Social Science Data Repositories in Data Deluge: A Case Study at ICPSR Workflow and Practices

Abstract:

Design/methodology/approach: We conducted two focus group sessions and one individual interview with eight employees at the world's largest social science data repository, the Interuniversity Consortium for Political and Social Research (ICPSR). By examining their current actions (activities regarding their work responsibilities) and IT practices, we studied the barriers and challenges of archiving and curating qualitative data at ICPSR.

Purpose: Due to the recent surge of interest in the age of the data deluge, the importance of researching data infrastructures is increasing. The Open Archival Information System (OAIS) model has been widely adopted as a framework for creating and maintaining digital repositories. Considering that OAIS is a reference model that requires customization for actual practice, this study examines how the current practices in a data repository map to the OAIS environment and functional components.

Findings: We observed that the OAIS model is robust and reliable in actual service processes for data curation and data archives. In addition, a data repository's workflow resembles digital archives or even digital libraries. On the other hand, we find that: 1) the cost of preventing disclosure risk and 2) a lack of agreement on the standards of text data files are the most apparent obstacles for data curation professionals to handle qualitative data; 3) the maturation of data metrics seems to be a promising solution to several challenges in social science data sharing.

Original value: We evaluated the gap between a research data repository's current practices and the adoption of the OAIS model. We also identified answers to questions such as how current technological infrastructure in a leading data repository such as ICPSR supports their daily operations, what the ideal technologies in those data repositories would be, and the associated challenges that accompany these ideal technologies. Most importantly, we helped to prioritize challenges and barriers from the data curator's perspective, and contribute implications of data sharing and reuse in social sciences.

1 Introduction

As the research paradigms in science disciplines become data-intensive and collaborative (Hey, Tansley, Tolle, 2009), researchers are promoting data as the "infrastructure of science," critical in forming "the basis for good scientific decisions, wise management and use of resources, and informed decision-making" (Tenopir et al., 2011). Although disciplinary cultural differences exist between social sciences and natural sciences, the former discipline is changing to require greater access to data and more transparency (Guest, 2012; Elman & Kapiszewski, 2013). All of this calls for a strong emphasis on data depositing and sharing.

Despite the recent surge of interest in the age of the *data deluge*, managing digital resources inside a repository for the purposes of preservation and access is neither novel nor unique. Since the late 1990's, digital library communities have been designing and improving the concept of a trusted digital repository, which by its definition should possess key attributes such as "reliable", "long-term access", "managed resources" and for the "designated community"; all are recognized as critical requirements for data management and curation services (Borgman et al., 2007).

The Open Archival Information System (hereafter: OAIS) has been a well-known and widely-adopted conceptual model for creating and maintaining a digital repository. OAIS was proposed two decades ago and, ever since, has become a consensus and a standard for "maintaining digital information over the long-term" (Lavoie, 2004, p.2). The OAIS model can be viewed at three different levels of granularity. The first level describes the external world with which OAIS interacts. The second level defines the internal workflow of OAIS, including six functional entities: ingesting, archive storage, data management, preservation planning, access, and administration (i.e., day-to-day operation). The third level defines the format of possible inputs to the OAIS services.

Considering the important status of OAIS in digital repositories, and the fact that OAIS is a conceptual reference model that requires customization or "translation" into actual practice or a service (Vardigan & Whiteman, 2007), it is important to closely examine the practices in data repositories and review how they adopt the OAIS model. So far, we have seen reports from data repository management teams documenting their adoption of OAIS for data curation services, but there are few third-party studies examining how data curation practices map to the OAIS model. Therefore, the first research question in this preliminary study is:

RQ1. What are the current practices in a data repository?

In order to closely examine how data repository services support social science data sharing, it is necessary to gather information about how data professionals carry out current practices at a research data repository. We conducted a case study on the world's largest social science data repository, the Interuniversity Consortium for Political and Social Research (hereafter: ICPSR). We further mapped the gathered information to the OAIS environment and OAIS's suggested functional entities in order to examine the current practices with a scaffolding reference. We view the case study with ICPSR as an opportunity to examine the support technologies in data repository services. Therefore, our second research question is:

RQ2. What are the current challenges of the underlying technologies at a data repository? What are the desired information technologies (ITs) perceived by employees to support their data repository services?

In addition to RQ1 and RQ2, we also report several interesting findings on the challenges and opportunities in social science data sharing-reuse cycles. We attempt to address the critical inquiry: What are the challenges or barriers encountered by data curation professionals when handling social science data? What general challenges do they see regarding social science data sharing?

By investigating these research questions, we are able to evaluate the gap between current practices and the straightforward realization of the OAIS model. We are also able to identify whether the current technological infrastructure in a leading data repository such as ICPSR is sufficient to support their work; if not, what their desired ITs would be; and the challenges of supporting any ideal technologies. Most importantly, we can prioritize the identified challenges and barriers from the data curator's perspective and obtain a holistic view on data sharing practices in social sciences.

2 Literature Review

We take a funnel approach (i.e., from broader to narrower topics) to review related literature. First, we review the background of the increasing importance of data repositories for research data management in the data deluge age. Then in Section 2.2, we review articles related to the overall operation and workflow in data repositories, specifically in the adoption of OAIS by data repositories. Aligned with our focus on technical challenges, in 2.3, we focus on the technical infrastructure and review its evolution in social sciences.

2.1 The data deluge and research data management

The requirement of research data management can date back to the e-Research[1] movement in the mid-2000s. In the discussions regarding e-Research movements, rapid increasing computational capabilities enables more demands of data-driven scientific discoveries. (e.g., Griffin, 2015). More

[1] The predecessors of e-Research are cyber-infrastructure and e-Science, terms that were coined in the early 2000s to highlight the importance of information technology that supports scholarly activities. According to Borgman (2007), the United States tends to use the term "cyber-infrastructure," whereas Asia, Europe, Australia, and other areas favor the term "e-Science." The prefix "e" in e-Science is usually taken to stand for "electronic," but can also be understood as "enable" or a concept of "enhancement" (p.20).

plans, controls, and management are needed to face the "Data Deluge[2]" and advance data scholarship. In response to the popularity of e-Research and data scholarship, since 2002, the NSF has engaged in organizing councils and digital scholarship workshops, producing several high-impact reports, including Cyberinfrastructure Vision for 21st Century Discovery and Understanding Infrastructure: Dynamics, Tensions, and Design in 2007. This series of movements and endeavors reflects the government's view of research data management: the data deluge requires greater control of data management. Later, in 2009, the US government announced a manifesto of digital stewardship and preannounced a mandate[3] in 2010 that all NSF applications should include a research data management plan.

The NSF's mandate on data management has also become a source to explore how PIs share and reuse their data. Mischo, Schlembach, & O'Donnell (2014) analyzed 1,260 DMPs from July 2011 to November 2013 at the University of Illinois. Mischo et al. found that the most common venues that PIs who preserve their datasets were personal websites (39.9%), personal servers (41.9%), local instructional repositories (e.g., IDEALS at UIUC, 52.9%), and repositories that do not locate on the campus, including disciplinary repositories (21.8%) and other non-UIUC organization (28%). Among all 1,260 DMPs, the authors calculated the occurrence of named repositories which were mentioned by PIs. The arXiv, GenBank, and NanoHub are among the most frequently mentioned. However, Mischo's project did not find significant differences in storage venues when comparing funded grants to unfunded proposals. Additionally, they found that NSF grant applicants underutilized disciplinary repositories. Similarly, Bishoff and Johnston (2015) analyzed 182 DMPs included in NSF grant proposals at the University of Minnesota, and they found a variety of PIs' data sharing strategies.

The NSF mandates signal the important role of data repositories. Because of the growing importance of the research repository in the data deluge age, it is imperative to examine its current state and potential challenges.

2.2 The adoption of OAIS in data repositories

Many data repositories in the archive communities have adopted the OAIS model. For example, as early as 2007, ICPSR published a series of articles and guidelines describing how it integrates the OAIS model into its work model. The outcome, called the "ICPSR Pipeline," adopts the OAIS reference model in the context of social science research data, and is well-documented in "Designing the Future ICPSR Pipeline Process" (Gutmann et al., 2009) and "ICPSR meets OAIS" (Vardigan & Whiteman, 2007).

On the other hand, data repositories such as the UK Data Archive (UKDA) at the University of Essex and The National Archives (TNA) tested and reported on how their systems and processes complied with the OAIS reference model (Beedham et al., 2005). Such work helps to provide guidance for digital repositories and further promote a more cooperative environment among data repositories.

Aside from assessing compliance, there are existing studies that use OAIS as a foundational framework to examine data repositories' practices. For example, Yoon and Tibbo (2011) conducted a content analysis on data submission package elements (SIP as "submission information package" in the OAIS model), and examined submission forms and submission guidelines collected from 14 data repositories in the social science domain.

[2] The term "data deluge" was coined in the early 2000s (e.g., Hey & Trefethen, 2003) in order to reflect the sheer volume and magnitude of research data in the digital age.

[3] Based on this preannouncement, all NSF grant applicants, on or after January 18, 2011, are required to submit a two-page research data management plan describing how to share and manage their data. US federal funding agencies further expanded this mandate in 2013 by adding new data management and data-sharing requirements to grant applications.

2.3 Technical infrastructure in social science

Information communication technologies have enhanced academia with more possibilities and opportunities since the early 90s (Bingham, 1990). In particular, improved technical infrastructure encourage research activities in social science from many aspects, e.g., increasing size and complexity of analyzable research data, availability of easy-to-use social science analysis tools and more channels and patterns for scholars to collaborate efficiently and widely. As revealed by Curty's interview with 13 social scientists (2016), the technological infrastructure of data repositories is important for data sharing and re-use.

However, the challenges underlying the technical infrastructure are not inevitable. Lazer et al. (2009) claimed that the unavailability of easy-to-use social science analysis tools and the insufficient sharing of data impede the advance of the computational social science.

Bingham (1990) identified the three particular aspects of research conduct influenced by technical infrastructure: 1) data collection and analysis, 2) communication and collaboration, and 3) storage and retrieval. Based on Bingham's framework, researchers developed an integrated data collection and analysis platform in social sciences: Common Language Resources and Technology Infrastructure (CLARIN), which is a research infrastructure embedding technical infrastructure to support researchers. This particular tool offers the social sciences and humanities research community with advanced tools to discover, explore, exploit, annotate, analyze or combine data on language resources (Krauwer & Hinrichs, 2014).

Re-focusing on social scientists' data sharing practices on an individual level, Jeng et al. (2016) recently investigated researchers' perceived technical infrastructure and reported four technical limitations that hinder data sharing, namely: platform availability, platform usability, facilities and technical standards. As for technical infrastructure concerning data storage and retrieval, Fecher et al. (2015) summarized three sub- factors as architecture, usability, and management software.

3 Methodology

3.1 Case study: ICPSR

ICPSR was established in 1962 and is the world's largest primary data archive of social science research. As of July 2016, ICPSR holds 8,053 studies, 68,033 datasets, and 196,881 files for download (ICPSR, 2016). As we mentioned in the previous section, ICPSR has adopted OAIS and represents its workflow as the "ICPSR Pipeline" (Beecher, 2009). Although these publications provide documentation for the adoption of OAIS, our examination of ICPSR's workflow and infrastructure is promising and legitimate for the following rationales. Firstly, we are external information science researchers interested in social science data management and services. We provide a different and novel perspective on the data management issues at ICPSR. Secondly, we used a focus group as our research method so that we can collect in-depth data directly from ICPSR practitioners, and we used a bottom-up approach to reconstruct data management and services in ICPSR. The depth of data collected through focus groups cannot be matched by reading published articles.

3.2 Research design and study protocol

Our focus group study uses participatory design and employs a special technique called visual narrative inquiry (Bowler et al., 2014). The detailed execution of this research method has been creatively revised by Mattern et al. (2015) and Lyon et al. (2016; 2017) to enhance the engagement of a focus group. Using a focus group approach in this case study draws upon participants' experiences and encourages interaction among group participants. Our study project

has been reviewed by the Institutional Review Board at the University of Pittsburgh and meets all the necessary criteria for an exemption (IRB#: PRO15050056).

The brief version of study protocol, as shown in Table 1, begins with Stage I: The study information was introduced and consent was obtained from the participants in the focus group. The participants were then invited to describe their backgrounds and explain how their backgrounds have led them to their job positions at ICPSR.

In Stage II: Session of Professional Activities, each participant wrote down their professional activities (i.e., activities related to their day-to-day responsibilities at their institution) regarding data curation or collection development at the institution, one activity per sticky note. The participants then had a discussion among themselves and explained these activities to each other. Next, participants worked on sorting these actions into clusters. Participants were encouraged to leave their seats and go to the whiteboard, self-grouping their sticky notes. They were allowed to use magic markers as a visual aid or re-position the sticky notes however they see fit.

In Stage III, participants were sent back to the table and asked, on another set of sticky notes, to write down the tools related to the sorted concepts on the whiteboard, such as specific software, online services, or homegrown programs. Participants were then encouraged to describe imaginary or desired information technologies.

In the final stage, participants were asked to elaborate about challenges and opportunities regarding data-sharing practices, and were additionally asked about ICPSR's professional activities. While the Appendix lists all actual questions, here are some examples of them in Stage IV:

- Please elaborate more about the differences when curating qualitative, mixed-method, and quantitative data, if any. (Group A)
- What are critical factors that may influence researchers' willingness to share their data? (Group A)
- How do you determine the scope of ICPSR's collection? (Group B)
- Does ICPSR provide other services or support to further connect the data depositors and data reusers? (Group B)

Table 1. Stages in focus group sessions

Stages	Description
I. Warm-up	• The mediators introduce the study information and acquire consent. • Participants describe their background and explain how their backgrounds led them to their current job positions.
II. Session of professional activities	• Each participant writes down their *actions* (professional activities related to their responsibilities at their institution) regarding data curation or collection development at the institution, one action per sticky note. • All participants leave the table and go to the whiteboard, self-grouping their sticky notes. Participants are free to use magic markers as a visual aid or re-position the sticky notes.
III. Underlying information technology activity-collecting *current ITs and desired ITs*	• Participants return to the table and, on another set of sticky notes, write down the tools related to the concepts on the whiteboard, such as certain software, online services, or homegrown programs. • Participants describe desired information technologies.
IV. Semi-structured interview	• Each participant elaborates more about their actions in curation, acquisition, and collection development.

Note: The detailed procedure is attached in the Appendix

3.3 Sampling and data collection

Our study consists of two focus groups and one individual interview, all of which were conducted in June 2016 onsite at the ICPSR headquarters in Ann Arbor, Michigan. In total, eight

ICPSR employees participated in the study, and seven out of eight were directors or senior managers (at least 10 years of experience).

The sampling method in this study is expert and convenience sampling, targeting data curation professionals and other professionals who work in ICPSR. To contact such a specific target population, the invitations were sent according to two categories ("Administration" and "Collection Development & Delivery") on the staff directory webpage or were referred by ICPSR employees.

Table 2 summarizes work experience (in years) and general responsibilities of our participants. Group A's session lasted about 75 minutes, Group B's session lasted about 65 minutes, and the individual interview lasted about 40 minutes.

Table 2. Participant Background

Groups	ID	Year of experience	General responsibilities in ICPSR
A	P01	>10 years	Curation
	P02	>10 years	Curation, data processing
	P03	<10 years	Curation, data processing
B	P04	>10 years	Acquisition, administration
	P05	>10 years	Customer relations, administration
	P06	>20 years	Curation, administration
	P07	>20 years	Administration
*	P08	>20 years	Administration

Note: * Individual interview was conducted.

The topics discussed in these focus groups and the interview were as follows.

Group A — "Curation Services." The emphasis of Group A was on data curation services. Participants include P01 to P03. Figures 1a-1d illustrate a more detailed breakdown of our focus group procedure. In Stage II (see the description in Table 1), each participant first wrote on their individual sticky notes and attached them to the whiteboard in the conference room (Figure 1a). Individual participants were welcome to write additional notes after a discussion with the other participants in their group. Participants were also invited to take advantage of visual aids to elaborate about their actions (Figure 1b). In Stage III, participants added underlying IT and desired IT on the whiteboard using yellow rectangular sticky notes (Figure 1c). In Figure 1d, participants continued adding different visual aids, such as the section that reads "OpenICPSR" with a dashed line to the final outcome.

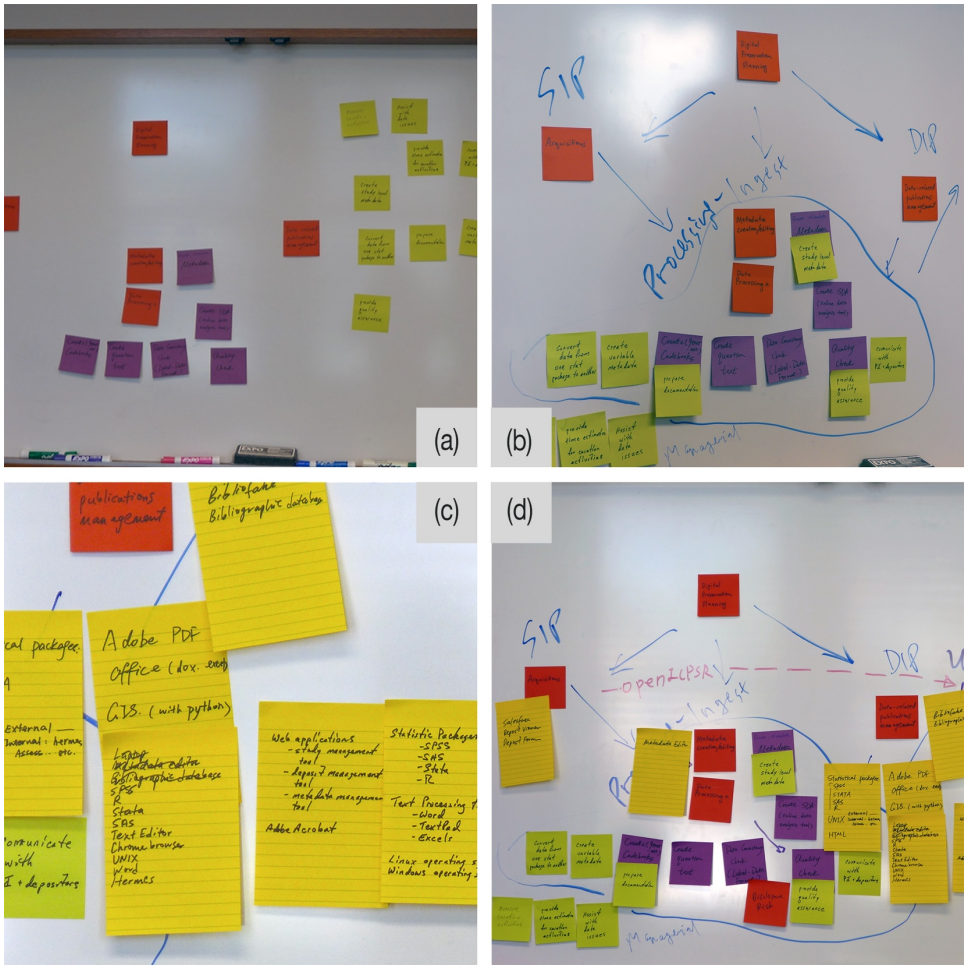

Figure 1. Group A activity breakdown

Group B — "Collection Development." The emphasis of Group B was on collection development at ICPSR. All participants in Group B are directors or managers, and their daily responsibilities extend beyond collection development, including acquisition, delivery, supervising, customer relations, outreach, administration, and preservation planning. Participants include P04 to P07 in Table 2. A more detailed breakdown can be found in Figures 2a-2d. Firstly, all Group B participants attached their notes to the whiteboard without any sorting or classification (Figure 2a). Later, the participants grouped similar actions into columns (Figure 2b) and named each cluster themselves (Figure 2c). Note that the focus group mediators did not directly participate or interfere with participants' sorting process. Finally, as shown in Figure 2d, the participants added their IT practice notes onto the white board.

Figure 2. Group B Activity breakdown.

Interview — In addition to the two focus groups, we interviewed an experienced director (P08 in Table 2) to add valuable perspective and to clarify some points regarding our research questions. Questions include:

1) a follow-up on how curation professionals communicate with data depositors about potential disclosure risk;

2) factors that can influence a researcher's willingness to share data with ICPSR;

3) the potential challenges and opportunities for social scientists when sharing their qualitative data.

After collecting data from the research sites, we digitalized all the sticky notes and entered data into a spreadsheet-style table. Specifically, the workflows or clusters created by participants in both focus groups were digitalized by a digital camera. These digital images allowed us to re-create and analyze the focus group results. All conversations that happened during the focus groups and the interview were recorded and transcribed. Participants' quotations on transcription files are managed using ATLAS.ti, a qualitative data analysis software.

4 Findings

In this section, we report interesting findings observed in the data collected from the two focus group studies (with Participants P01-P07) and in the individual interview with P08, including direct quotations. The results are divided into three sections. In 4.1, we discuss the overall practices (data curation actions) at ICPSR, while in 4.2 and 4.3 we answer RQ2 by reporting the current IT

practices and desired IT from data curation professionals' perspectives. Finally, in 4.4, we report our findings and observations regarding challenges and opportunities in social science data sharing.

4.1 Data curation workflow at ICPSR

As we collected participants' actions, results presented by the participants in Group A resemble the ICPSR Pipeline (Figure 1d). However, the results presented by the participants in Group B were mostly bottom-up action clusters (Figure 2d), which have little similarity with the OAIS structure.

Based on the clusters of sticky notes, we further integrate the participant-created action clusters with the OAIS model, presented in Figure 3. In Group A's reported actions, after receiving an SIP (submission information package) from the data depositors, data processors perform a series of activities to prepare the data for documentation, such as "building metadata" and "creating codebook." The various actions in the data processing stage seem to be interrelated and not necessarily sequential, as the participant P02 expressed, "*once we get everything together, then we start to put all these pieces together and they're all interrelated. You don't have to do one before the other.*"

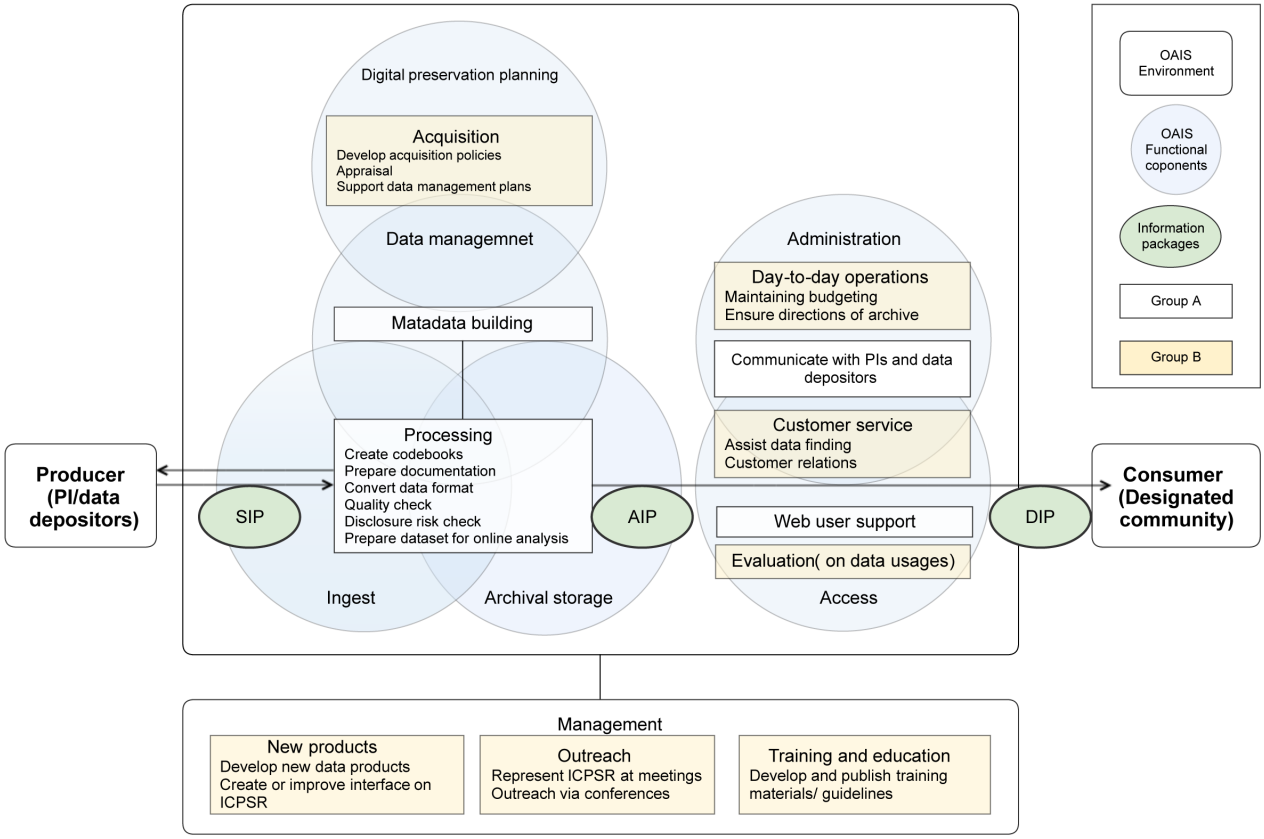

Figure 3. Participant-reported activities and OAIS components at ICPSR

Unlike Group A's use of a workflow to explain their actions, Group B sorts their actions (shown in yellow rectangles in Figure 3) into eight clusters: curation, new products, acquisition, outreach, evaluation, management, customer services, and training & education. We found that Group B's action clusters overlay with other OAIS functional components except for data processing and metadata building.

However, we find that only a portion of the action clusters can be perfectly covered by a single OAIS function entity. For example, in Figure 3, the actions in "Ingest," "Archival Storage" and "Data Management" are overlapping, suggesting that they require support from multiple entities. This is exactly the purpose of viewing OAIS as a reference: although the OAIS model provides a

high-level reference guideline, data archives or repositories should expect to work out the details and customize the model to reflect their own needs.

4.2 Current IT practices

Table 3 enumerates the reported technologies based on associated action clusters. We find that participants mentioned more IT tools related to "data processing" and their effort to develop "new products". We also see that office software (such as word processors, text editors, and spreadsheets) are the most common tools. On the other hand, participants reported that they prefer Linux-based operating systems in their work environment, and most of their work was done under the Linux environment: "*We do our work in the Linux environment but we have Windows environment that we can also work in as well"* (P02); "(We) *log on PC but using Linux"* (P01).

Table 3. Current information technologies reported by participants

Action clusters	Current IT	Participants
Acquisitions	Metadata editor, lead management tool, deposit viewer, deposit form, spreadsheet, email	P01, P04, P05
Web team	Bibliographical database (bibliofake), PDF applications	P01, P03
Processing	Word processor, spreadsheet, GIS scripts, SPSS, SAS, Stata, R, text editor, Linux, Windows, Study management tool, deposit viewer, metadata editor, PDF applications, web browser, Unix, Hermes, HTML	P01, P02, P03
New products	Online questionnaire software, usability testing tool, web-hosted service for webinars, responsive design tools, email, Unix, HTML, XML, word processor, funding database, lead management tool, deposit form, email	P04, P05, P06, P07
Outreach	Web-hosted service for conferences, presentation software, Google Analytics, word processor	P04, P05, P07
Evaluation	Text visualization tool, Google Analytics, data mining tools, data visualization tools, online questionnaire software	P04
Management	University financing reporting system, spreadsheet, word processor	P04, P06, P07
Customer service	Email tracking system, web-hosted service for webinars, email, social media, online video	P04, P05, P07
Training and education	Word processor, web-hosted service for webinars, email extension (Boomerang for Gmail)	P04, P05, P06, P07

According to Group A (in which participants used Figure 4 to explain the internal tools), we find that core actions in the data processing cluster mostly rely on internally-developed applications, which include:

- Herme (a file-converting tool that can convert data files from one format to another, such as from SPSS to CSV and SAS),
- Deposit Form (creating the package after data depositors or PIs finish the deposit;
- Deposit Viewer (allowing curators at ICPSR to view metadata about deposits),
- Metadata Editor, "the primary environment for creating, revising, and managing descriptive and administrative metadata about a study" [Beecher, 2009, para 5]); and the
- Bibliofake (a database created for storing "bibliographic information and exports it into a format in a system that can use to render that information on the website" [P01]).

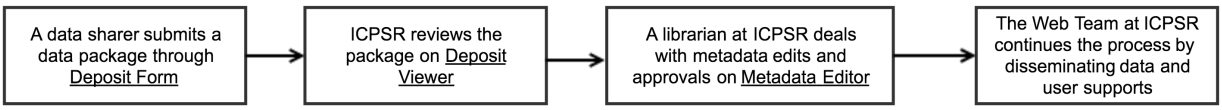

Figure 4. The internal workflow of processing data package at ICPSR (provided by P01)

We concluded that there is no single integrated platform that handles multiple action clusters simultaneously. On the other hand, some actions, such as processing, involve more tools and thus are more complex than others. As shown in Figure 5, P03 wrote down a couple tools that she uses during data-package processing. Participant P02 elaborated on what P03 wrote by saying: "*I'm mostly surprised these are all the stuff that we're doing"* (P02).

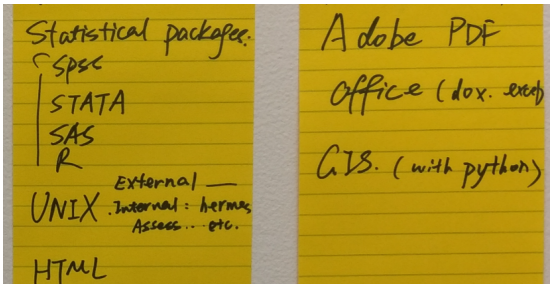

Figure 5. A data curator's toolbox for processing data packages at ICPSR (provided by P03)

4.3 Desired information technologies for data curation professionals

As shown in Table 4, Group A precisely describes the tools and technologies needed to address their daily challenges. For example, they would like to have technologies that can automatically extract all of the metadata from an input dataset; as one participant mentioned, "*Wouldn't it be great if there was a form where you uploaded a file and that system would automatically extract all of the metadata for that file" (P01).* They also desire tools that can help "flag" possibly sensitive or harmful content, and technologies that can automatically discover possible identifier combinations. Almost all participants in Group A mentioned the disclosure check: *"You always have to decide, 'Is it harmful?" What's the level of harm that's going to happen and what's the level of sensitivity?"*" (P02). *"[S]ometimes you miss human sense of what kind of information is dangerous. I know there are tools for disclosure risk but they are not efficient and they cannot identify information [that] we actually identify as disclosure risk" (P03).*

Action clusters	Current challenges	Ideal IT solutions	Participant
Processing	Metadata are manually extracted.	Technologies that can automatically extract most of the metadata from an input dataset	P01
	Disclosure risk or sensitive content are manually checked	Technologies that can help 'flag' possible sensitive or harmful contents; automatically find out possible combination of identifiers	P02 & P03
	Quality control	Tools that can speed up the process for ensuring data quality by checking if file crushes, errors, executing dataset and scripts	P02 & P03
Administration	Hard to estimate "cost" for every single case	Technologies that can estimate needed resources before assigning laboring and money.	P06

Management	Hard to synchronize with other departments in the institution	One united and transparent system that can instantly and actively inform or facilitate communication and synchronization between internal departments or separate archives; that can reduce time between contacts	P04 & P06
Training and education	--	A platform that can enhance user engaging and allows customization for training purpose	P05

Table 4. Current challenges and ideal IT solutions reported by participants

On the other hand, since all the participants in Group B are in management positions, their descriptions of ideal technologies are less specific but more comprehensive than those provided by Group A. For example, they desire automated tools to estimate the cost of each study, and systems that can unite multiple departments. Participant P04 called for tools that can "*make things connect and interact across because now we have all of these silos, systems with the University (U of Michigan) with ICPSR.*" She also anticipated this one-stop-shopping system can be developed sooner: *"…the hope is that over the next few years, we'll be putting in a new enterprise system, securities and if this will connect some of those things better or just take one place that you put everything and go in and grab what you need" (P04).*

4.4 Challenges and opportunities of social science data sharing

In this section, we discuss the challenges and opportunities regarding social science data sharing. Table 5 lists out the challenges and opportunities that we identified through the focus group sessions (P01-P07) and the interview with P08. Challenges and opportunities occur at various levels, ranging from individual researchers, their discipline communities and data infrastructures, to the national level.

We would like to note that a *cross-level* investigation is needed because a challenge that exists in one level may be solvable by an opportunity existing in another level.

Table 5. Challenges and opportunities toward social science data sharing in different levels

	Challenges	**Opportunities**
Individual level	▪ Social scientists' individual concerns about data sharing: o PI's confidentiality concerns (P01, P08) o PI's confidence of data sharing (P01, P08) ▪ Lack of reward model (e.g., data are not recognized as research products) (P01)	--
Community level	▪ Lack of agreement on the standard of text data files in qualitative studies (P01, P02) ▪ Low awareness of data sharing in social sciences (P01, P02)	▪ Data metrics (P01, P03, P05, P06, P07)
Infrastructural level	▪ Labor-intensive process of data curation, especially for qualitative data (P01, P02, P03, P04) ▪ Hard to fulfill various community needs at once (P04, P05, P06)	▪ Active curation (P04) ▪ Enclaves and embargo settings (P01, P02, P08)
National level	Can be both challenges or opportunities: ▪ Regulations and mandates on data sharing at the national level (P07)	

4.4.1 Challenges

We explain in detail each of the six identified challenges in social science data sharing from data curators' perspectives.

Labor-intensive process of data curation, especially for qualitative data. Preparing qualitative data for sharing requires extra time and effort. For data curation professionals, open-ended responses can be text-heavy, and the processing cost for time and labor is hard to estimate. For example, participants P01 and P03 had a conversation and described the effort of processing qualitative responses, "*If you have to read through 10,000 responses"* (P01); "*Sometimes they mention the names, other people name their names or the exact date of something happened, that's the information we don't want them to (reveal)"* (P03).

Lack of agreement on the standard of text data files. Participants also suggested that it is necessary to adopt and inform data depositors about sustainable digital file formats and standard metadata for qualitative data. Regarding qualitative data curation, ICPSR widely accepts a series of text-based files, whereas the PDF is an exception. "*We have a very good handle on that where we put it into an ASCII text file or set ups with qualitative stuff. It's not as cut and dried to use Word as a proprietary format, to use XML, or PDFs, or if you put it in a PDF, is it searchable?* (in a rhetorical tone)" (P01).

The designated community problem: difficult to fulfill various community needs at once. Data curators often face the designated community problem—that is, they find it difficult to clearly identify the target users of a data repository. For example, P06 expressed that from time to time they would ask themselves about who the designated community of ICPSR is: "*there's customers* (research institutions who pay the annual membership fee to ICPSR) *and there's users* (data reusers), *and then people who use our data are often not the people who pay for it"* (P06). Therefore, the team may need to use additional labor and time to repeatedly review potential stakeholders.

Social scientists' individual concerns about data sharing. Several observations made by the data curators can help explain why a social scientist might refuse to share their data. On the top of the list, it seems that social scientists are most worried about "sensitive data" and have "confidentiality concerns": "(One barrier) *is fear of confidentiality or privacy issues, feeling like they have some sensitive information or data that they won't be able to release and so but they don't know about these other channels that are available*" (P01). In addition, qualitative approaches usually deeply involve the researchers' worldviews; such subjectivity might influence how qualitative researchers view and value their research data, and thus may sometimes result in resistance to archive and share their data. Participant P08, speaking from an administrator's perspective at ICPSR, shared his thoughts on qualitative data sharing and still believes qualitative data sharing is possible: *"data sharing tends to be weakest in qualitative fields because qualitative researchers many of them for various ideological and ontological reasons believe they can't share their data, But it's not true that that's not universal*" (P08).

Awareness of data sharing is increasing but still low. The majority of faculty and graduate students in social science fields do not share data or are aware of its importance. Participant P01 related this phenomenon to the low awareness of perceived benefits: "*not everyone or even not the majority maybe know that publishing data or putting your data into a repository is a good thing"* (P01). On the other hand, the **lack of a reward model** can be another critical hindrance for researchers' data sharing in general. The same participant compared data products with research articles*: "[Y]ou've probably gone through the tenure process where your reviewers, if you publish a data collection, or let's say you publish an article, but you also spent… a lot of time publishing a data product. That data product is used by thousands of people around the world. That article maybe was read by ten people but it was in science or nature, that would be a tenure, the data product, from what I understand, doesn't get nearly the eyeballs or attention"* (P01).

4.4.2 Opportunities

Despite the challenges, we also observe four encouraging opportunities for social science data sharing from data curators' perspectives. Among these opportunities, data metrics were on the top of the list and were mentioned by participants in both focus groups.

Secure dissemination services, such as an enclave policy. Several participants (P01, P02, and P08) mentioned the enclave policy at ICPSR. "*We do have a restricted data use policy. People can apply and receive the data from our secure downloads if they can have it or if it's just really restricted, we can put it in a physical enclave or we have a digital enclave where people can log into it and only use the data there"* (P02). Research data infrastructure also pays attention to potential disclosure risks, and data repositories such as ICPSR often offer secure dissemination services. Such security mechanisms are an opportunity to address an individual's confidentiality concerns, mentioned above.

The maturity of data metrics. Despite imperfections, citation-based bibliometric methods have been widely used to evaluate scholars for promotion, tenure, hiring, or other recognizing mechanisms (Borgman, 2007). However, data citation or data publication is not a common recognizing mechanism in academia. In our study, participants across both focus group sessions mentioned the lack of recognition of data citation repeatedly: "*It's funny that you look at the citation or reference of a book or a journal article and that's very well established in research and academia but this you can't say nearly the same for our data collection. It's not yet considered a first rate research product and as a result it affects other aspects of the research life cycle"* (P01). Although NSF (2013) has recognized data as a research product since 2013, it is taking time for academia to form an agreement to adopt data publications as research products. To encourage data sharing in social sciences, the community can consider data sharing a kind of academic contribution by adopting data metrics. P05 in Group B expressed her positive attitude about the connection between providing data metric services and a PI's willingness to share data at ICPSR: "… *individual PI, they might be excited to see downloads and citations and search…They can say, look at how much impact we have had… [B]ut again it's all still relatively new"* (P05).

Call for an "active curation." To speed up the process of data curation, participant P04 mentioned the concept of *active curation*, a new model of accomplishing data curation piece by piece (Myers et al., 2015). The traditional curation model usually requires everything to be available before proceeding to the next step, whereas active curation is an incremental model where metadata and elements can be added over time: "*That's where my wishes came from, reducing the time it takes to get data in the door, supporting active curation, so maybe we can get the data in before they have to actually deposit it or let others use it, but if we can help them along the way"* (P04). This opportunity not only reduces curation time, but also ultimately allows PIs to proactively update their datasets. This is beneficial for PIs who are hesitant to share data because they are afraid that errors or mistakes in their data will be pointed out.

Call for a national policy. Participant P07 mentioned the UK, which has national policies that encourage UK researchers to submit datasets to the national archives: "*Yeah, and many other countries like UK, there is requirement that people deposit their data in a particular place"* (P07). There is no nation-wide data sharing infrastructure as of 2016 in the US, and there is no universal guideline for selecting a data archiving platform. The existence of a national policy can simplify PIs' effort to select a data archiving platform, but it would be challenging to build supporting infrastructure for such a policy.

5 Discussion and Implications

The adoption of OAIS. The curation and collection development practices that we collected greatly resemble the ICPSR Pipeline, which adopts the OAIS model as its high-level framework. We observe that although data infrastructures such as ICPSR have the freedom of being detached from the OAIS model, their practices still strongly resemble OAIS at large. This observation may indicate

two things: 1) the OAIS model is robust and reliable in actual service processes for digital curation and digital archives; 2) a data repository's workflow resembles digital archives or even digital libraries. However, even though we observe that ICPSR's workflow and practices resemble the OAIS model, the institution also revises the model to meet their own needs. For example, some activities require interaction between multiple function components; we observe overlapping areas between "Ingest," "Data Management," and "Archival Storage" as well as "Administration" and "Access". This finding may reflect the employees' actual skillsets or work allocations in a data repository, and thus can serve as a reference for other data repositories.

Current IT practices are specific and their coherence needs to be improved. As for current IT practices, we find that ICPSR's main actions in data processing are handled by internally-developed tools, which is consistent with the observation made by Jeng and Lyon (2016). That is, social science projects tend to require a unique set of IT functionalities, and thus it is common to develop customized tools for a specific task rather than using general-purpose tools for multiple tasks. Current IT challenges include data disclosure checks and the coherence problem. For disclosure management, a more intelligent tool is desired, which can help improve efficiency and the decision-making process by providing additional information, such as highlighting possibly risky texts. Addressing the coherence problem requires a better platform on which people can work together smoothly. However, we did not ask participants to elaborate on the desired IT's possible functionalities and appearance, so a future specific participatory study is anticipated to capture more details.

Gaps in data-sharing practices in social sciences at scale. The aforementioned discussion reveals several challenges as well as opportunities. Again, although a particular challenge exists on one level (e.g., PIs' concerns about data sharing at the *individual* level), it may be resolvable by an opportunity existing on another level (e.g., the maturity of data metrics at the *community* level). Data curation remains challenging to scale due to privacy concerns and its labor-intensive process. To resolve this scaling issue and handle big data in social sciences, researchers require better and automated tools to help detect or perform disclosure checks. In addition, consistent with prior work (Jeng, He, and Oh, 2016), data curators also express their worries about the low awareness about data sharing in social sciences. However, it is unclear what the root cause of this is, given that every stakeholder appears to support of data sharing. As a bottom-up approach, we suggest that it might be helpful to expose early-career social scientists (i.e., senior graduate students, post-doctoral researchers, and assistant professors) to trainings on research integrity, data transparency, and the spirit of open research.

6 Conclusion

Through two focus group sessions and one individual interview with eight ICPSR employees, we evaluated the gap between ICPSR's current practices, IT practices, and their adoption of the OAIS model. We also revealed the current ITs which support data curation professionals' daily operations, the ideal technologies these professionals desire, and the challenges with these ideal technologies. Most importantly, we helped to prioritize barriers from data curators' perspectives and we contributed implications about data sharing and reuse in social sciences.

Based on participants' point of views, several challenges and opportunities regarding data sharing in social sciences are also observed. Our reported findings reveal several challenges (such as data ownership and confidentiality concerns); however, to reiterate, a particular challenge may exist on one level (e.g., PIs' concerns about data sharing at the individual level), and be resolvable by an opportunity existing on another level (e.g., the maturity of data metrics at the community level). Data sharing and curation in social sciences remain challenging to scale due to privacy concerns and

a labor-intensive process, especially with regard to qualitative data sharing. Better and automated tools would be required to help detect or perform disclosure checks.

This case study on ICPSR might be limited because the study only focused on one repository and participants were self-selected. In addition, interviewing more participants on data-curation-related responsibilities (e.g., the Web team, the IT team, or metadata librarians) would allow us to yield a more robust outcome and present a more holistic workflow at a data repository.

One future work is to compare our results with related work based on the investigation on social scientists' data-sharing and reuse practices (e.g., Jeng, He, and Oh, 2016; Yoon, 2016; Curty, 2016). A cross-level (i.e., individual, institution, community, and infrastructure) triangulation is exceptionally needed for capturing the whole picture of data sharing and reuse practices in social sciences. Another future direction is to compile a list of design principles for improving the design of a data curation system, based on the collected IT practices and ideal technologies in this study.

7 Acknowledgements

The authors thank the iFellowship, guided by the Committee on Coherence at Scale (CoC) for Higher Education, sponsored by the Council on Library and Information Resources (CLIR) and Andrew W. Mellon Foundations; as well as Beta-Phi-Mu Honor Society, which provided research funding for this project. This study is also partially supported by the project titled Research on Knowledge Organization and Service Innovation in the Big Data Environments funded by the National Natural Science Foundation of China (No. 71420107026).

The authors also thank Drs. Nora Mattern, Liz Lyon, Sheila Corrall, Jian Qin, Jung Sun Oh, and Stephen Griffin for their invaluable comments and suggestions on this research project. Last but not least, the authors thank all participants and people who helped facilitate the field study at ICPSR for their valuable input and assistance.

8 References

Beecher, B. (2009, November 19). The ICPSR Pipeline Process. Retrieved October 31, 2016, from http://techaticpsr.blogspot.com/2009/11/icpsr-pipeline-process.html

Beedham H, Missen J, Palmer M, Ruusalepp R (2005) Assessment of UKDA and TNA compliance with OAIS and METS standards. Joint Information Systems Committee (JISC), United Kingdom, http://data-archive.ac.uk/media/1692/OAISMETS_report.pdf.

Bingham, J. L. (1990). Information Technology and the Conduct of Research. Bulletin of the Medical Library Association, 78(3), 326.

Bishoff, C., & Johnston, L. (2015). Approaches to data sharing: An analysis of NSF data management plans from a large research university. Journal of librarianship and scholarly communication, 3(2), eP1231.

Bohémier, K. A., Atwood, T., Kuehn, A., & Qin, J. (2011). A content analysis of institutional data policies. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (pp. 409-410). ACM.

Borgman, C. L., Wallis, J. C., Mayernik, M. S., & Pepe, A. (2007, June). Drowning in data: digital library architecture to support scientific use of embedded sensor networks. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (pp. 269-277). ACM.

Bowler, L., Knobel, C., & Mattern, E. (2015). From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media. Journal of the Association for Information Science and Technology, 66(6), 1274-1293.

Curty, R, G. (2016). actors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study. International Journal of Digital Curation (IJDC), 11(1): 96-117.

Elman, C., & Kapiszewski, D (2013). A Guide to Sharing Qualitative Data. Center for Qualitative and Multi Method Inquiry (CQMI), Syracuse University.

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

Griffin S. (2015). Libraries in the Digital Age: Technologies, Innovation, Shared Resources and New Responsibilities, Chapter in "Communication and Technology", Volume 5 of the series "Handbook of Communication Science", Ed. by Cantoni, L., Danowski, J., De Gruyter Mouton.

Guest, G., Namey, E. E., & Mitchell, M. L. (2012). Collecting Qualitative Data: A field manual for applied research. Sage.

Gutmann, M. P., Evans, B., Mitchell, D., & Schürer, K. (2009). The Data Archive Technologies Alliance: Looking towards a Common Future. In IASSIST Conference.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). The Fourth Paradigm; Data-Intensive Scientific Discovery. Redmond, Washington: Microsoft Research.

ICPSR. (2016). Size of ICPSR's Holdings. Retrieved October 31, 2016, from https://www.icpsr.umich.edu/icpsrweb/content/about/history/

ICPSR. ICPSR: A Case Study in Repository Management. Retrieve from https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html

Jeng, W. & Lyon, L. (2016). A report of data-intensive capability, institutional support, and data management practices in social sciences. International Journal of Digital Curation (IJDC), 11(1): 156-171.

Jeng, W., He, D., & Oh, J. (2016). Toward a conceptual framework for data sharing practices in social sciences: A profile approach. In the proceedings of the ASIS&T 2016 Annual Meeting.

Kim, Y. (2013). Institutional and Individual Influences on Scientists' Data Sharing Behaviors. Unpublished dissertation. Syracuse University.

Krauwer, S., & Hinrichs, E. (2014). The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) (pp. 1525-1531). European Language Resources Association (ELRA).

Lavoie, B. F. (2004). The open archival information system reference model: Introductory guide. Microform & imaging review, 33(2), 68-81.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915), 721.

Lyon, L., Jeng, W., & Mattern, E. (forthcoming). Research Transparency: A preliminary study of disciplinary conceptualisation, drivers, tools and support services.

OCLC (2002). Trusted Digital Libraries: Attributes and Responsibilities. Retrieved from https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

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.

Mischo, W. H., Schlembach, M. C., & O'Donnell, M. N. (2014). An Analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals. Journal of eScience Librarianship, 3(1), 3.

Myers, J., Hedstrom, M., Akmon, D., Payette, S., Plale, B. A., Kouper, I., ... & Kumar, P. (2015). Towards sustainable curation and preservation: The sead project's data services approach. In e-Science (e-Science), 2015 IEEE 11th International Conference on (pp. 485-494).

National Science Foundation (NSF). (2013, January). National Science Foundation's Merit Review Criteria: Review and Revisions. Retrieved October 31, 2016, from https://nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_sigchanges.jsp

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).

Vardigan, M., & Whiteman, C. (2007). ICPSR meets OAIS: applying the OAIS reference model to the social science archive context. Archival Science, 7(1), 73-87.

Yoon, A., & Tibbo, H. (2011). Examination of Data Deposit Practices in Repositories with the OAIS Model. IASSIST Quarterly, 35(4). Chicago

Yoon, A. (2016). Data reusers' trust development. Journal of the Association for Information Science and Technology.

Appendix. Focus Group Protocol

Group A (data curation professionals): 60 minutes

Time	Activity	Mediator actions	Question prompts
00:00-00:03	Review information and consent	Distribute introduction script Obain consents on: ▪ proceed the focus group ▪ use recorders, and ▪ data will be shared	Thank you for your participation. I believe your input will be valuable to this research and in helping grow all of our professional practice. Approximate length of interview: 60 minutes, two group activities and three major questions
00:03-00:15	Warming up	Mediator actions ▪ Set timer ▪ Set recorder Taking note: ▪ Education background ▪ Career history ▪ Year of experience ▪ Primary activities	Please take us back through a little history in your career that brought you to this current position. Also, we would like to know more about your current work at ICPSR. Prompts: ▪ How long have you been involved in your current job? (What year were you involved) ▪ What primary tasks does your job involve?
00:15-00:35	Concept construction	Distribute post-its (different colors) Process: individual write post-its stick to write board sort cluster 📷 Take a picture Distribute easel pad 📷 Take a picture Distribute post-its (yellow post-its) 📷 Take a picture	**Question 1**: What are your activities as a curation professional to support data curation? Prompt: before/ after data submitting Process: individual write post-its→ stick to write board → sort→ draw cluster→ ask participants if there is anything left. **Question 2**: Now we have n clusters, could you explaining the relationships among the activities **Question 3**: What are the tools that you use for your actions in curation? Prompts: ▪ Computer equipments ▪ Software

			▪ Online services ▪ Internal toolkits? Question 4: Can you think of any desired tools or technology (tools may not exist) which can facilitating your actions at ICPSR? (talking only, do not distribute sticky notes)
00:40-00:55	Questions about qualitative data curation	--	**Question 5A**: Have you ever curated qualitative data? If yes, jump to 5B If no, have you heard about your colleagues or others in ICPSR curating qualitative data? Do you have any observation? **Question 5B:** Please tell us about the difference when curating qualitative, mixed method, and quantitative data, if any. Is there any special case or example that you would like to share? **Question 6:** Based on your observations and experience as curation professionals in ICPSR, what are the critical factors that may influence a PI's willingness to share his/her data? Prompts: ▪ Has a PI ever told you about or you have heard---the factors could influence PI's willingness? ▪ Are they from: ▪ Individual incentives ▪ Research culture ▪ Institution
00:55-00:60	Debriefing	--	Suggestions about research instrument? Was anything unclear?

Group B (collection development professionals): 60 minutes

Time	Activity	Mediator actions	Question prompts
00:00-00:03	Review information and consent	Distribute introduction script Obain consents on: ▪ proceed the focus gorup ▪ use recorders, and ▪ data will be shared	Thank you for your participation. I believe your input will be valuable to this research and in helping grow all of our professional practice. Approximate length of interview: 60 minutes, two group activities and three major questions
00:03-00:15	Warming up	Mediator actions ▪ Set timer ▪ Set recorder Taking note: ▪ Education background ▪ Career history ▪ Year of experience Primary activities	Please take us back through a little history in your career that brought you to this current position. Also, we would like to know more about your current work at ICPSR. Prompts: How long have you been involved in your current job? (What year were you involved) What primary tasks does your job involve?
00:15-00:30	Concept construction	Distribute post-its (different colors) Process: individual write post-its stick to write board sort cluster 📷Take a picture Distribute easel pad 📷Take a picture Distribute post-its (yellow post-its) 📷Take a picture	**Question 1:** What are your responsibilities in supporting collection development and delivery in ICPSR? Prompt: before/ after data submitting Process: individual write post-its→ stick to write board → sort→ draw cluster→ ask participants to clarify if there is any sticky note unclassified. **Question 2**: Are there any tools that you use? Prompts: Computer equipments Software Online services Internal toolkits? (yellow post-its) Question 3: Can you think of any desired tools (tools may not exist) or technology which can facilitating your actions at ICPSR?

			(talking only)
00:30-00:55	Questions about collection development and vision		Now we have a couple questions related to collection development, collection delivery, management, and marketing topics in ICPSR. **Question 4:** How do you determine the scope of ICPSR's collection? We read about ICPSR's collection development policy, we read about the high-priority areas including sexual orientation, social media, immigration, and so on. How does ICPSR decide which areas should be given priority? Prompts: ▪ Are these decisions from ICPSR's interval decision? ▪ members' opinions or feedback? ▪ Recent research hot topics (recent publications)? ▪ or community or specific researchers' demands? ▪ How does ICPSR decide to add a new interest? **Question 5:** This question is related to appraisal standards in ICPSR. Please tell us about how ICPSR applies the selection and appraisal criteria for data from mixed-method study or qualitative study. Are they different from quantitative one? Is there any special case or example that you would like to share? Prompts: When will data be referred to the QDR? **Question 6:** This questions is about OpenICPSR. Given the differences between OpenICPSR and ICPSR, please share your experience with us about how ICPSR handles or manages these two different collections. Is OpenICPSR within the scope of ICPSR? Prompts: ▪ Do ICPSR members mention anything about ICPSR? (Their experience with OpenICPSR?)

			▪ What is your observation? ▪ Is there any plan for further promoting Open-ICPSR to ICPSR members? **Question 7:** Currently ICPSR supports search interface and track utilization for data sharers and reusers. Does ICPSR provide other services or support to further connect the data depositors and reusers?
00:55-00:60	Debriefing	--	Suggestions about research instrument? Was anything unclear?