Implementing a Data Citation Workflow within the *State Politics and Policy Quarterly (SPPQ)* Journal

Principal Investigator: Thomas Carsey
Project Manager: Sophia Lafferty-Hess
The University of North Carolina at Chapel Hill
H. W. Odum Institute for Research in Social Science
Davis Library, CB #3355
Chapel Hill, NC 27599
919.843.5990
carsey@unc.edu

Introduction

Public funding agencies, scholarly journals, and open data advocates have pushed strongly over the past decade for increased archiving and sharing of research data. A recent Office of Science and Technology Policy Memo (2013) highlights the importance of providing open access to datasets and scholarly publications as a method of promoting innovation, accountability, transparency, and efficiency. Today, while well-established practices are in place to handle the review, publication, dissemination, and citation of research findings through scholarly journals, a system to integrate the review, publication, dissemination and citation of research data within the scholarly publication structure still requires attention.

Over the past year through support from a challenge grant from the Alfred P. Sloan Foundation and the Inter-university Consortium for Political and Social Research (ICPSR), the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill has undertaken a project to implement a prototype data citation workflow within the current *State Politics & Policy Quarterly (SPPQ)* journal publication workflow. This project developed a human-driven workflow to archive, share, and link underlying replication data to their associated scholarly publications. The workflow developed through this project identified the roles and responsibilities of various stakeholders (i.e., journal staff, data repository staff, publishers, and authors) and examined key challenges involved in linking published articles and data underlying the articles.

The Principal Investigator (PI) of this project, Thomas M. Carsey, is currently the director of the Odum Institute and the editor of *SPPQ,* which is published by Sage, a premier commercial publisher. Carsey led the project team and worked closely with the staff of the Odum Institute Data Archive to develop a workflow that would efficiently support both the archiving and sharing of replication data, and the creation of a stable link between data and scholarly publications. *SPPQ* uses the Dataverse Network (DVN) virtual archive platform to archive and make accessible replication data for their articles. The DVN provides a robust archival infrastructure for *SPPQ* researchers' data that supports discovery and access to research data. The DVN has an infrastructure in place that mints persistent identifiers for datasets and automatically generates data citations.

This report discusses the following: (1) key aspects of the prototype workflow, (2) lessons learned, (3) automation suggestions, and (4) next steps. Much of the content will cover lessons learned, which include the importance of relationships, the need to attend to small details when developing a robust workflow, the need to decrease the time spent on data archiving and sharing by editors and authors whenever possible, and the challenges of designing a workflow that uses multiple systems. These lessons learned will help inform future work on integrating data citations into scholarly publication systems and stimulate productive conversations with various stakeholders including data archivists, editors, publishers, and researchers.

Prototype Workflow

Prior to developing and implementing the prototype workflow, the project team analyzed the current state of the *SPPQ* editorial workflow, which uses the popular ScholarOne/Manuscript Central online submission system. The project team then identified and diagramed potential workflows that could accommodate data archiving and linkages, and assessed the strengths and weaknesses of specific features. One key question centered on the timing of data submission (i.e., prior to manuscript acceptance or after manuscript acceptance). The project team also considered the assignment of roles and responsibilities of stakeholders. For instance, should the editorial staff, archive staff, or author create the catalog record in the Dataverse Network? Who should verify the data once submitted? Who will send the data citation to Sage after the data is deposited and verified?

Submission of Data: The project team determined that authors would submit data after the manuscript was accepted for publication because requiring data submission during the review process was currently not part of the *SPPQ* data replication policy. Furthermore, this introduced concerns about the disposal or retention of data for manuscripts not accepted for publication. It also raised concerns about whether staff time and resources should be devoted to managing replication data for submissions that may not ultimately be accepted.

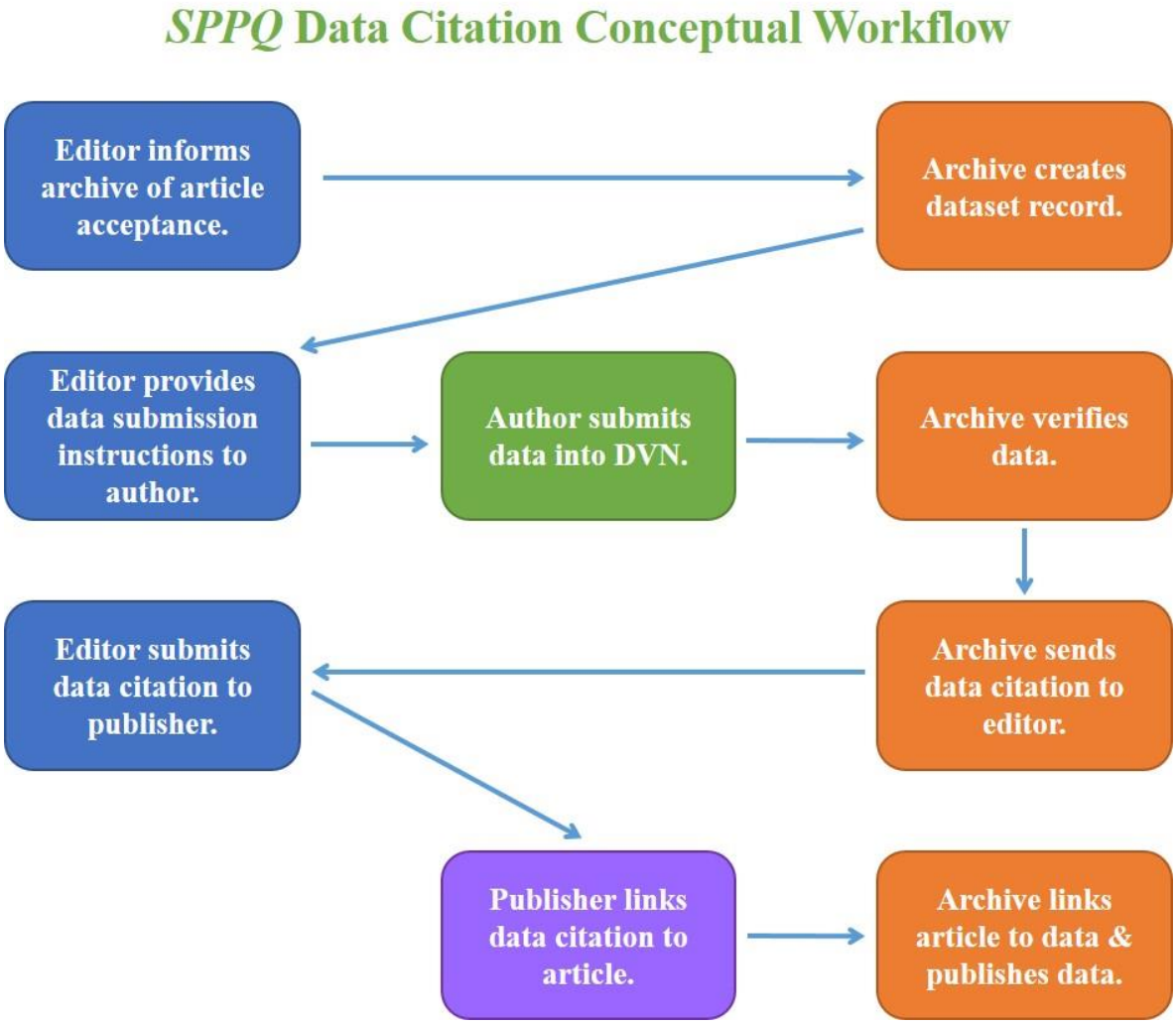
The question of when to receive replication data and other materials also raised conceptual questions about the role of data in the review process of an article. A recent Data Pub blog post by John Kratz (2014) discussed various ideas that surround the issue of data validation and review including the suggestion that data be reviewed in conjunction with the manuscript during the peer review process; however, as Kratz notes, this is far from a universal opinion as others see data validation as a separate process from peer review.

Ultimately, each journal must identify in their replication policy when the submission of replication materials will occur. Given the current heterogeneity regarding data review as part of peer review, archiving and data linking workflows will need to be flexible in order to support the various policies and procedures of different journals.

Roles and Responsibilities: In regard to roles and responsibilities, the project team determined that the Odum Institute Data Archive staff would be responsible for creating the catalog record, generating descriptive metadata, verifying the data, transmitting the data citation to the editor, inputting the DOI for the article once published, and publishing the data (see the conceptual workflow below and the Appendix for the full workflow diagram). Necessary information such as the link to the catalog record and user login information would be sent to the *SPPQ* editor through a series of email communications (see the Appendix for email templates).

The archive staff was given a key role in the workflow for the following reasons: (1) the archive staff has significant expertise with the DVN and data archiving best practices; (2) the archive staff is more static with less turnover than editorial staff, thereby providing more consistency; and (3) the archive staff could potentially be more efficient since often editors are academics

who have a variety of commitments and responsibilities and may not have additional time to spend creating catalog records and verifying data.



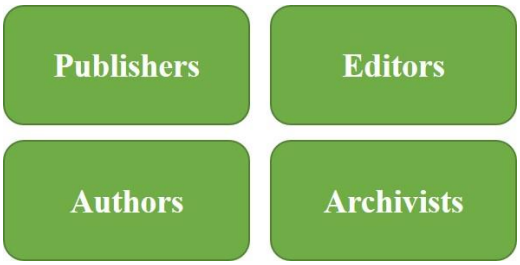
Data Verification: The data verification portion of the workflow is still currently in the pilot stages, and a reassessment of its viability will be performed as the project progresses. Verification and validation of data can be complex and highly variable depending on data type and discipline. The development of the prototype workflow enabled the project team to consider a variety of issues that surround verification including: (1) the role of subject expertise, (2) the importance of specifying procedures to ensure responsibilities and expectations are made explicit among stakeholders, and (3) the impact of varying journal policies (i.e., what type of data is required to be shared—replication data, source data, all original data?, what type of supporting materials should be shared—computer code, meta-data, etc.?).

For the *SPPQ* workflow, it was determined that verification would involve performing a cursory check of data files, ensuring proper documentation is included, and checking that any accompanying code properly replicates the findings in the paper.

Sustainability: A drawback of the developed workflow is that it will require significant archival resources including staff resources to assign metadata and verify data, as well as technical resources to support the preservation infrastructure. Moving forward, journals will need to consider providing funding to archives to fulfill this role.

Lessons Learned

The Importance of Relationships



To effectively preserve and make research data underlying articles available requires collaboration between publishers, journal editors, authors, and archivists. These relationships are the foundation for creating a robust data-sharing infrastructure for the future. They also can add value to current practices through knowledge sharing and can result in the expansion of services, which benefits a broad array of stakeholders.

One of the key outcomes of this project was a strengthening of the relationship between Sage Publications, the Odum Institute Data Archive, and *SPPQ*. This project presented a unique opportunity to discuss data sharing conceptually with Sage representatives and to discuss Sage's practical role in creating a stable link between articles and research data. Some specific lessons learned from the discussions and interactions with Sage are discussed below.

Making Connections with the "Right" People: A critical part of adding data citations within the Sage system involved first identifying and then communicating with specific individuals who had access to add the data citation to the Sage website. Making these connections and building lines of communication created an important infrastructure for continuing the process of linking data and publications including adding data citations for previous *SPPQ* articles. Establishing these lines of communication also allowed the project team to discuss the easiest mechanisms from the publisher's perspective for transferring data citations for new manuscript submissions.

Understanding the Various Perspectives of Stakeholders: Archivists, commercial publishers, editors, and researchers all have varying perspectives, assumptions, and beliefs regarding data archiving and sharing. For instance, data archivists may incorrectly assume that all stakeholders understand data management and sharing concepts, terms, and best practices. In the end, collaborations present an opportunity to learn and grow from interactions between diverse groups.

As an example, one step in this project involved working with Sage staff to create a link between previously published articles and their underlying replication data stored in the *SPPQ* DVN.[1] The archive staff provided a complete data citation to Sage staff and had anticipated the entire data

[1] To see an example of a completed link for a previously published *SPPQ* article, click here.

citation would be placed on the newly created "Replication Materials" page with the persistent identifier hyperlinked to the data's DVN location. On the first trial run, Sage only created a hyperlink entitled "Replication Materials" that navigated to the data stored in the DVN, but did not include the data citation. After more completely explaining the importance of the data citation and explicitly asking for it to be included, Sage staff were happy to place the entire data citation on the "Replication Materials" page. This exchange exhibited how archive staff had not initially clearly communicated the importance of the data citation for scholarly acknowledgement and had not thoroughly understood that, from a publisher's perspective, the inherent value of a data citation may not be readily apparent.

These relationships also create an opportunity to advocate for data sharing best practices and provide rationale for their importance in the world of scholarly publication. They also allow for discussions between publishers, editors, and archivists regarding data replication policies. This helps each stakeholder better understand the perspectives held by others, thereby facilitating the discovery of common goals and objectives.

Conversations between the various stakeholders in relation to policies, procedures, and best practices is an essential step in the creation of a research infrastructure that supports the publication of all forms of scholarly communication as well as a system that makes publications and research data available concurrently in an efficient and effective fashion. To continue building these relationships will require an appreciation of all parties' perspectives and a willingness to engage in open and forthright communication.

The Devil is in the Details

Developing a workflow is both a conceptual and practical process. After defining the ideal prototype workflow, one may consider the "hard" work done. However, after the primary decisions are made, a myriad of smaller decisions that may not be explicitly addressed in the planning phase must also be made during implementation of the plan. These smaller decisions are often essential for creating a robust and scalable workflow. While conceptually the project team addressed the "big" decisions prior to implementation, one of the challenges was addressing the smaller but essential details as they arose. Two specific examples are discussed below.

User Names: The creation of DVN accounts for authors to submit data, while seemingly simple, became more complex upon implementation. The project team had to consider questions such as:

- If the paper has multiple authors, should the archive staff create accounts for every author or only the listed contact author for the article?
- What user name should be created? Is there a standard we should be considering (i.e., ORCID, etc.)?
- What if the author already has a DVN account? Should we attempt to discover this information prior to making a new account?

The project team in many ways took a "More Product Less Process (MPLP)" approach to address these questions and decided to make an account for the contact author using their email address as the user name (Greene & Meissner, 2005). However, these questions also point to some of the challenges of using a human-driven workflow that requires authors to interface with the DVN to upload data.

Data Citations: Another question that arose when doing the first trial of adding a data citation to the Sage online platform involved the formatting of the data citation. The importance of creating interoperable and persistent data citations has recently been expressed by the Joint Declaration of Data Citation Principles, which states that "data citation…is good research practice and is part of the scholarly ecosystem of supporting data reuse" (Preamble, 2013). The DVN automatically generates a standardized data citation that includes all the necessary elements for discovery and attribution including author, date, title, distributor, a persistent identifier (a registered handle), a Universal Numeric Fingerprint (UNF), and versioning information (Altman & King, 2007).[2]

The project team considered whether the DVN data citation should be further formatted to more closely coincide with the American Political Science Association (ASPA) Style Manual used by *SPPQ*. The project team decided to maintain the original data citation generated by the DVN to lessen the workload on archive staff, and because it already contains the necessary elements to support proper scholarly acknowledgment. However, formatting of data citations in relation to discipline-specific style manuals is an issue that will need to be addressed as archivists and publishers move forward with integrating data citations into scholarly publishing workflows. This also presents an opportunity to develop new tools to automatically re-format generated data citations to comply with style manuals across disciplines while also maintaining the essential elements to support reuse and interoperability.

Time is a Precious Commodity

One of the driving factors of the proposed workflow was to lessen the time burden of sharing data and creating the data citation link for both the authors and editor. While editors and authors are required to play important roles in the workflow, wherever possible the archive has been assigned the bulk of the activities.

The rationale for this decision came from an assessment that adding significantly more work onto the editor may be unrealistic since often editors of journals are busy academics who have many different commitments and responsibilities. In addition, researchers often cite a lack of time and resources as a reason for not making data available electronically (Tenopir et al., 2011; Swan & Brown, 2008). Finally, editors and authors generally lack the expertise possessed by

[2] An example data citation generated from the DVN:
Carsey, Thomas; Harden, Jeffrey, 2010, "Replication data for: New Measures of Partisanship, Ideology, and Policy Mood in the American States", http://hdl.handle.net/1902.29/11598
UNF:5:EKgHvTNfkkS86dNzABIhNw== Odum Institute for Research in Social Science [Distributor] V3 [Version]

archivists. In the developed workflow, the archivist's role was partially seen as facilitating data sharing by lessening the burden on other stakeholders.

The workflow attempts to lessen these stakeholders' time-commitment through the following mechanisms:

- Archive staff create the catalog record for the replication data
- Archive staff input descriptive metadata
- Archive staff send all pertinent information to the editor using standardized email templates
- Archive staff are available to answer data deposit questions from authors as needed
- Archive staff verify that data submitted are well-formed and usable
- Archive staff provide the editor with a formal data citation generated by the DVN to be provided to the publisher
- Archive staff add the persistent identifier (DOI) for the article to the data catalog record once the article is published
- Archive staff publish the data

Despite these time-saving mechanisms and the creation of email templates for use in the human-driven workflow, editors will still have to spend time sending and responding to email communications and addressing verification issues, and authors will have to spend time preparing data for sharing and uploading data into the DVN. To further simplify this workflow will require automation, which could include automation of email communications and the removal of the need for authors to interface with the DVN system.

The Challenge of Working with Multiple Systems

The developed workflow requires the use of multiple different systems including ScholarOne (manuscript submission system), the DVN (data archive preservation and access system), and High Wire (Sage's online publishing system). Because of this, the project team first had to educate one another about the basics of the systems and their varying affordances. The project team then attempted to devise a workflow that took advantage of the each system's affordances whenever possible.

However, one key challenge and less-than-ideal aspect of the current workflow is that authors are required to interface with more than one system. While much of the archival work within the DVN is performed by the archive staff, authors are asked to upload their replication data into the DVN and adjust metadata as needed. This is far from ideal; the project team would have preferred that the authors only have to interact with the ScholarOne system; however, to do this would have required the development of tools that were outside the scope of this project. In the automation section below, suggestions for the development of a more integrated system is discussed.

Automation Suggestions

The workflow developed through this project allowed the project team to explore various options for further automation and improvement. Other groups working on integrating research data sharing infrastructures and journal publication workflows have been developing automated tools for simplifying and streamlining this process. Specifically, the PKP – Dataverse Integration Project, a collaboration between Stanford University's Public Knowledge Project (PKP) and The Institute for Quantitative Social Science (IQSS) at Harvard University, has developed a DVN Application Programming Interface (API) that can be used in conjunction with a plugin created for the Open Journal Systems (OJS) to allow authors to upload data in OJS while automatically depositing the data into the DVN (The Institute for Quantitative Social Science, 2012).

The project team suggests the development of a new Sage plugin that uses the DVN API to streamline the data deposit and data citation process for Sage journals within the ScholarOne system. This plugin would accomplish the following:

- Support data sharing and preservation through a robust system developed for a premier commercial publisher
- Support a broad array of current journal data sharing policies
- Remove the need for authors to interface with the DVN directly
- Support the automatic generation and delivery of standardized email communications
- Promote data access by simplifying the mechanisms for sharing research data
- Ensure that data is stored in an archive that supports long-term preservation of digital content
- Ensure consistent and stable data citations are generated, applied, and linked to scholarly publications

These suggestions are not exhaustive, and, prior to beginning development, more research would need to be performed to examine how exactly the systems would be integrated and to determine the specific functional elements of the plugin. It is also suggested that discussions and a knowledge-sharing initiative with the PKP – Dataverse Integration Project team and other data repositories involved with integrating archiving and journal submission workflows be initiated to support collaboration and cross-disciplinary learning. Another potential collaborator is Dryad, which already has in place an integrated journal submission system for data archiving and linking for journals in evolutionary biology and ecology (Dryad, 2013). Through collaboration and not "reinventing the wheel," the data archiving and sharing community can develop tools that can be used by open access and commercial publishers, journals across disciplines, and in varying systems.

Next Steps

The prototype workflow for *SPPQ* was undertaken as a pilot project to help inform the integration of data citation workflows into journal editorial systems. The primary goal of this project was to extract key lessons from the development of the workflow and to present possible ways to improve the workflow through automation. Although funding for this project has concluded, data archiving and creating persistent links between publications and underlying research data is still of utmost importance to *SPPQ* and the Odum Institute Data Archive. Carsey finishes his appointment as the editor of *SPPQ* at the beginning of June 2014; however, the incoming editors have expressed interest in continuing the project. The Odum Institute Data Archive staff also believes many important lessons can still be learned through continuing this project including:

- Further examining the issue of data verification and the role of the archive in verifying/validating data
- Continuing the process of knowledge-sharing and relationship building with publishers
- Continuing to build understanding of the culture surrounding scholarly communication
- Continuing to examine the challenges that arise when helping *SPPQ* editors and authors make underlying data available in conjunction with publications

Conclusion

The lessons learned throughout this project will help to inform future initiatives and can be used to stimulate further conversations about the challenges and opportunities that surround integrating data archiving, sharing, and citation in the process of scholarly publication. The relationships that were fostered during the current project will continue to grow and expand. These relationships present an opportunity to help build the future of scholarly communication that Van de Sompel et al. (2004) called for. However, these relationships will require communication, respect, and an understanding of the various perspectives, cultures, and practices held by different stakeholders.

The project team also learned that all good workflows attend to both the large conceptual decisions as well as the small details. Implementing a data citation workflow in *SPPQ* also exhibited that editors and authors often have many commitments and responsibilities, meaning that effective implementation of a data sharing workflow likely requires enlisting archival experts to undertake as many responsibilities as possible. Finally, the ideal workflow would lessen the need for stakeholders to use multiple systems. This ideal system would be automated and would allow authors to submit data within ScholarOne and facilitate the automatic deposit of data in a secure and trustworthy data repository.

Bibliography

Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, *13*(3/4).

Data Citation Synthesis Group. (2013). Joint declaration of data citation principles. Retrieved from https://www.force11.org/datacitation

Dryad. (2013). Submission integration: Overview. Retrieved from http://wiki.datadryad.org/Submission_Integration:_Overview

Greene, M. A., & Meissner, D. (2005). More product, less process: revamping traditional archival processing. *The American Archivist*, *68*(2), 208–263.

Kratz, J. (2014, May 8). Fifteen ideas about data validation (and peer review). Retrieved from http://datapub.cdlib.org/2014/05/08/fifteen-ideas-about-data-validation-and-peer-review/

Office of Science and Technology Policy. (2013). *Increasing access to the results of federally funded scientific research memorandum*. Washington, D.C.: Executive Office of the President. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., Frame, M., & Neylon, C. (2011). Data sharing by scientists: practices and perceptions. *PLoS ONE*, *6*(6), e21101. doi:10.1371/journal.pone.0021101

The Institute for Quantitative Social Science. (2014). FAQ on the PKP-Dataverse Integration Project. Retrieved from http://projects.iq.harvard.edu/ojs-dvn/book/faq-ojs-dataverse-integration-project

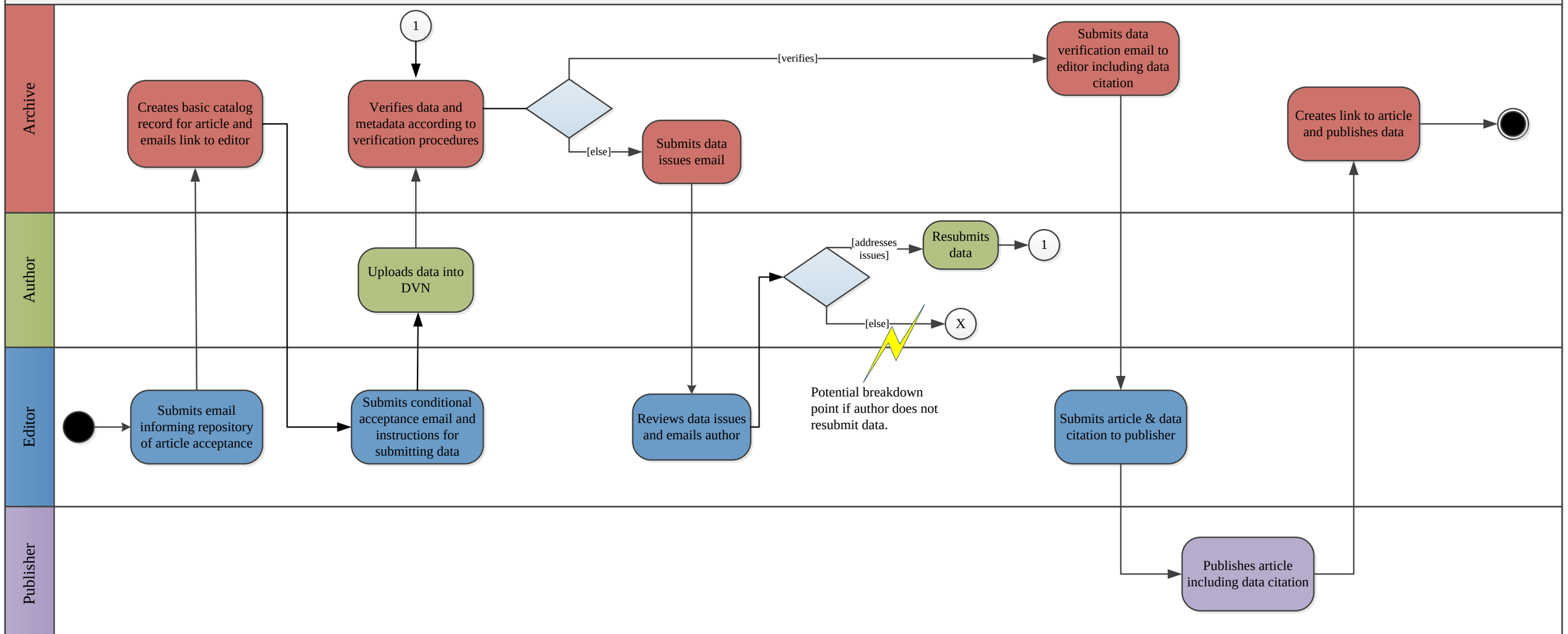
Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of research data outputs*. Research Information Network.

Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking scholarly communication. *D-Lib Magazine*, *10*(9). doi:10.1045/september2004-vandesompel

Appendices

SPPQ Data Citation Workflow

Data is requested AFTER manuscript is accepted for publication



Email Templates

Manuscript Acceptance Email to Archive from Editor

Subject: [Journal Name] Manuscript Acceptance ID # [xxxxx]

Dear XXXX,

The following manuscript has been conditionally approved for publication*:*

Journal:
Article title:
Author(s):
Contact author:
Manuscript ID:
Abstract:

Please create the catalog record for this manuscript, an email notifying the author of the conditional acceptance of their manuscript and an invitation to submit their data will be submitted upon receiving a link to the catalog record.

Thanks,
XXXXX

Manuscript Catalog Link Email to Editor from Archive

Subject: [Journal Name] DVN Catalog Record Link ID # [xxxxx]

Dear XXXX,

The conditional manuscript record has been created for:

Journal:
Article title:
Author(s):
Manuscript ID:

Below is the catalog record link, username, and login information for the author:

Catalog record link:
Username: authors@email
Password:

After the data has been deposited, we will email you concerning any data verification issues.

Thanks,
Odum Institute Data Archive staff

Data Submission Invitation to Author from Editor

Dear XXXXX,

Congratulations your manuscript, XXXXX, has been accepted for publication within the State Politics and Policy Quarterly!

SPPQ currently has in place a data replication policy that asks all authors to make replication data publicly available. We would like to encourage you to deposit the data underlying this article in the SPPQ Dataverse housed within the Odum Institute Dataverse Network. The Odum Institute Data Archive is a well-established and trusted archive in the social science field. As a member of the Data Preservation Alliance for the Social Sciences (Data-PASS) and the Library of Congress National Digital Stewardship Alliance (NDSA), Odum provides a strong archival and data distribution infrastructure for research data.

A Dataverse Network user account and password has been created for you. Upon initial login to the Dataverse, please change your password for security purposes. Below is the account information you will need to access the *SPPQ* Dataverse:

Catalog record link:
Username: authors@email
Password:

After your data is deposited, the system will create a unique data citation, which will be included alongside the published article.

If you have any questions about depositing your data please contact the Odum Institute Data Archive at odumarchive@unc.edu.

Thanks,
XXXXX

Data Issues Email to Editor from Archive

Subject: [Journal name] DVN Data Issues ID # [xxxxx]

Dear XXXXX,

After completing the verification process, issues were found with the data or documentation for the following manuscript:

Journal:
Article title:
Author(s):
Manuscript ID:
Catalog record link:

We encourage you to review the issues described below (the data can be found using the catalog record link) and contact the author with further instructions concerning archiving their data.

Issues:

If you have any questions please let us know.

Thanks,
Odum Institute Data Archive staff

Data Verification Email to Editor from Archive

Subject: [Journal name] DVN Data Verification ID # [xxxxx]

Dear XXXX,

The following manuscript's data has been verified and persistent identifiers have been assigned:

Journal:
Article title:
Author(s):
Manuscript ID:
Catalog record link:

The manuscript data citation is as follows:

> Carsey, Thomas; Harden, Jeffrey, 2010, "Replication data for: New Measures of Partisanship, Ideology, and Policy Mood in the American States", http://hdl.handle.net/1902.29/11598 UNF:5:EKgHvTNfkkS86dNzABIhNw== Odum Institute for Research in Social Science [Distributor] V3 [Version]

Please provide this citation to the author and publisher to link the published article to the supporting data.

Thanks,
Odum Institute Data Archive staff