the original data, scientists can confirm or refute research findings, which helps prevent scientific error or misbehaviors such as fraud or selective reporting (Borgman, 2007; Vickers, 2006). Second, scientists can also test new hypotheses based on the shared data sets (Borgman, 2010; Fienberg, 1994; Vickers, 2006), and can conduct meta analyses (Vickers, 2006), which eventually lead to scientific innovation (Borgman, 2010; Campbell et al., 2002; Tenopir