In between data sharing and reuse: Shareability, availability and reusability in diverse contexts

Ayoung Yoon
Indiana University Purdue University Indianapolis
ayyoon@iupui.edu

Wei Jeng
University of Pittsburgh
wej9@pitt.edu

Renata Curty
Universidade Estadual de Londrina
renatacurty@uel.br

Angela Murillo
University of North Carolina at Chapel Hill
amurillo@email.unc.edu

ABSTRACT

Although data availability cannot be considered the sole predictor of effective reuse, because only accessible and well-managed data can make reuse possible, data reuse is contingent on the availability of data. It is critical to understand the nature of shareability, availability, and reusability, and their synergy and relationships, to further understand the dynamics of data reuse practices in multiple environments and contexts.

This panel aims to closely examine aspects related to data shareability, availability and reusability, based on the assumption that each condition poses a cumulative effect on each other and impacts the efficiency and efficacy of the data reuse process. The panelists will present their findings and perspectives in a diverse context regarding data availability, between academic and non-academic; data shareability and data reusability, social sciences and earth science, researchers' and journal publishers' perspectives. Presentations will be followed by an interactive session taking the team-based approach, with the expectation to engage participants in discussion and experience-sharing, and to contribute in terms of practice and research with the current knowledge and applications.

Keywords

Data availability, data reuse, data sharing, reusability, data curation.

INTRODUCTION

The potential for data reuse to impact scientific research and help solve many social problems is increasingly recognized. Data reuse can be only accomplished through data sharing and availability (Hey, Tansely, & Tolle, 2009), and has immediate and long-term benefits to discovery and innovation, including the ability to extract additional value from data, ask new questions of existing data, reexamine existing scientific theories and models, and to enable reproducible research (Borgman, 2012). On this note, Fabic, Choi, & Bird (2012) demonstrate how increased availability and proliferated data stimulate peer-reviewed research as well as evidence-based policy and program development.

There is little discussion about how data producers and re-users handling data shareability, availability, and reusability outside of the context of higher education institutions, research centers, and laboratories. It is critical to understand the nature of shareability, availability, and reusability, and their synergy and relationships, to further understand the dynamics of data practices in multiple environments and contexts. The availability of data, the ability and motivation to share data and reuse data has increased significantly with the proliferation of data repositories and infrastructures dedicated to house data for further reuse. This panel aims to closely examine aspects related to data shareability, availability, and reusability, based on the assumption that each condition poses a cumulative effect on each other and impacts on the efficiency and efficacy of the data reuse process; the desired outcome of data sharing endeavors.

Data shareability, availability, and reusability

Even though data availability cannot be considered the sole predictor of effective reuse, because only accessible and well-managed data can make reuse possible, data reuse is contingent on the availability of data. In other words, ample and improved data availability acts as a key precondition for ultimate reuse. It is important to pinpoint though, that not all data are shareable and therefore, can be easily made available. The degree of shareability is established taking into account legal, ethical, and technical aspects with regard to the nature of data. For example, data involving vulnerable populations---such as minors, survivals experienced domestic violence, or subordinates in the workplace---sharing data without proper de-identification or anonymization may face negative consequences if identities are disclosed (Kaiser, 2009).

ASIST 2017,. Crystal City, VA | Oct. 27-Nov 1, 2017

Rarely studies on data reuse formally define the term data availability or unfold its attributes. Data availability can be understood as the state or property of data being handy and ready for use. However, availability and accessibility cannot be used interchangeably. Different aspects of data have to be considered to make data available and prior to make them ultimately accessible. Borrowing from the definition in the context of information technology and management (Techopedia, n.d.), we understand data availability as the process of ensuring that data are available to end-users in a timely fashion, to the degree that the data are readily and continuously usable by the necessary management procedures, infrastructure, technologies, and policies.

Along these lines, data availability involves different aspects to consider, and the data cannot be reliably available by individuals in a long term (Vines et al., 2010). Different stakeholders, such as regulatory parties, funding agencies, journal editors, and repository managers, have shown their concerns on the data's public availability (e.g., Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011; Guttmacher, Nabel, & Collins, 2009; Godlee, 2009; Sommer, 2010). This shows that various factors and conditions influence to data availability, not just researchers' attitude on data sharing. Others influential factors include ethical and legal concerns, federal regulations, journal policies, management procedures, and access restrictions of proprietary data.

While data availability is a prerequisite for potential reuse and a condition, which can be controlled by data providers, data reusability, that is; the degree to which data are capable of being reused in a given context and to fit a given purpose (Faniel & Jacobsen, 2010; Curty, 2015), involves an evaluation process from the reuser perspective. Not only data should conform to some quality attributes in order to be considered "reusable", but also fit to the purpose of interest. Thus, reusability can be only appraised from the potential reuser perspective, who will juxtapose their best judgment about the attributes of the available data to their reuse intention/purpose.

PANELIST AND TOPICS

Topic 1: Data availability for community data reusers

Ayoung Yoon is an assistant professor at the School of Informatics and Computing at Indiana University Indianapolis (IUPUI). She is also a RDA/US data share fellow for 2016-2017. Her research areas include data curation, data sharing/reuse, and open data practices. She is particularly interested in constructing effective cyberinfrastructure for data sharing and reuse through proper curation and preservation practices. Currently she is conducting several research projects focusing on communities' needs of data sharing, reuse, and curation.

Her presentation in this panel will focus on the issue of data availability for community data reusers. The potential for data to help address current societal problems (e.g., education, health, economic development, environment) is significant not only at the federal and state levels but also in smaller communities, neighborhoods, and individual lives. While the proposition for this potential is that data are and will be shared with and reused by and for communities at different levels, many data are not systematically or routinely shared for reuse with communities, particularly data collected by university and research institutes. From her recent research project that investigated community members' data practices in a small city in the Midwest, she will discuss the challenges associated with community members' data reuse focusing on data availability issue with their strategies to access and utilize data for their community development. Because community members' data needs and reuse practices are different from scientists, her findings will broaden the scope of discussion on data availability.

Topic 2: Before making data available: handling data involving vulnerable populations

Wei Jeng completed her Ph.D. studies at the School of Information Sciences (iSchool), University of Pittsburgh in 2017. Her dissertation study, entitled Qualitative data sharing in social sciences, investigates social scientists' data-sharing practices, especially those dealing with qualitative data. Given the increasing need for academic communities to manage an enormous amount of data, her long-term research goal is to provide insights for improving research infrastructure for scholars in all disciplines, particularly the social sciences, humanities, and related scholarly communities.

In this panel, Jeng will share her research findings and perspectives on two aspects:

- The tension between data shareability and confidentiality concerns regarding qualitative data. Jeng will report her findings regarding a group of social scientists' qualitative data sharing practices, especially for those who handle sensitive data or un-shareable data.
- The best practices for sharing data involving vulnerable populations. Jeng will share her proposed strategies for handling research data with minors, using a research project she is recently working as an example. The research project explores teens' data awareness and involves interview and observation data with young people between 6 and 17.

Topic 3: Data reusability for earth and environmental science data reusers

Angela P. Murillo completed her doctorate at the School of Information and Library Sciences at the University of North Carolina at Chapel Hill. Her dissertation research investigated facilitators and inhibitors of data sharing and data reuse in the context of earth and environmental sciences and newly developing infrastructures. Additionally, she has completed research in scientific data

management, metadata for data science, and science data curation. Her research area includes: scientific data management, scientific data cyberinfrastructure, and reuse and sharing.

In this panel, Murillo will present findings from a qualitative study, which identified facilitators, and inhibitors of data reuse among earth and environmental scientists. This study addresses the questions: how do scientists determine data reusability and what information inhibits or facilitates data reusability. Murillo will discuss what information about the data scientist's need to determine reusability, as well as what inhibits and/or facilitates data reusability. Murillo will describe how scientists evaluate data to determine reusability. Finally, she will provide guidance on best practices for ensuring reusability of shared data.

Topic 4: Amplifying data availability through enhanced publications

Renata Curty is an assistant professor in the Information Science Department at the State University of Londrina (Universidade Estadual de Londrina), in Southern Brazil. Her research relates to scholarly communication, data curation and research data reuse. Her current research project focuses on how digital platforms' technical features influence and shape data reuse intentions and behaviors among scientists.

In this panel Curty will present findings of an ongoing study on enhanced publications and their role on effective data reuse. She will start by introducing different approaches to make research data systematically available and accessible to a broader audience, taking into account different levels of ties research data may have with the original research and its outputs, as well as with related research. Then, she will discuss how each of these approaches is expected to facilitate, more or less, one or more data reuse purposes (i.e. aggregation, integration, meta-analysis, replication, repurposing, and synthesis). Later, the presentation will focus on preliminary results of an analysis of enhanced scientific journals, in quest of answering the following questions: What technical features and policy strategies enhanced journals usually adopt in order to make data available? What are the common challenges and limitations such journals face in order to promote data availability?

EXPECTED CONTRIBUTIONS

The panel aims to bring different perspectives on data shareability, availability, and reusability, also taking into account multiple data sharing-reuse contexts, within and beyond the traditional scientific fields (i.e., universities and other research institutions). Panelist will discuss issues relevant to data availability in a relation to shareability and reusability. The topics covered by the panelists will include accessibility, ethics and policy, technology, and discipline practices. The panel also expects attendees to speak out the experiences and challenges on handling the availability of their own data or the data they have reused. Our contributions will be two-fold. In terms of research, we aim to contribute to conceptualizing terms that are still poorly defined in the literature (i.e. data shareability, availability, and reusability), which will bring a more rich and diverse discussion to the current data sharing and reuse research. In terms of practice, we expect this panel to be valuable to librarians, policymakers, open data advocates, and data repository stakeholders to reflect on community data reusers, scientific data reusers, vulnerable populations, and infrastructure and policy requirements. This will serve as a foundation to build more sustainable data infrastructures and facilitate data openness.

PANEL'S STRUCTURE

The structure of the proposed panel is as follows:

I. Introduction (5 minutes):

Introduce the panel theme, core definitions, purposes, and objectives of the panel.

II. Presentations (30 minutes):

Each one of the panelists (Yoon, Jeng, Murillo, and Curty) will present their individual research projects and empirical findings, individually.

III. Q&A session (10 minutes):

We will host an open Q&A session about panel presentations and discussion.

Interactive Session (40 minutes):

The interactive session will be moderated by Curty and follow the Team-based Learning (TBL) approach, which consists of a three-step cycle: 1) preparation, 2) readiness assurance testing, and 3) application-focused exercise.

Preparation. The first step will be covered by presenters, where panelists will briefly introduce or recap the aspects related to data shareability, availability and reusability.

Readiness assurance testing. This stage will be approached in two parts. First, a set of questions based on real-life and controversial situations will be handled to each participant. These questions will be based on key themes and core concepts covered in the presentations. Second, participants will be grouped to discuss the questions and their answers, with the goal to collectively agree upon the best answer.

Application-focused exercise. In the third and final stage, the moderator will go through each of the questions and discuss them. Then, participants will propose, in groups, possible strategies and alternatives for solving or handling situations regarding data shareability, availability and reusability. The ultimate goal of this interactive session is to promote individual metacognition, as well as stimulate discussion, and engage participants on experience sharing about controversial topics and real-life situations practitioners and researchers may encounter on a daily basis.

Wrap-Up (5 minutes):

Debriefing and wrapping up the topic with all attendees.

REFERENCES

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public Availability of Published Research Data in High-Impact Journals. PLOS ONE, 6(9), e24357. https://doi.org/10.1371/journal.pone.0024357

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078. https://doi.org/10.1002/asi.22634

Collins 2009 Why data-sharing policies matter.Proc Natl Acad Sci U S A10616894

Curty, R. G. (2015). Beyond "data thrifting": An investigation of factors influencing research data. (Doctoral Dissertation). Retrieved from: http://surface.syr.edu/cgi/viewcontent.cgi?article=1266&context=etd

Fabic, M. S., Choi, Y., & Bird, S. (2012). A systematic review of Demographic and Health Surveys: data availability and utilization for research. Bulletin of the World Health Organization, 90(8), 604–612. https://doi.org/10.2471/BLT.11.095513

Faniel, I. M. & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work (CSCW), 19(3), 355-375.

Godlee F (2009) We want raw data, now. BMJ 339: b5405.F. Godlee2009We want raw data, now.BMJ339b5405

Guttmacher AE, Nabel EG, Collins FS (2009) Why data-sharing policies matter. Proc Natl Acad Sci U S A 106: 16894.AE GuttmacherEG NabelFS

Hey, A. J. G., Tansley, S. & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery: Microsoft Research Redmond, WA.

Kaiser, K. (2009). Protecting respondent confidentiality in qualitative research. Qualitative Health Research, 19(11), 1632-1641.

Sommer J (2010) The delay in sharing research data is costing lives. Nat Med 16: 744.J. Sommer2010The delay in sharing research data is costing lives.Nat Med16744

Techopedia. (n.d.) https://www.techopedia.com/definition/14678/data-availability

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age. Current Biology, 24(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

.

The columns on the last page should be of approximately equal length.