Making DA-RT a Reality

Thomas M. Carsey, *The University of North Carolina at Chapel Hill*

Calls for greater data access and research transparency have emerged on many fronts within professional social science. For example, the American Political Science Association (APSA) recently adopted new guidelines for data access and research transparency. APSA has also appointed the Data Access and Research Transparency (DA-RT) ad hoc committee to continue exploring these issues. DA-RT sponsored this symposium. In addition, funding agencies like the National Institutes for Health (NIH) and the National Science Foundation (NSF) have expanded requirements for data management and data distribution. These pressures present challenges to researchers, but they also present opportunities.

I define data access as the degree to which scholars who produce a research product (such as a published paper) make the data used in producing that research product available to others. Such access might be limited to just the subset of data used to produce the research product in question, or it might include the entire data set. Access might require signing a data use agreement, collaborating with the original research team who collected the data or waiting until an embargo period has cleared. Proprietary, privacy, or other issues may also limit or prevent data access. In short, there are many shades of gray in evaluating data access.

I define research transparency as the degree to which the process used by scholars to produce a research product is made clear and open to others. Data access is part of research transparency, but it also includes clear descriptions of and access to codebooks, decision rules for collecting, coding, and analyzing data, and, where appropriate, computer code. To borrow a metaphor, research transparency involves describing and providing access to both the ingredients used in producing a research product and the recipe for combining them.

Fostering greater data access and research transparency rests on a strong normative foundation. It increases the exchange of ideas, expands learning from individual studies, permits greater public scrutiny of results, and expands the impact of research in both academic and nonacademic settings. At a more fundamental level, the ultimate goal of research is to generate new knowledge and disseminate it widely. Scholars search for knowledge, but they must share their discoveries, subject them to the scholarly judgment of others, and permit others to build on them. Knowledge advances collectively, not individually, and this collective effort requires data access and research transparency. These principles increase the credibility of academic research and provide greater legitimacy to the research process. Ultimately, adherence to data access and research transparency principles increases the value of the research we produce.

The articles making up this symposium offer insights on a range of issues associated with data access and research transparency. In this article, I turn the conversation toward concrete actions individual scholars and scholarly organizations can take if they accept the premise that data access and research transparency are essential to the collective production of knowledge.

Scholars have limited time and resources, and they face a broad and growing set of demands, constraints, and pressures from a host of sources. In simple terms, scholars cannot do everything they might like, or that we might like of them—they must make choices that involve trade-offs of time, resources, and effort. Universities, academic professional associations, journal editors, and publishers also face time and resource constraints. As a result, concrete proposals to promote greater data access and research transparency must affect the cost-benefit analysis scholars, editors, publishers, and universities use when making choices about how to allocate their efforts. Although I noted earlier a strong normative rationale for greater data access and research transparency, in this article I present some ideas I hope will help to increase the benefits to scholars that come from providing greater data access and research transparency while lowering the costs of doing so.

Finally, while data access and research transparency touch all aspects of social science research, the articles in this symposium highlight important differences between traditional quantitative digital data, qualitative nondigital data, and the various mixes in between. Most of the issues I discuss here are relevant for all kinds of data, although their direct application might be easiest to envision in relation to research based on the application of some data reduction algorithm or similar procedure to digitized information for revealing particular patterns or attributes in the data.

EXPANDED VISIBILITY AND IMPACT

Many scholars advocate greater data access and research transparency because they want to promote research that replicates one or more existing studies. For some researchers, this raises the specter of facing public critique or refutation of their own research claims. From that perspective, some scholars might view providing greater access to their research data as a potential risk. This fear must be countered by recognizing that research that fosters replication is, by definition, influencing the larger research community. Even work that is eventually refuted provides at least part of the impetus for the subsequent critique. Common folklore asserts that the modal number of citations a refereed publication receives is zero. If that is anywhere close to reality, we should place value on research that stimulates replication studies even if its main conclusions are subsequently refuted. Making existing studies easier to replicate also makes them easier to build on, increasing the impact of the original study. Remember, knowledge advances collectively.[1]

doi:10.1017/S1049096513001753

https://doi.org/10.1017/S1049096513001753 Published online by Cambridge University Press

If replication is critical to the progress of knowledge, scholarly journals should be encouraged to publish replication studies, and departments and universities should give credit to researchers who produce such work. Doing this would encourage more scholars to engage in such activities. What constitutes a publishable replication study is an open question.[2] Space in leading journals may not be best used for replication studies that use the exact same data and exact same methods to reproduce the exact same results as a previously published paper. However, scholars should be encouraged to confirm or challenge findings using similar data and similar methods as a means of assessing the robustness of a published result. Efforts that uncover findings that cannot be replicated also deserve publication. Foreshadowing a point raised later in this article, the publication of replication papers might be effectively accomplished through publishing such materials online.

If replication is critical to the progress of knowledge, scholarly journals should be encouraged to publish replication studies, and departments and universities should give credit to researchers who produce such work.

We need more research to document the impact of promoting data access and research transparency principles. Some evidence indicates that articles that provide easy access to replication data and code are more influential than comparable articles that do not (Piwowar, Day, and Fridsma 2007). Similarly, the use and reuse of data sets enhances the visibility of the original project (Pienta, Alter, and Lyle 2010). However, we do not know if the adoption of more visible data access and research transparency policies increase the visibility, attractiveness, and impact of the journals that do so. We also need research that explores the metrics for measuring the impact of data access and research transparency policies. Universities are making greater use of citation counts to evaluate scholarly impact, and services like Google Scholar make gathering such information easier. If we want to encourage greater data sharing, citations to data sets should be part of a scholar's citation count. This also raises the question of whether all citations should count equally. If scholars cite Article A in a string of five or six citations to support an important, but secondary, point, but use Data Set B in their central analysis, one could argue that Data Set B has a greater scholarly impact than does Article A. Including citations to data sets in a scholar's citation count, and developing a metric to assess the impact of a data set shared by a scholar, would promote greater sharing of data.

DATA AS A RESEARCH PRODUCT

Hiring, promotion, and tenure committees at universities evaluate the actual and potential research productivity of scholars primarily in terms of the papers, books, and other publications they produce. Some also give weight to grants sought or received. The weight placed on each publication or grant is often affected by its perceived actual or potential influence. We socialize graduate students and junior faculty to respond to these metrics, and successful ones do. If the same value were attached to the production and distribution of data sets, scholars would face the same incentives to produce and share data sets that they currently face to produce and publish papers.

A recent development at NSF magnifies this point. Grant submissions to NSF have long required short biographical sketches for the principal investigator (PI) and all co-PIs that are meant to help reviewers evaluate the qualifications of a research team. In 2012, NSF changed one of the required biographical sketch subsection headings from "Publications" to "Products," explicitly identifying data sets as one of the types of research products scholars might list. Such products, including data sets, must be citable. This decision by NSF acknowledges the value of research data in its own right, but it also highlights the need to provide scholars with a method of data citation. Several efforts are underway to provide rules and methods for data citation—I am most familiar with the effort connected with the Dataverse Network (DVN) project.[3] A byproduct of these efforts, should they succeed, will be metrics for evaluating the impact of data production and data sharing.

Getting data sharing and data citation counts included in hiring, promotion, and tenure decisions will require leadership. Professional associations, like APSA, could make public statements regarding the value of doing so. Leading universities/departments might help establish a trend by adopting such strategies and encouraging other universities/departments to follow. Internal and external reviewers could highlight data contributions made by candidates for promotion and/or tenure in their evaluation letters. I do not support making a significant change in expectations for those already approaching tenure and promotion decisions, but this is one area where strong incentives to promote greater data access and research transparency could be established.

LINKING ARTICLE PUBLICATION WITH DATA PUBLICATION

Sharing data openly as a public good differs from publishing data as a product of research. The publishing industry is grappling with these issues via debates about open access to journals, open publishing, and the like. As the open access debate unfolds, it provides a good opportunity to consider linking the distribution of research findings and research data.

We have well-established workflows for authors, editors, reviewers, and publishers for the publication of academic articles.[4] These workflows change as technologies change, and some important differences exist across disciplines and journals, but for the most part, these workflows are well understood by nearly

https://doi.org/10.1017/S1049096513001753 Published online by Cambridge University Press

everyone involved. First-time authors and first-time editors face a learning curve, but publishers, former editors, and experienced colleagues are readily available for consultation and support. In short, we know how to publish papers.

We know much less about how to publish data sets. For example, Gherghina and Katsanidou (2013) report that only 19 out of 120 journals in political science and international relations have a published policy on data accessibility. Several efforts are under way to help foster better integration between paper publication and data publication, but a great deal of work remains to be done.[5] The goal is to make it easier for authors, editors, and publishers to publish data linked with research articles that use that data. Doing this requires addressing both technical and workflow issues.

Regarding workflow issues, several questions require answers. For example, should authors be expected to submit replication data and code as part of their initial submission, only when invited to revise and resubmit, or only when a paper is accepted for publication? Similarly, should reviewers be asked to review the data and code as part of evaluating a paper under consideration for publication? How much access to data should reviewers be granted prior to publication? Does access to data reduce anonymity of authors for journals using double-blind review? Should journal editors or publishers be expected to verify replication materials? What happens to replication materials if a submission is ultimately rejected for publication? Do editorial and production staff have the necessary expertise to evaluate and manage the review of replication materials and data publication?

On the technical side, the development of online article submission platforms for peer-reviewed journals has been a huge benefit for both authors and editors.[6] Similarly, tools like the DVN provide individual researchers with access to a web-based submission system for archiving and sharing research data. Both the DVN and commonly used journal submission systems allow for posting supplementary documents, which means, in theory, that journal submission systems could accept replication materials and the DVN could accept reprints of published papers.[7] However, no platform currently exists that integrates the features of both types of systems. Authors, editors, and publishers need a single interface that integrates article submission and publication with data submission and publication. This would make it easier for journals to adopt and monitor replication policies, easier for authors to comply with those policies, and easier for other researchers to find published data and articles related to their own work. Again, development projects are underway to resolve some of these issues, and organizations, like the Odum Institute, are involved in some of these efforts. Asking authors, editors, reviewers, and publishers to deal with replication materials necessarily increases their workloads. For these efforts to succeed, we must produce workflows and related tools that make this work as easy as possible for all involved.

Authors, editors, and publishers need a single interface that integrates article submission and publication with data submission and publication. This would make it easier for journals to adopt and monitor replication policies, easier for authors to comply with those policies, and easier for other researchers to find published data and articles related to their own work.

IMPLICATIONS FOR RESEARCH TRAINING

Lasting adoption of data access and research transparency principles requires that we integrate these values into our graduate training programs. To do it well, this training needs to start in the standard scope and methods course that most graduate programs offer in the first semester. Such courses often consider research ethics, competing notions of science, and various methods of doing research and collecting qualitative and quantitative data. These courses should incorporate the values of data access and research transparency from ethical and scientific perspectives, but they should also explore developing the associated pragmatic skills. The more that ideas associated with data access and research transparency are blended with discussions of developing research questions, formulating initial research plans, and developing research designs, the easier it will be for students to incorporate these principles in their work.

Beyond this initial course, most PhD programs in political science offer one or more quantitative methods courses, and many offer additional courses in both qualitative and quantitative methods. These courses vary in their focus on methodological theory versus application, but they often devote little or no time to broader issues of data management, data access, and the generation of transparent research replication materials. Whether the task involves proper documentation of the R code used to estimate a statistical model or proper documentation of field notes from a participant-observation study, these types of applied skills need to be folded into our methods training.

One growing trend within quantitative methods sequences is assigning students the task of replicating an existing published study, with sometimes the added element of providing some extension to that study as well. I have given such assignments regularly for the last decade. Unfortunately, one of the lessons students generally learn from this is how poorly existing research is documented and how difficult it is to replicate published results. Data is not made available, different versions of the data exist but are not clearly documented, decisions used to transform or recode variables are not well documented, and code used to conduct the actual analysis is

https://doi.org/10.1017/S1049096513001753 Published online by Cambridge University Press

not provided. Most of the original authors try to be helpful, but occasionally they are entirely nonresponsive. My students often come away from this assignment frustrated, shocked, and rather disappointed by what they see as practices that undermine the credibility of the research they read. However, I also hope these experiences instill a resolve in them to make sure that their own work meets higher standards for data access and research transparency.

Finally, the principles of data access and research transparency should play a central role in our substantive courses. The typical graduate seminar engages and evaluates existing literatures relative to the theoretical and/or substantive topic at hand. The transparency and replicability of the research process used by authors should be a normal part of how young scholars are trained to evaluate existing studies. Similarly, student seminar papers should be evaluated, in part, on the transparency of the research methods used.

Adapting how we train students to incorporate data access and research transparency principles from the outset has many potential benefits. Researchers who learn to think about these issues at the start of their careers, and who see value in doing so at the start of each research project, will be better able to produce research consistent with these principles. Furthermore, meeting these goals should be easier for scholars trained this way from the outset—in fact, it will hopefully feel "automatic" or "natural" for students and scholars who experience and internalize this type of training. The best way to have an enduring impact on how research is conducted in the future is to affect how researchers are trained in the present.

The best way to have an enduring impact on how research is conducted in the future is to affect how researchers are trained in the present.

There may be economies of scale that can be realized in this process. Common issues reach across social science disciplines, so departments might be able to collaborate. Numerous organizations with interdisciplinary training missions might also provide services. Some of the training necessary might be delivered online or through workshops rather than in traditional classes and seminars. In the end, however, the method of development and delivery is secondary to the more fundamental issue of deciding that data access and research transparency should be central elements of graduate education.

THE VALUE OF METADATA

Most discussions of data access and research transparency focus on the data itself. This focus has been amplified in recent years by both scholarly and public attention to the explosion of "Big Data." While raw data is essential, metadata is of equal importance. Metadata is best thought of as information about the data, or data about the data. Metadata provides meaning to data by describing it. Metadata includes information on who collected the data, when it was collected, where it was collected, how it was collected, and so forth.[8] Metadata also provides information on what each variable represents, and even what each value for each bit of data represents. Thus, while a variable in a data set might consist of a column of zeros and ones, that data lacks meaning until you know that it was collected by a particular polling firm in October of 2012 via a telephone survey of registered voters, that the variable itself captures each respondent's intention to vote, and that a value of one indicates that the respondent intends to vote while a value of zero indicates that the respondent does not intend to vote. Such metadata is often described as a codebook for a data set, but modern data archiving links data and metadata directly rather than collecting metadata in a separate codebook.

Successful implementation of data access and research transparency principles requires careful attention to the production, documentation, and sharing of metadata. Metadata allows researchers to communicate information about their data sets as well as learn about other data sets. Metadata is the currency of data archives—it allows scholars to share, search, and discover what data exists and determine whether it might be of use to them. Sometimes researchers might need to limit access to the data itself, but they can still allow for the public distribution of metadata. Data might include identifying or other sensitive information that cannot be made public. Scholars might also want a period of time to exploit their data before sharing it with others. Proprietary restrictions might be needed on some data. However, scholars may be able to share metadata in each of these instances that would provide for greater transparency for their research. Discovery of metadata by others might also create opportunities for collaboration or some other limited access to the data itself through a data use agreement. Thus, metadata is essential in its own right, but can also bridge the gap between complete data sharing and no transparency at all. Finally, training efforts regarding research transparency and data access should include explicit discussion of metadata from both a conceptual and applied perspective.

TURNING OBSTACLES INTO ADVANTAGES

NIH has required data sharing plans for grant proposals exceeding $500,000 since 2003. More recently, NSF released guidelines for meeting a new data management plan requirement. On May 9, 2013, President Obama issued an Executive Order, "Making Open and Machine Readable the New Default for Government Information," that requires the Office of Management and Budget to issue an Open Data Policy designed to make government data more widely available. These are just a few of the new policies pushing data openness and data sharing. Professional societies like the APSA are calling for more data sharing, and an increasing number of journals are adopting data sharing and replication policies. Such efforts certainly pose challenges in terms of privacy and data security,

https://doi.org/10.1017/S1049096513001753 Published online by Cambridge University Press

and meeting these challenges will require researchers to work differently.

Individual scholars, departments, universities, disciplines and their professional associations, journals, publishers, and data archives can either resist these changes or they can lead them. They can wait to see what happens, or they can shape what happens. Change always creates disruptions, but it also presents innovators with opportunities. Those who incorporate data access and research transparency principles into their training programs and their own research practices, and those who invest time and effort into leading these efforts, are those who stand to gain.

Successful incorporation of data access and research transparency principles into the practice of research will ultimately make the entire research process, often called the research lifecycle, more efficient and productive. Scholars who are trained to track what they do from the beginning will make more efficient use of their research time. Those who seize the opportunity to develop training methods and new tools that advance these principles will attract more research dollars and more scholarly attention to their work. Journals that build easy-to-use systems for sharing all of the products of research should enhance the visibility of the work they publish. Departments and universities that reward scholars who exemplify data access and research transparency principles will be better able to attract and retain researchers committed to those principles.

Successful incorporation of data access and research transparency principles into the practice of research will ultimately make the entire research process, often called the research lifecycle, more efficient and productive.

Expanding efforts to promote greater data access and research transparency does add some additional burdens to researchers. However, effective implementation of these principles in training, workflow, and technical solutions will minimize those burdens and may increase opportunities for success by promoting visibility. While these pressures are emerging from a number of sources, a comprehensive evaluation of how social scientists conduct research and train future researchers offers a chance for healthy adaptation and reform. Again, the key to success is increasing the benefits associated with providing greater data access and research transparency while lowering the costs of doing so.

LOOKING FORWARD

A significant obstacle to successful adoption of data access and research transparency practices is the uneasy sense that changes will continue to happen faster than we can adapt. Thus, I close with some thoughts about four other trends affecting political science research that might be connected to greater data access and research transparency.

1. Increased use of open-source research tools. Research in the social sciences—particularly quantitative research—has seen an explosion in the use of open-source tools. The increased use of R for doing statistical analysis and of $\LaTeX$ for writing research reports alone has changed how many scholars and journals operate.[9] Open-source tools for analyzing data make it easier to provide open access to the data itself because it reduces the need for access to commercial software to exploit the data.
2. Increased use of online resources to distribute scholarly products. Numerous journals now publish accepted articles online before they are published in print. Publishers increasingly sell electronic access to journals to both individual and institutional subscribers, and more journals are providing online-only options for their subscribers. Given the economies of online publishing and the growing demand for online access to research materials, this trend is likely to continue. The good news is expanding online publication of research articles should make it easier to connect those studies to digital archives of the data such articles use.
3. Continued tension between data security and data openness. The demand for access to data of all types is increasing across all sectors of society, not just in the academic research community. Data about the attitudes, opinions, and behaviors of people—the bread and butter of social science research—is increasingly available, but it is posing new challenges for the protection of privacy and any other potential harm that might befall research subjects. The ability to mine data from online activities, use of digital devices, and so forth exposes research subjects to greater privacy risks. These broader ethical issues increase the pressure to make data access and research transparency more central elements of our graduate training programs.
4. Big Data. The term "Big Data" has gone from novel to overused very quickly. While definitions of Big Data differ, there is no denying the explosion of data about social processes that has become available. This trend seems destined to continue, meaning that any long-term solution to data access and research transparency concerns must consider how well it scales up to massively large data sets and the associated complex analytic methods used to analyze such data.

Fortunately, tools and ongoing research projects are focusing on addressing these trends. For example, tools like Sweave and knitr allow researchers to embed R code directly into a $\LaTeX$ document so that when the document is compiled in $\LaTeX$, the R code is automatically executed.[10] This allows for research reports that include an analysis of the most recently available data to be generated on the fly. A byproduct of this

https://doi.org/10.1017/S1049096513001753 Published online by Cambridge University Press

approach is that it embeds the replication code necessary to produce the reported analysis directly in the document, though researchers would need access to the uncompiled $\LaTeX$ document to see it.

Another tool currently available is a package that can be installed in R called Shiny.[11] Shiny allows researchers to create simple web applications that present output from R functions online. The code to produce those results can also be shown. Thus, research papers could be presented with interactive tables and/or figures that appear online and include the code used to produce them. This provides another mechanism whereby researchers can directly share more than just the final table or figure they wish to include in their paper—they can present the code that accesses the data necessary to produce that table or figure as well. If the underlying data is updated, the table or figure can be automatically updated as well.

The DVN includes several features designed to facilitate data sharing, data citation, and research replication. The DVN has extensive capabilities to help users produce quality metadata. It also includes some built-in analysis tools, a means of providing a unique digital identifier as part of a citation to data sets, and even the capacity to produce subsets of data and the corresponding code associated with any analysis a researcher might run within the DVN.

As director of the Odum Institute, and through involvement with APSA's DA-RT ad hoc committee, fortunately I have been engaged in some of the efforts directed at promoting greater data access and research transparency. Although I see many challenges, I strongly support the normative, ethical, and scientific values associated with greater research transparency. Still, the success of efforts designed to promote greater data access and research transparency will depend on whether they lower the costs and raise the benefits of adopting data access and research transparency principles, and whether we adapt our training programs so that these principles drive the establishment of updated norms about the proper conduct of research and dissemination of knowledge.

ACKNOWLEDGMENTS

An earlier version of this paper was prepared for presentation at the DA-RT conference held at Syracuse University, June 22–23, 2013. I would like to thank the conference attendees for their feedback, and especially Colin Elman and Auther Lupia for their comments and suggestions. I would also like to thank Odum Institute staff members Thu-Mai Christian, Jonathan Crabtree, and Sophia Lafferty-Hess for their helpful feedback. Of course, final responsibility for the contents of the article rests with me. ■

NOTES

[1] We can make this easier by encouraging scholars to frame their own research in terms of how it builds on existing studies rather than in terms of the problems with existing work.

[2] The essay by John Ishiyama in this symposium devotes careful attention to this issue.

[3] Those interested in learning more about the DVN should start here: http://thedata.org/.

[4] The same applies for books and edited volumes as well.

[5] See Vision (2010), the NERC Science Information Strategy Data Citation and Publication Project (http://ijdc.net/index.php/ijdc/article/view/208) and the DVN Integration project (http://projects. iq.harvard.edu/ojs-dvn) for examples. The Odum Institute also has a pilot project underway, supported in part by an ICPSR/Sloan Foundation Challenge Grant, to develop recommendations on integrating the article and data publication workflows.

[6] As one who has served as a journal editor both with and without access to an online system, I can attest to this claim.

[7] It is more likely that each would simply use links to the other.

[8] Archives have developed a number of conventions and standards for the production of metadata. The DVN, for example, permits the generation of metadata that follows DDI, Dublin Core, FGDC and MARC standards.

[9] The DVN is also open-source software.

[10] Interested readers should consult the Sweave website: http://www.stat.uni-muenchen.de/~leisch/Sweave/ and the knitr website: http://yihui.name/knitr/.

[11] Interested readers should start here: http://www.rstudio.com/shiny/.

REFERENCES

Gherghina, Sergiu, and Alexia Katsanidou. 2013. "Data Availability in Political Science Journals." *European Political Science* March: doi:10.1057/eps.2013.8.

Pienta, Amy M., George C. Alter, and Jared A. Lyle. 2010. "The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data." http://hdl.handle.net/2027.42/78307.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. 2007. "Sharing Detailed Research Data is Associated with Increased Citation Rate." *PLoS ONE* 2 (3): doi:10.1371/journal.pone.0000308.

Vision, Todd J. 2010. "Open Data and The Social Contract of Scientific Publishing." *Bio-Science* 60 (5): 330–331.

https://doi.org/10.1017/S1049096513001753 Published online by Cambridge University Press