Toward a Conceptual Framework for Data Sharing Practices in Social Sciences: A Profile Approach

Wei Jeng
School of Information Sciences
University of Pittsburgh
wej9@pitt.edu

Daqing He
School of Information Sciences
University of Pittsburgh
dah44@pitt.edu

Jung Sun Oh
School of Information Sciences
University of Pittsburgh
jsoh@pitt.edu

ABSTRACT

This paper investigates the landscape of data-sharing practices in social sciences via the data sharing profile approach. Guided by two pre-existing conceptual frameworks, Knowledge Infrastructure (KI) and the Theory of Remote Scientific Collaboration (TORSC), we design and test a profile tool that consists of four overarching dimensions for capturing social scientists' data practices, namely: 1) data characteristics, 2) perceived technical infrastructure, 3) perceived organizational context, and 4) individual characteristics.

To ensure that the instrument can be applied in real and practical terms, we conduct a case study by collecting responses from 93 early-career social scientists at two research universities in the Pittsburgh Area, U.S. The results suggest that there is no significant difference, in general, among scholars who prefer quantitative, mixed method, or qualitative research methods in terms of research activities and data-sharing practices. We also confirm that there is a gap between participants' attitudes about research openness and their actual sharing behaviors, highlighting the need to study the "barrier" in addition to the "incentive" of research data sharing.

Keywords

Research data sharing, knowledge infrastructure (KI), Theory of Remote Scientific Collaboration (TORSC), social science, qualitative data

INTRODUCTION

Sharing information, ideas, and research materials has always been recognized as one of the fundamental features of scholarly collaboration and scientific discovery (Franceschet & Costantini, 2010). Among all the sharable resources, research data is viewed as a valuable cornerstone because it allows scholars to make sense of inquiries, gain insight from evidence, develop humanity, and explain the world around us (Corti, Van den Eynden, Bishop, & Woollard, 2014). Given the recent mandates from institutions, publishers, and funding agencies, as well as the encouragement from professional associations for data management and sharing plans (ROARMAP, 2014), sharing research data has become a movement, an expectation, and also a common-sense practice.

However, previous studies have revealed that some STEM (Science, Technology, Engineering and Math) researchers are reluctant to share data for several reasons: unbalanced cost-effectiveness (too much effort but few perceived returns), perceived risks (such as fear of data misinterpretation and misuse), and lack of incentives (Tenopir et al., 2011; Kim & Stanton, 2016).

The same barriers encountered by STEM researchers also plague social science researchers. Worse yet, the latter usually face additional challenges due to the high ethical standard expected by the social science community (Israel & Hay, 2006), the lack of funding and technical infrastructure in general (Jeng & Lyon, 2016), and the higher probability that they will handle qualitative data, which are often considered too complex to reuse and share (Yoon, 2014). Given the presence of these additional obstacles and the unique characteristics of social science data, studies are needed that specifically focus on social-science researchers in order to understand their specific data-sharing practices.

Traditionally, professional communities in the data curation and data management fields rely on *profiling tools* to gather descriptions about researchers and their research data in a "concise but structured document" (Witt, Carlson, Brandt, & Cragin, 2009, p.3). The researchers or practitioners who use such a profiling tool can later illustrate a landscape or current state based on the collected responses. We find this profiling approach useful in studying data-sharing practices, as it assists a range of stakeholders (e.g., institutions, discipline communities, and data infrastructures such as repositories or data centers) to better understand individual researchers' current preparedness to share data and their actual data-sharing behaviors.

However, existing profiling tools are limited in many ways from understanding the social science data-sharing landscape. First, these tools are not designed for data sharing. Most of them focus on data curation (e.g., Data Curation Profile), digital preservation (e.g., Cornell Maturity Model), data management (e.g., CMM for SDM), and data infrastructure (e.g., CCMF). Second, because

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

existing tools are made for big science or data-intensive research (e.g., CCMF), they are not fully suitable for social sciences or humanities without substantial modifications (Jeng & Lyon, 2016). Finally, these tools do not scale well to collect larger sample sizes, as it takes a long time to complete the questions.

To fill the need for a customized profiling tool to investigate social scientists' data-sharing practices, we develop a comprehensive profiling instrument encompassing all facets of social-science research, including mixed-method research and qualitative data that have been thus far under-investigated. We stress the importance of grounding the instrument development in theoretical frameworks, and adopt pre-existing conceptual frameworks related to digital scholarship. To validate the effectiveness of this profiling instrument, we apply it into a case study. By doing this, we want to discover whether the breadth and depth of the profiling instrument can sufficiently cover individuals, data, technology, and their discipline culture.

This study aims to address the following research questions:

- How can a data-sharing profiling tool be developed based on existing conceptual frameworks that support digital scholarship, particularly Knowledge Infrastructures and Scientific Collaboration Theory?
- What does this profiling tool reveal about social scientists' data-sharing practices, including their perceived technological infrastructure, research culture, and motivations in terms of data sharing? Particularly, is there any difference among social scientists who prefer quantitative, mixed-method, or qualitative methods?

Two conceptual frameworks -- Knowledge Infrastructures (KI) and Theory of Remote Scientific Collaboration (TORSC) -- are used as a theoretical lens, leading to the development of four overarching dimensions: data characteristics, perceived technical infrastructure, perceived organizational context, and individual characteristics and motivation.

Under the guidance of KI and TORSC, we further examine several well-known profiling tools, including the Community Capability Model Framework, Data Curation Profile, and survey instruments presented in Tenopir et al. (2010), Wallis, Rolando, & Borgman (2013), and Kim & Stanton (2016). The goal is to construct set questions related to data sharing in social sciences that are understandable by social scientists and require reasonably minimal effort to answer.

In the remaining sections, we introduce two conceptual frameworks, followed by reviewing two highly-relevant current profiling tools and related work. In the Methodology section, we discuss how we constructed the detailed questions and conducted a case study using our profile. In the Result and Discussion sections, we report findings from the case study and summarize our research insights.

LITERATURE REVIEW

Conceptual Frameworks Supporting Digital Scholarship

Both Knowledge Infrastructures (KI) and Olson's Theory of Remote Scientific Collaboration (TORSC) are well-known theories for supporting digital scholarship.

Knowledge Infrastructures (KI). The term "knowledge infrastructure" builds on early developments in e-Research movements and information infrastructure (Borgman, 2015). Transformed from information infrastructure (Bowker, Baker, Millerand, & Ribes, 2010), knowledge infrastructures refer to "robust networks of people artifacts and institution that generate, share, and maintain knowledge about human and natural worlds" (Edwards, 2010, p. 17, as cited in Borgman, 2015). Knowledge infrastructures include seven elements – people (individuals), shared norms and values, artifacts, institutions (organizations), routines and practices, policies, and built technologies – all of which work together as a complex ecology (Edwards et al., 2013; Borgman, Darch, Sands, Wallis, & Traweek, 2014).

Theory of Remote Scientific Collaboration (TORSC). Data sharing can be viewed as a type of scholarly collaboration. G. Olson and J. Olson (2000) discuss four concepts that lead to success in remote scientific collaboration: 1) common ground, 2) coupling work, 3) collaborative readiness, and 4) technology readiness. These four concepts have been adopted in the fields of information science and behavioral science when researchers want to discuss the essence of scholarly collaboration and communication (Borgman, 2007). In 2008, Olson and his research team developed TORSC (Theory of Remote Scientific Collaboration), which extends their previous framework to include general collaboratories. The updated framework comprises five overarching categories: the nature of the work, common ground, collaboration readiness, management/planning/decision-making, and technology readiness (Olson, Zimmerman, & Bos, 2008, p.80; J. Olson & G. Olson, 2013). TORSC complements the theoretical foundation of KI by considering more elements of scientific collaboration.

Inspired by the above-mentioned frameworks, we propose a novel framework designed to investigate scholars' data-sharing practices. As shown in Table 1 on the next page, our proposed framework consists of four dimensions: characteristics, perceived technical infrastructure, perceived organizational context, and individual characteristics & motivations. These act as the highest level in our profile.

Profiling Tools for Data Curation and Management

For the questions and measurement items under each dimension, we reviewed several current data-practice profiling tools, two of which are the Community Capability Model Framework (CCMF) and Data Curation Profile (DCP).

Framework to support digital scholarship		Dimensions influencing data-sharing practices (proposed by this study)	
Knowledge Infrastructure (KI)	Theory of Remote Scientific Collaboration (TORSC)		
▪ People (individuals) ▪ Shared norms and value	▪ Collaboration readiness	Individual facet	Individual motivations and characteristics
▪ Artifacts	▪ The nature of the work	Context facet	Data characteristics
▪ Institutions (organizations) ▪ Routines and practices ▪ Policies	▪ Common ground ▪ Management, planning, and decision making		Organizational and research culture
▪ Built technologies (system and networks)	▪ Technology readiness		Technical infrastructure

Table 1. Proposed framework to study data-sharing practices.

Community Capability Model Framework (CCMF). This framework aims to examine the infrastructure of an academic discipline's data curation, management, and sharing practices (Lyon, Ball, Duke, & Day, 2012). The CCMF Toolkit was released as an instrument, in a spreadsheet style, that includes a consent form, 10 open-ended questions about an interviewee's data profiles, and 55 other questions related to critical factors of data capabilities. In terms of the applications of this toolkit, both Brandt (as cited in Lyon, Patel, & Takeda, 2014) and Jeng and Lyon (2016) apply CCMF to study agronomy and social-science scholars' data practices, respectively.

Data Curation Profiles (DCP). DCP supports practitioners and researchers who would like to assess and analyze researchers' data, and considers the discipline's characteristics (Cragin et al., 2010; Witt et al., 2009). One apparent usage for each completed data curation profile is as a resource to help practitioners and researchers quickly capture how specific data will be generated, reused, and used in a certain research area. Lage, Losoff, and Maness (2011) adopted the DCP tool to examine research data practices in the University Libraries at the University of Colorado-Boulder. Their findings, presented as eight persona profiles, help academic librarians and data librarians understand clients' data needs, barriers, and data-related activities.

Because CCMF focuses more on technological and organizational infrastructure, we adopt CCMF's actual questions to strengthen the "Technology Infrastructure" and "Organizational and Research Culture" dimensions in Table 1. The components in DCP are primarily used for collecting "Data Characteristics".

However, while the actual questions in CCMF and DCP provide a good starting point to facilitate our profile design, they both lack considerations about individual motivations. Thus, we adopt other related work in the topic of research data sharing to fill this gap.

Research Data Sharing

The related literature on research data sharing can be examined on two levels with different granularities: general (including social scientists and STEM scientists) and social science specifically.

The report by the Research Information Network (RIN, 2008) is likely the most comprehensive report investigating researchers' data sharing in the past decade (Witt et al., 2009). The report examines six subject areas and two interdisciplinary areas (mainly in STEM fields), and interviews 10-15 scholars in each area. The RIN project identifies researchers' data needs, motivations, constraints, and attitudes in ensuring data qualities. It also points out several gaps, such as the lack of a reward model and researchers' skillsets for preparing data sharing.

Tenopir et al. (2011) investigates 1,329 scientists' data needs, sharing practices and intentions. They find that social-science researchers are less likely to make their data electronically available to others when compared with STEM scholars. Overwhelmingly, 79.4% of the social-scientist participants agreed or somewhat agreed that they had concerns about data being used in ways other than intended.

Kim's research team conducted a national survey with more than 1,000 researchers in 43 disciplines in 2013 (Kim, 2013; Kim & Stanton, 2016). Their research indicates that perceived career advancement and individual researchers' altruism have positive associations with their data-sharing frequencies. On the other hand, perceived effort might hinder their sharing frequencies. Kim and Adler (2015) extracted the sample of social scientists from Kim's earlier work (2013) and specifically discuss social scientists' data-sharing behaviors. They hypothesize that the pressure from funding agencies and journal publishers would influence social scientists' data sharing. However, they found no statistical evidence supporting this hypothesis specifically.

Fecher, Friesike, and Hebing (2015) conducted a thorough literature content analysis with 98 selected articles, and finally built a theoretical model (i.e., Figure 4 in Fecher, Friesike, & Hebing, 2015) to explain the process of sharing data. They also provide a complete view of a data-sharing workflow, which has inspired follow-up studies to investigate the relationships between components in the workflow.

Dimensions	Attributes	Examples questions	# of items	Source
Data Characteristics	DC1. User of data	Target audience of data	9	Witt et al., 2009 (DCP);
	DC2. Data source	Observational data, survey data, experimental data, simulation data (generated from test models)	7	University of Virginia Libraries
	DC3. Data types	Text, relationship, images, or audio		
	DC4. Data volume	File size, number of files in a study	3	Lyon et al., 2012 (CCMF)
	DC5. Data sensitivity	Data that are sensitive or confidential		
	DC6. Data's shareability	Data that are sharable	1	Proposaed by this study
	DC7. Data ownership	Ambiguity of data ownership	1	Parry & Mauthner, 2004
Technical Infrastructure	TI1. Platform availability	Existing disciplinary data repositories	3	Fecher et al. 2015; Mennes et al., 2012
	TI2. Platform usability*	Easy-to-use platform, tools and application' usability	0	Fecher et al. 2015; Mennes et al., 2012
	TI3. Facilities	Access to technical tools or resources	6	Coti et al., 2013
	TI4. Technical standards*	Metadata standard	0	Lyon et al., 2012 (CCMF)
Organizational and Research Context	OC1. Funding sufficiency	Funding for the support of data sharing	1	Lyon et al., 2012 (CCMF)
	OC2. Research data service (RDS) supports	Existing library RDS support	3	Proposaed by this study
	OC3. Internal human resources	Human resources involved in RDM services	7	Lyon et al., 2012 (CCMF)
	OC4. Legal and policy	Mandates	1	Lyon et al., 2012 (CCMF)
	RC1. Discipline culture	The culture of open sharing	6	Proposaed by this study
	RC2. Discipline norms	Discipline norms and ethical considerations in terms of subject protection	2	Israel & Hey (2006); Israel (2015)
	RC3. Research skills	Valued research skills	9	Proposaed by this study
	RC4. Research activities	Research activities involoved	11	Mattern et al., 2015
Individual Characteristics and Motivations	IC1. Researchers' demographics	Prior experience, positions, etc.	8	--
	IC2. Cost effectiveness	Sufficient time for preparing datasets, documentation, ensuring the interoperability; administrative work, potential misuse or misinterpretation of the data	5	Kim & Stanton, 2016; Wallis et al, 2013; Tenipir et al., 2011; 2015
	IM1. Extrinsic motivation	Expected reward for career, citations	3	
	IM2. Scholarly Altruistism	Altruistic behaviors (e.g., sense of achievement for sharing great research)	2	
Research Product Sharing Practices	DS1. Data sharing (channels and frequencies)	• Publishing with journal venues • Institutional repositories • Publically accessible web sites • Academic social media platforms • Discipline repositories • Sent to others upon request	6	Kim & Stanton, 2016; Tenipir et al., 2011; 2015
	DS2. Manuscript sharing (channels and frequencies)	• Institutional repositories • Publically accessible web sites • Academic social media platforms • Discipline repositories • Sent to others upon request	5	Proposaed by this study, questions were based on DS1

Table 2. Proposed profiling instrument for capturing data sharing practices in social sciences (99 items)

Note: *Items are dropped when the case study is carried out.

In summary, existing studies only include social scientists as a small portion of their participants (e.g., 15.3% in Tenopir, and 14.6% in Kim & Stanton), and the scope of their studies is broader, addressing social sciences only marginally.

METHODOLOGY

Constructing the Profile Instrument

The profile instrument in this study consists of four dimensions at the top level, and then many actual measurements (see Table 2) to examine data characteristics, technical infrastructure, perceived organizational context, and individual characteristics and motivations to survey social scientists' actual data-sharing behaviors. Besides the four dimensions adopted from KI and TORSC, we append a sub-section describing a group of questions related to social scientists' actual data-sharing behaviors.

Data Characteristics

We believe that the nature of the research data can influence the intention or decision to share. Therefore, our instrument includes questions regarding data characteristics (e.g., source and volume), as well as approaches and strategies to manage, archive, and reuse data. Furthermore, social science data can be produced from observations, experiments, and simulations (e.g., from test models). The distinctive source of data might also raise issues of confidentiality or ambiguity of data ownership (Parry &

Virginia RDS	Modified items in this study
Observational	Observational data captured in real time (e.g., fieldnotes, social experiments)
N/A	Data directly obtained from the study groups/informants (e.g., survey responses, diaries, interviews, oral histories)
Experimental	Experimental data (e.g., log data)
Simulation	Simulation data generated from test models, where models are more important than output data (e.g., economic models)
Derived or compiled	
N/A	Documentation-based data: records, literature, archives, or other documents (e.g., court records, prison records, letters, published articles, historical archives)
N/A	Secondary data (e.g., government statistics, data from IGOs or NGOs, other's data)
N/A	Physical materials (e.g., artifacts, samples)

Table 3. An example of customized items: data types in social science

Mauthner, 2004). These factors may hinder data sharing in social sciences. In the end, we developed seven questions for this dimension (see DC1- DC7 in Table 2).

For questions regarding social scientists' data type (source), our tool adopts the University of Virginia Library Research Data Services' version (n.d.) but carefully tailors it to fit the context of social science research activities. For example, in Table 3, we added four new categories for data type in order to enhance the measurement: data directly obtained from the participants, documentation-based data, secondary data, and physical materials.

In addition to data source, we also capture data volume. Social-science data are inherently complex and can be "big" (Dey, 1993). The volume and complexity of data (especially those involving a variety of sources) might discourage scholars from sharing data (Jahnke, Asher, & Keralis, 2012). On the other hand, some data might contain sensitive or copyrighted information, which has disclosure risks and cannot be shared without proper handling.

Technological Infrastructure

From a technical point of view, there are three limitations that impede the intention to share data in the social sciences: TI1- *platform availability*, TI2- *platform usability*, TI3- *facilities*, and TI4- *technical standards*.

Platform availability examines whether there is a common, easy-to-locate platform on which scholars can publish data. However, even if such a platform exists, its service might not always be easy to adopt and use (Fecher et al. 2015). Therefore, related work emphasizes the importance of an easy-to-use data-sharing platform. Such a platform should contain several well-designed features, such as a simple upload mechanism or automatic data verification (Poline et al., 2012; Mennes, Biswal, Castellanos, & Milham, 2013).

Platform usability enables us to examine whether existing platforms are difficult to access or use due to inadequate support, e.g., lack of access to a data analysis tool or lack of research data management resources. Researchers encounter resistance or fail to obtain support within their associated institutions. Due to insufficient technical support or associated resources, some institutes lack technical training programs or administrative support for researchers.

The lack of well-defined technical standards could be a factor that discourages sharing and reuse. Prior work has suggested that in order to achieve long-term accessibility and usability of research data, it is necessary to adopt sustainable digital file formats, standard metadata, and comparable software (Corti et al., 2014). In addition, for each dataset shared via non-standard formats or procedures, researchers interested in reuse have to investigate additional resources for interpretation. In other words, researchers can benefit from well-defined standards that specify suggested or mandatory file formats, discipline-dependent metadata for datasets, sufficient minimal data description, etc.

Organizational and Research Culture

Table 2 lists the items related to organizational and research culture (i.e., OC1-OC4, RC1-RC4) that can influence social scientists' data-sharing practices. Based on the literature regarding research norms in social sciences, we argue that community plays an important role, influencing an individual's data-sharing decision and motivation. Organizational and research context can be discussed in two ways: as an institution in which scholars are employed or affiliated, or as the research norms from the discipline's community practices.

Certain internal research cultural factors, such as unfamiliarity with appropriate methods of secondary analysis and lack of a sharing culture (Jeng & Lyon, 2016; Kim & Stanton, 2016), are also incompatible with sharing. Institutional supports for data management or data curation has a critical impact on scholars' behaviors.

From a research norm perspective, social-science researchers have expressed several concerns about sharing their data, especially when qualitative data are involved. For example, some are hesitant to share their data due to ethical considerations (RC2- *Discipline norms and ethical considerations*), such as worrying about misconduct or misuse (Kim & Stanton, 2016) and the level of required privacy protection (Yoon, 2014; Jahnke et al., 2012). Researchers are unsure whether they have the right to publish the data or to what extent it should be sanitized to protect participants' privacy.

In addition to disciplinary norms, we would like to capture valued research skills (RC3) and research activities (RC4), inspired by Mattern et al. (2015)'s study. Mattern et al. gathered information about how social scientists visualized their research patterns, and found that social scientists do not follow the similar research process. RC3- *Research Skill* and RC4- *Research activities* aim to deepen this observation and to further examine whether social scientists' research activities are associated with their data-sharing practices.

Individual Characteristics and Motivations

Individual factors such as academic position and other characteristics always play a critical role in scholars' data-sharing decisions (IC1- *Researchers' demographics*). IC2- *Cost effectiveness* is another layer of consideration for selective factors that influence researchers' data-sharing behaviors. Given low expected benefits or high expected effort, researchers lack incentives to share or reuse data (Kim, 2013; Kim & Stanton, 2016). Prior work identifies the challenge researchers face to provide "rich-enough" documentation of context or insufficient time for others to use unfamiliar data (Corti et al., 2014). Tenopir et al. (2011) also indicate that "[t]he leading reason (of why their data are not available electronically) is insufficient time" (p. 9).

A lack of reward models can be viewed as a barrier for data sharing. Scholars greatly rely on a reward system in which recognitions, research funds, and credits can return to those who make contributions to creating knowledge (Kim, 2013). However, the current reward model in the social science field is still associated with publications in formal venues (e.g., journals which received higher SSCI impact factors). Data-sharing reward models (IM1- *Extrinsic motivation* in Table 2) within social-science disciplines are still not widely recognized. Based on prior studies (e.g., Kim & Stanton, 2016), we also include IM2- *Scholarly altruism*, for these two factors (IM1 and IM2) might strongly influence social scientists' data-sharing behaviors.

Data Sharing Practices

We adopt the measurement that Kim's team used (2013; Kim & Stanton, 2016) as an outcome of social scientists' data-sharing practices. Kim's measurement covers online channels that researchers can use to give others access to their research data, as well as the frequencies in which they have done so. In addition to data-sharing frequencies, we are also curious about social scientists' manuscript (pre-print) sharing conditions as a reference point. The question examples are listed in DS1- *Data sharing* and DS2- *Manuscript sharing* in Table 2.

The final version of our profile includes 99 items (four open-ended questions, seven items in multiple selections, and 88 items in multiple choice format). Among the 88 multiple-choice questions, 54 use a 5-point Likert scale which allows for future factor analysis. Note that TI4- *Technical standards* and TI2- *Usability* were removed from the case study because at that point we were unsure whether our participants share their data to a discipline repository or an institutional repository; it was therefore too early to gather detailed information about how they assess metadata standards and the usability of these repositories.

Case Study on Social Scientists' Data Sharing

As stated, we conducted a case study using a profile instrument to examine social scientists' data sharing. This case study used a convenience and representative sampling method for data collection, recruiting early-career researchers who were available to participate. Our rationale for targeting early-career researchers is that they tend to be fully engaged in every research stage of their projects, including data collection, processing, and analysis, whereas senior researchers might focus more on constructing ideas and interpreting data. The target population includes all currently-enrolled PhD students and post-doctoral researchers in all social-science-related department units at two major research universities, the University of Pittsburgh (PITT) and Carnegie Mellon University (CMU) in the U.S. Survey invitations were sent to 553 potential participants in 20 social-science-related units at these two universities. Among the invitation emails sent to PITT participants (498 out of 553), 17 were immediately rejected by the email service system, possibly due to account expiration after users left the organization.

With an online questionnaire link (Qualtrics), an invitation for completing the profile was sent in December 2015, and a reminder was sent in February 2016. We received responses from 93 out of the 536 successfully-delivered invitations, resulting in a 17.4% response rate. This rate is highly comparable to that of related work (with response rates of 9-16%) (Kim & Stanton, 2016; Tenopir et al., 2010). Among the 93 responses, 66 completed the full profile. These 66 completed profiles were the final samples included in this study. After removing two extreme values

		Self-identified preferred research methods			TOTAL
		QUANT	MIX	QUAL	
Discipline Groups	Eco & Business	12	1	0	13
	Info & Communication	1	5	2	8
	Policy & Political Sciences	7	6	0	13
	Psychology & Decision sciences	12	2	0	14
	Education	7	4	0	11
	Sociology & social work	1	0	4	5
	History	0	2	0	2
	Total	40 (60.6%)	20 (30.3%)	6 (9.1%)	66

Table 4. A cross-tabulation of preferred research methods and disciplines (n=66)

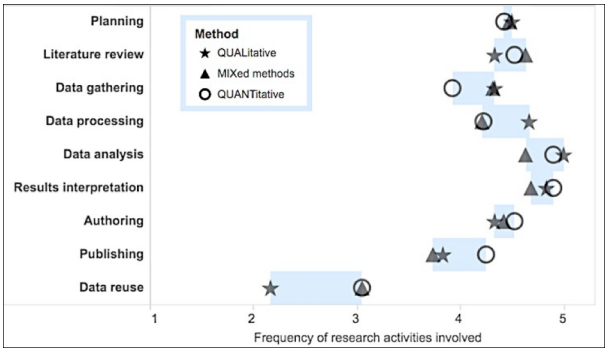

Figure 1. Frequency of research activities involved in social scientists' general research projects

(23.4 hours and 8.82 hours), the average completion time for the remaining 64 participants is 13.4 minutes.

RESULT FINDINGS

Research Activities

Table 4 summarizes the distribution of our sample participants by preferred research methods and discipline groups. Both *Policy & Political Science* and *Education* have a non-negligible portion favoring QUANT and MIX approaches. Participants in *Economics & Business* overwhelmingly select QUANT approaches as their preferred method. *Information & communication* participants identify MIX approaches as the ones they mostly take.

For participants in each method group (i.e., QUAL, MIX, and QUANT), we analyzed how frequently they perform individual research activities. These research activities include Planning, Literature Review, Data Gathering, Data Processing, Data Analysis, Result Interpretation, Authoring, Publishing, and Data Reuse (Mattern et al., 2015).

Figure 1 summarizes the results of the research activities involved in participants' general research work, where legends ★, ▲, and ○ represent the qualitative, mixed, and quantitative groups, respectively. Participants are asked to what extent certain research activities might be involved in their research. The frequency is measured on a scale from 1 (never) to 5 (all of the time). The light blue band indicates the range (difference) among observed values.

The results provide several interesting findings. First, counterintuitively, there is no significant difference between qualitative and quantitative methods, even for data-related activities such as data processing and analysis. There is a significant difference between the frequencies of data analysis on different research methods at the $p<.05$ level conditions [(2, 62) =4.32, p=0.018]. Post hoc comparisons using the Tukey HSD test suggest that the mixed method approach (M = 4.63, SD = 0.114) is significantly lower than the other two.

Second, the MIX group does not always fall in between QUAL and QUANT—an interesting pattern that we would like to investigate in future work.

We also observe different averages in the "publishing" and "data reuse" stages. A subsequent ANOVA test suggests that researchers whose primary method is quantitative data report more frequent publishing activities than the other two methods.

Research Data Characteristics

For social scientists' research data, we report results from four research data characteristics: data volume, data type, whether the data can be shared, and the intended audience of the data.

Among the 61 participants who completed the responses and reported data volume, two-thirds deal with data on the scale of megabytes (N=44), thereby confirming that they are small-data rather than big-data projects. Specifically, 26 participants report volumes between 0-100MB, 18 report 100MB-1GB, 15 report 1GB-10GB, and five report to have more than 5GB. The average data volume is 4.25 GB per research project, with a median of 200 MB, indicating the existence of outlier values much higher than the average. Although the majority (61 out of 66) report an estimated size, there are still five participants who answered "unknown." In the Discussion section, we share insights for how we can modify this question to further improve the response rate.

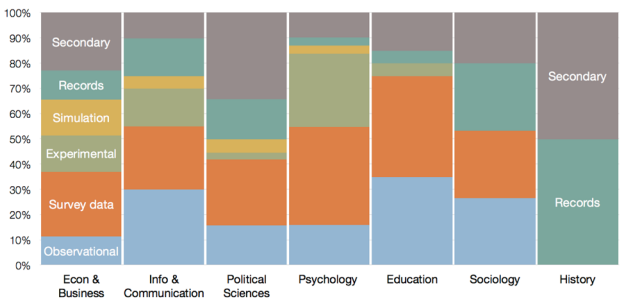

Figure 2. Data types and discipline categories

The average data volume of QUANT projects is 5.4GB, much larger than that of QUAL (2.6GB) or MIX (2GB). However, through an ANOVA, we did not find evidence to support the hypothesis that there is a significant difference of data volume among these three research methods.

Figure 2 illustrates the distribution of data types in each discipline. Although Economics is biased toward QUANT in terms of a primary research method (see Table 4), its data type is diversified. The data type reported by Education, Sociology, and History researchers are less diverse and centered around qualitative data, such as records and observational data.

We further investigated whether research methods are associated with shareability of the research data. When asked if their data is sharable, the majority of participants report that their data is completely shareable (N=14, 21.2%) or mostly shareable (N=28, 42.4%). However, about 5% of participants think their data is not allowed to be shared. Table 5 summarizes the answers reported by participants in the different method groups. Although the QUAL group appears to skew toward "not shareable" compared with the QUANT and MIX groups, the difference is not statistically significant in our chi-square test, where χ^2 (4, N = 61) = 8.92, p=0.06, at the 0.05 level. Note that because the chi-square test requires the expected value in each cell to be greater than 5, our analysis only includes data for Completely sharable, Mostly sharable, and Partially sharable.

	Preferred methods			Total
	Quant (n=40)	Mix (n=20)	Qual (n=6)	
Completely Sharable	10	4	0	14
Partially Sharable	17	10	1	28
Partially Sharable	9	5	5	19
Not allowed to share	2	1	0	3
Other	2	0	0	2

Table 5. A cross-tabulation of data sharability and research methods

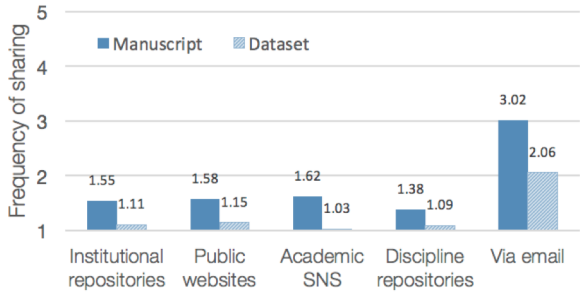

Figure 3. The frequency of sharing research products on five sharing channels

As for the target audience for the data, "researchers in the same discipline" wins by a landslide, mentioned by 93.9% (62 out of 66) of the participants. In second place, surprisingly, is "graduate students" (40 out of 66, 60%), suggesting that participants perceive the value of teaching and learning from research data. The third and fourth are the practitioner (25 of 66, 37.9%) and policy maker (25.8%), respectively. Besides these top four choices, government administration, research participants, and researchers outside the field are also mentioned by more than 20% of participants. Note that the participants are allowed to select more than one target audience, and thus the total exceeds 100%.

Current Practices of Data Reuse and Sharing

Figure 3 reports the frequency of sharing data in the past three years on five channels, including Institutional Repositories, Public Websites, Academic SNS, Discipline Repositories, and Via Emails. The frequency is scaled between 1 (never) and 5 (all of the time). In an attempt to establish a meaningful baseline, we also asked about the frequency of sharing manuscripts (including pre-prints) in addition to sharing datasets, because manuscripts can be seen as the most common product generated by research.

Unsurprisingly, the frequency of manuscript sharing is slightly higher than that of dataset sharing. However, the sharing frequency remains consistently low for the five channels and the two types of research products. Before manuscript sharing becomes a common practice, it might be difficult for researchers to take an additional step toward dataset sharing. To validate this hypothesis, in the future we would like to study the relationship between the frequency of data sharing and preprint sharing.

Community culture	1	2	3	4	5
Researchers in my discipline expect people to share data.	11.5%	27.9%	18.0%	32.8%	9.8%
It's common to see people share their data in my discipline community.	11.3%	27.4%	17.7%	32.3%	9.7%
There is a standard procedure for data sharing.	36.8%	33.3%	15.8%	8.8%	5.3%
There are well-known data repositories everyone knows.	24.6%	21.3%	26.2%	16.4%	11.5%
Discipline cares a great deal about the protection of human participants	6.0%	1.5%	16.4%	26.9%	49.3%

Table 6. Perceived community culture

Perceived benefits	1	2	3	4	5
More citations	1.5%	10.6%	**48.5%**	27.3%	12.1%
Career advancement	3.0%	13.6%	**40.9%**	33.3%	9.1%
Collaboration opportunity	1.5%	3.0%	7.6%	**62.1%**	25.8%
Fulfill others' research need	0%	3.0%	**33.3%**	30.3%	**33.3%**
Inspire researchers outside your field	0%	1.5%	19.7%	**45.5%**	33.3%

Table 7. Perceived benefits

Perceived Organizational, Research Culture, and Technical Infrastructure

Table 6 shows a list of possible community cultures and to what extend the participants agree that they are indeed community cultures, where 1 represents strongly disagree and 5 represents strongly agree. To our surprise, there is a cognitive gap between the perceived culture and the reality. While participants are inclined to agree that the community expects people to share data, and while they agree that it is common to see people sharing data, Figure 3 tells a different story.

The majority of participants (strongly or slightly) disagreed with the existence of a standard procedure and well-known, recognized data infrastructure. The result is consistent with Jeng and Lyon's (2016) findings that standards are one of the least-developed capabilities in social science disciplines.

As for the perceived technology infrastructure and supports in participants' work environment, only a small portion of participants report that tools or resources for facilitating data reuse (13%) and data sharing (5.8%) are sufficient, suggesting that the related services have room for improvement to prepare social scientists to reuse and share data.

Individual Motivations

The participants were also asked about perceived benefits and rewards of sharing data, as reported in Table 7 (1: strongly disagree; 5: strongly agree). More than 85% of

participants (strongly or slightly) agreed that opportunity for collaboration is one benefit of data sharing. However, it is interesting that a large percentage of participants (more than 40%) took a neutral stance regarding citations and career advancement. It is worth noting that two of the perceived benefits (i.e., *Fulfill others' research need* and *Inspire researchers outside your field*) are altruistic. If we consider only the "strongly agree" column, these two altruistic reasons outperform the rest, and they are each backed by 33.3% of participants.

DISCUSSION

Through the findings of the case study, we gain several insights regarding the development process of a data-sharing profile and the status of data sharing in social sciences.

Insights on Developing a Profiling Tool for Data Sharing

First, we find that institutional, departmental, and discipline levels are often interwoven; thus, it is hard to precisely categorize questions in TI, OC, and RC (Table 2). For a particular infrastructure, such as funding resources or technical resources, researchers can either obtain them from external funders (e.g., a discipline community) or from the local institution. We leave the problem of precise categorization to future work.

Another observation is that some questions in our profile are context-specific. For example, data volume (the total size of data), data sensibility, and data sharebility can vary significantly depending on the projects themselves. Another example is the research stage of a project. In a real-world situation, a researcher might work on multiple research projects in parallel: some projects might be closed, whereas others might still be in early stages and not ready for any form of sharing. Since the situations can differ from project to project, it is imperative to ask the participant to focus on one project that has been completed when reporting a cross-sectional study. Specifically, we think that the participants ought to recall one of their completed and most representative projects when they answer the questions. In practice, this can be achieved by applying a survey software system's piping functions, such as the Piped Text function in Qualtrics or Piping function in Survey Monkey.

Data Sharing Practices in Social Science

We confirm that social scientists rarely share data, which is largely consistent with prior work. However, as our baseline, manuscript sharing in social sciences is not much more active than data sharing. It is also intriguing that no statistical difference was found between qualitative, mixed, and quantitative methods with respect to data-sharing behaviors. We plan to collect a larger sample for further investigation.

We also find that scholarly altruism is a common reason for data sharing, whereas extrinsic motivations (e.g., gaining citations and career advancement) are less relevant.

Most importantly, we reveal a chasm between social scientists' attitudes, beliefs, and actual behaviors. This observation is consistent with prior work by Jeng and Lyon (2016): social science scholars highly value data sharing and witness data sharing in their fields, but they do not actually share their own data.

The lack of extrinsic motivations and the gap between attitudes/beliefs and actual behaviors have been observed repeatedly in the literature. Thus, we believe there is a critical need to study not only motivations and incentives, but also the "barrier" in the way of social scientists' data sharing.

CONCLUSION

This study presents a profile instrument that captures individual social scientists' research activities, data-sharing practices, data characteristics, and perceived technical support. In the case study, we find that there is no significant difference among quantitative, mixed, and qualitative methods than we predicted in terms of research activities and data-sharing practices for early-career social scientists. We also confirm that there is a gap among participants' attitudes and actual behaviors. However, we are unable to draw disciplinary conclusions from the current case study.

Future work includes two threads. First, we plan to conduct the survey on a larger scale. To achieve this, we plan to convert this profile into a questionnaire that is suitable on a national level and does not need to be supervised. Second, we would like to extend this profile to behavioral science or humanities, and even to the social aspects in STEM (e.g., nursing and public health), to test the generalizability of our profiling instrument.

ACKNOWLEDGMENTS

The authors thank the iFellowship, guided by Committee on Coherence at Scale (CoC) for Higher Education, sponsored by Council on Library and Information Resources (CLIR) and Andrew W. Mellon Foundations, provides the research funding for this study. The authors also thank Drs. Nora Mattern, Liz Lyon, Sheila Corrall, Jian Qin, and Stephen Griffin for their invaluable comments and suggestions on this research project.

REFERENCES

Borgman, C. L., Darch, P. T., Sands, A. E., Wallis, J. C., & Traweek, S. (2014). The ups and downs of knowledge infrastructures in science: Implications for data management. *Proceedings from JCDL/TPDL 2014*.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge MA: MIT Press.

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup, M. Allen, (Eds.) *International Handbook of Internet Research* (pp. 97–117). Dordrecht, Netherlands: Springer Netherlands.

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: A guide to good practice*. London: Sage.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, 368(1926), 4023–4038.

Dey, I. (1993). *Qualitative data analysis: A user friendly guide for social scientists*. London: Routledge.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., ... & Calvert, S. (2013). Knowledge infrastructures: Intellectual frameworks and research challenges. Retrieved from http://knowledgeinfrastructures.org/

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing?. *PLOS ONE*, 10(2), e0118053.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, 4(4), 540–553.

Israel, M., & Hay, I. (2006). *Research ethics for social scientist*s. London: Sage.

Israel, M. (2015). *Research ethics and integrity for social scientists: Beyond regulatory compliance*. London: Sage.

Jahnke, L., Asher, A., & Keralis, S. D. C. (2012). The problem of data. Washington, DC: Council on Library and Information Resources (CLIR).

Jeng, W. & Lyon, L. (2016). A report of data-intensive capability, institutional support, and data management practices in social sciences. Proceedings from the *11th International Digital Curation Conference* (IDCC).

Kim, Y., & Adler, M. (2015). Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International Journal of Information Management*, 35(4), 408-418.

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*. 67(4), 776-799.

Lage, K., Losoff, B., & Maness, J. (2011). Receptivity to library involvement in scientific data curation: A case study at the University of Colorado Boulder. *portal: Libraries and the Academy*, 11(4), 915-937.

Lyon, L., Ball, A., Duke, M., & Day, M. (2012). Community Capability Model Framework. Retrieved from http://communitymodel.sharepoint.com/

Lyon, L., Patel, M., & Takeda, K. (2014). Assessing requirements for research data management support in academic libraries: introducing a new multi-faceted capability tool. Proceedings from Libraries in the Digital Age (LIDA), 13.

Mattern, E, Jeng, W., He, D., Lyon, L., & Brenner, A. (2015). Using participatory design and visual narrative inquiry to investigate researchers' data challenges and recommendations for library research data services. *Program: electronic library and information systems*. 49(4): 408-423.

Mennes, M., Biswal, B. B., Castellanos, F. X., & Milham, M. P. (2013). Making data sharing work: The FCP/INDI experience. *Neuroimage*, 82, 683-691.

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-computer interaction*, 15(2), 139-17.

Olson, G. M., Zimmerman, A., & Bos, N. (2008). *Scientific collaboration on the Internet*. Cambridge, MA: The MIT Press.

Olson, J. S., & Olson, G. M. (2013). Working together apart: Collaboration over the internet. Synthesis Lectures on Human-Centered Informatics, 6(5). San Rafael, CA: Morgan & Claypool.

Parry, O. & Mauthner, N. S. (2004) Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. *Sociology*, 38(1), 139-152.

Poline, J. B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., ... & Kennedy, D. N. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, 6.

Research Information Network (RIN) (2008). To share or not to share: Publication and quality assurance of research data outputs: Main report. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

ROARMAP: Registry of Open Access Repositories Mandatory Archiving Policies. (n.d.). Retrieved from http://roarmap.eprints.org/

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE*, 6(6), e21101.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE*, 10(8), e0134826.

University of Virginia Library Research Data Services. (n.d.). Retrieved April 2, 2015, from http://data.library.virginia.edu/data-management/plan/format-types/

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLOS One*, 8(7), e67332.

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, 4(3), 93-103.

Yoon, A. (2014). "Making a square fit into a circle": Researchers' experiences reusing qualitative data. Proceedings of the American Society for Information Science and Technology, 51(1), 1-4.