IJDC | *General Article*

Building a Bridge Between Journal Articles and Research Data: The PKP-Dataverse Integration Project

Eleni Castro
Harvard University

Alex Garnett
Simon Fraser University

Abstract

A growing number of funding agencies and international scholarly organizations are requesting that research data be made more openly available to help validate and advance scientific research. Thus, this is an opportune moment for research data repositories to partner with journal editors and publishers in order to simplify and improve data curation and publishing practices. One practical example of this type of cooperation is currently being facilitated by a two year (2012-2014) one million dollar Sloan Foundation grant, integrating two well-established open source systems: the Public Knowledge Project's (PKP) Open Journal Systems (OJS), developed by Stanford University and Simon Fraser University; and Harvard University's Dataverse Network web application, developed by the Institute for Quantitative Social Science (IQSS). To help make this interoperability possible, an OJS Dataverse plugin and Data Deposit API are being developed, which together will allow authors to submit their articles and datasets through an existing journal management interface, while the underlying data are seamlessly deposited into a research data repository, such as the Harvard Dataverse. This practice paper will provide an overview of the project, and a brief exploration of some of the specific challenges to and advantages of this integration.

Received 13 January 2014 | *Accepted* 26 February 2014

Correspondence should be addressed to Eleni Castro, 1737 Cambridge Street, K318, Cambridge, MA 02138 USA. Email: ecastro@fas.harvard.edu

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Why Connect Published Work to Research Data?

As data sharing technology, data management practices, and policies from funding agencies and scholarly organizations have evolved over the last few years (Wipperman, 2013), a growing number of academic journals are joining the effort to publish and disseminate research data associated with their published articles. A recent article by Gherghina and Katsanidou (2013) stated that "any moves towards data sharing are dependent upon the cooperation of journals." However, this cannot be accomplished by journals alone; "[t]he most immediate of these obstacles is the lack of a consolidated infrastructure for the easy sharing of data" (Vlaeminck, 2013). Vines et al., in a recent study on the value of preserving research data, assert that authors who do a good job archiving their data in a public repository produce papers that "are more valuable to the scientific community, and to the publication they appear in" (2013). By partnering with research data repositories to develop the technology and best practices necessary to make research data readily available (with published results) and accessible for the long term, subsequent researchers will be able more easily to replicate and reuse research data, which helps the overall validation and advancement of science.

The PKP-Dataverse Integration Project

Through a two year Sloan Foundation grant (2012-2014), the Public Knowledge Project[1] (PKP)-Dataverse Integration Project is working on connecting journal articles with their underlying research data. This is being done by integrating two well-established open source systems: PKP's Open Journal Systems (OJS) (Willinsky, 2005; MacGregor, Stranack, & Willinsky, 2014), developed by Stanford University and Simon Fraser University; and Harvard University's Dataverse Network web application (King, 2007; Crosas, 2013), developed by the Institute for Quantitative Social Science (IQSS) (King, 2014). This integration will contribute to an increase in the replication and reuse of research outputs by improving the infrastructure for, practice of, and incentives related to data publication and citation. Technically, this project will result in the development and dissemination of an OJS Dataverse plugin and Data Deposit API; allowing authors to submit their article and datasets seamlessly through a journal management system, while the underlying data are automatically deposited into research data repositories like the Harvard Dataverse[2]. This plugin will complement and/or replace OJS' current 'supplementary files' option, and will also provide access to the Dataverse Network's data preservation, citation and analysis tools at the journal/article level. This work coincides with the increased use of Dataverse and other research data repository platforms, as more researchers, funding bodies and universities advocate for the continued availability of research data.

1 Public Knowledge Project: http://pkp.sfu.ca
2 Harvard Dataverse: http://thedata.harvard.edu

 doi:10.2218/ijdc.v9i1.311

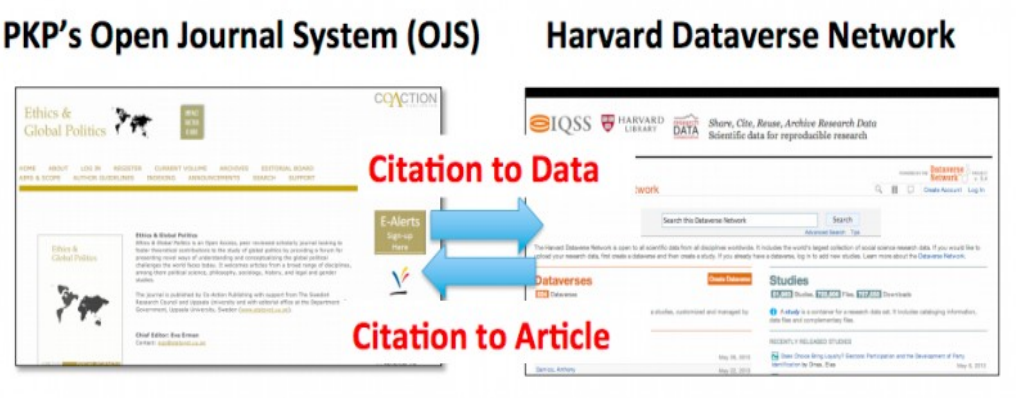

Figure 1. High-level diagram of the integration between an OJS journal and Dataverse.

PKP's Open Journal Systems (OJS)

OJS, developed by PKP in a partnership with Stanford University and Simon Fraser University, is an open source software platform for the management and publishing of peer-reviewed journals. This system assists with every stage of the refereed publishing process, from submissions through to online publication and indexing. Through this system, PKP seeks to encourage open access publishing and improve both the scholarly and public quality of refereed research. As of December 2012, there were over 5,000 active[3] journals utilizing OJS. According to the OJS website[4], it is estimated that there are over 1.5 million articles published in OJS journals.

Dataverse Network

The Dataverse Network, developed at Harvard University's IQSS, is an open source web application for sharing, citing, analyzing and preserving research data. It facilitates making data available to others through a persistent data citation (with a DOI), and by allowing researchers to more easily replicate and reuse others' work. Researchers, data authors, publishers, data distributors and affiliated institutions all receive appropriate credit. A Dataverse can consist of multiple dataverses and/or Datasets[5], where each one is a virtual archive with the possibility of storing data generated by a single researcher, or replication data for a journal, or the archive of an entire research project, or all the data created by the dissertations of an institution (Crosas, 2013). The Harvard Dataverse, powered by the Dataverse Network application, currently contains over 52,000 studies, is free and open to all researchers worldwide.

3 'Active' is defined as publishing at least ten articles across any number of issues in a year.
4 OJS: http://pkp.sfu.ca/ojs/
5 In the context of the Dataverse Network application, a 'Dataset' is a container within which researchers can upload their research data and its overall descriptive metadata.

Advantages to Integration

Among the many advantages to integrating Open Journal Systems and Dataverse, one of the most significant is the time savings to authors, editors, administrators and users of both platforms. For this reason we have focused our work on integrating with existing journal publishing and data deposit workflows, outlined later in this article. Above all, we want the output of this project to be useful; we are attempting to make no inherent judgments about existing publication workflows other than that they currently take too long and require too many clicks and logins across different platforms. The ancillary benefits to resolving these issues is immediately apparent: only needing to input metadata in one place not only increases the likelihood that it will be input, it greatly reduces the possibility of discrepancies or broken links between systems by creating permanent links between articles and data, enhancing the visibility of both. Additionally, the Dataverse API and the OJS plugin structure (using SWORD) are both designed to be reused for various other use cases. Thus, this project contributes to ongoing standardization, providing a model to be followed (as it does for data citation).

Challenges to Integration

One of the largest challenges to integration is admittedly the fragmentation of Open Journal Systems. Our testing to date, which necessarily has involved a cross section of the most avid and eager users of OJS, has already encountered some small bugs from older versions of the platform that cannot be easily upgraded due to site-specific customizations. However, we expect to have resolved many of these in time for a wider release, with OJS 3 to follow, bringing a hoped-for stabilization of production versions not long after. Another, non-software issue comes from heterogeneous article and data deposit policies. In some use cases, we expect journals to be linked to Dataverses that operate under different licensing terms, and are managed by different institutions. For cases like these, we have had to ensure that different deposit (and access or reuse) agreements are presented to authors and creators at the correct stages of the workflow, so that, for example, an author not consenting to the terms of a linked Dataverse is not precluded from publishing in an OJS journal. In the current working version, Dataverse Terms of Use are exposed via the API and can optionally be automatically duplicated within the OJS plugin interface.

Progress to Date on Project

From September 2012 to January 2014, this project has succeeded in reaching out to over 50 journals that cover various disciplines across the social sciences, biomedical sciences, geophysics, and humanities. Some provided feedback on designing data publication workflows and use cases for integration. Fall 2013 saw the development completion of initial versions of the Data Deposit API and OJS Dataverse plugin. During our second major development cycle in the winter of 2013-4, the aforementioned journals took part in testing the integration between OJS and Dataverse, and have provided valuable feedback on improving the plugin and API's functionality.

 doi:10.2218/ijdc.v9i1.311

Data Publishing Workflow Development

As previously mentioned, a small subset of OJS journal publishers and editors were consulted to provide feedback on any relevant data publishing use cases, which could be consistent with their journal publishing workflows. The following two figures illustrate different data publishing workflows being considered for this project. Workflow A (Figure 2) is for authors that deposit their research data at the same time as they submit their article to a journal by using the same platform for both. Should an article not be accepted for publication, authors will have the option to deposit their research data into a repository of their choice. Workflow B (Figure 3) is for authors that submit their article to a journal but already have their research data in a repository. So, instead of depositing the data again, they only need to include a formal data citation with a persistent link pointing to their data. Additionally, although not pictured below, it was also taken into consideration that authors would be able to deposit their data after they submit their article to a journal. This particular workflow would be useful if journal managers prefer to approve the article *prior* to requesting that the authors deposit their research data.

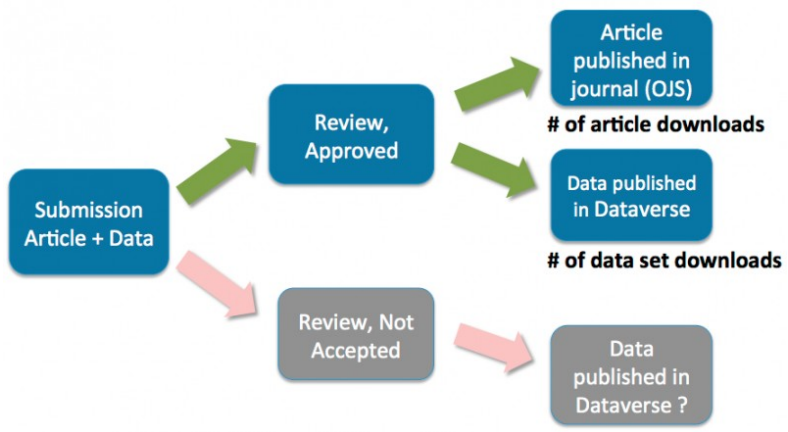

Figure 2. OJS-Dataverse data publishing workflow A.

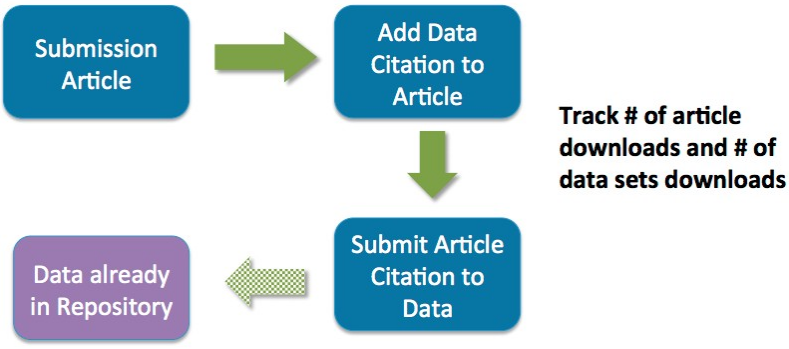

Figure 3. OJS-Dataverse data publishing workflow B.

Data Deposit API

Development of the initial version of Dataverse's Data Deposit API[6] was completed in early Fall 2013. The API is based on the SWORD[7] protocol, which stands for Simple Web-service Offering Repository Deposit, and is a profile of the Atom Publishing Protocol (AtomPub). This protocol is an open standard and is widely accepted by the scholarly community for the purposes of interoperability. SWORD was developed specifically to lower the barriers of depositing data and metadata from one 'scholarly system' to another. The most recent version, SWORDv2, was developed to support the whole deposit lifecycle of scholarly resources, which now allows for updating, replacing and deleting resources (Lewis, de Castro, & Jones, 2012).

In addition to interoperating with OJS, the Data Deposit API is now supported by a package developed for the R programming language by Thomas J. Leeper called the 'DVN R Package'[8]. Also, the Dataverse team is actively working with Research Compendia[9] and Open Science Framework (OSF)[10] to integrate with their systems[11].

OJS Dataverse Plugin

The OJS Dataverse plugin is designed to integrate as closely as possible within OJS' existing workflow. It utilizes the SWORDv2 library that ships with OJS 2.4.3 and newer versions[12], and when enabled, replaces OJS' existing supplemental file deposit form with the one shown in Figure 4, allowing a richer subset of DDI[13] metadata (shared by the Dataverse Network) to be applied to individual files. Authors are able to choose whether or not supplemental data should be made available for peer review at the time of deposit. Supplemental files that have been uploaded with this new deposit form are sent directly to a linked Dataverse, and all changes made to these files at a later date are automatically reflected in Dataverse. The plugin hooks into OJS' notification engine to provide alerts from successful or unsuccessful deposits.

The plugin also displays an article's data citation, along with a link to the data deposited in Dataverse (using permanent identifiers, such as DOI, where available) from the OJS sidebar when viewing an article, as shown in Figure 5. Future versions will allow this citation to be inserted automatically and directly into the reference list of an article. As discussed, the plugin can be configured to make supplemental data available at different points of an article's OJS publication workflow, depending on editorial preference; finer-grained permissions are available from the Dataverse interface.

The plugin is undergoing two full development rounds pending community feedback, and will be updated for OJS 3 as well.

6 Dataverse Data Deposit API: http://thedata.harvard.edu/guides/dataverse-api-main.html
7 SWORD: http://swordapp.org/
8 DVN R Package: http://cran.r-project.org/web/packages/dvn/index.html
9 Research Compendia: http://researchcompendia.org/
10 Open Science Framework: https://osf.io/
11 External Applications Working With Dataverse: http://thedata.org/book/apps
12 The plugin is supported as far back as the most recent release of the OJS 2.3.x branch, if the SWORDv2 library is installed manually.
13 Data Documentation Initiative: http://www.ddialliance.org/

 doi:10.2218/ijdc.v9i1.311

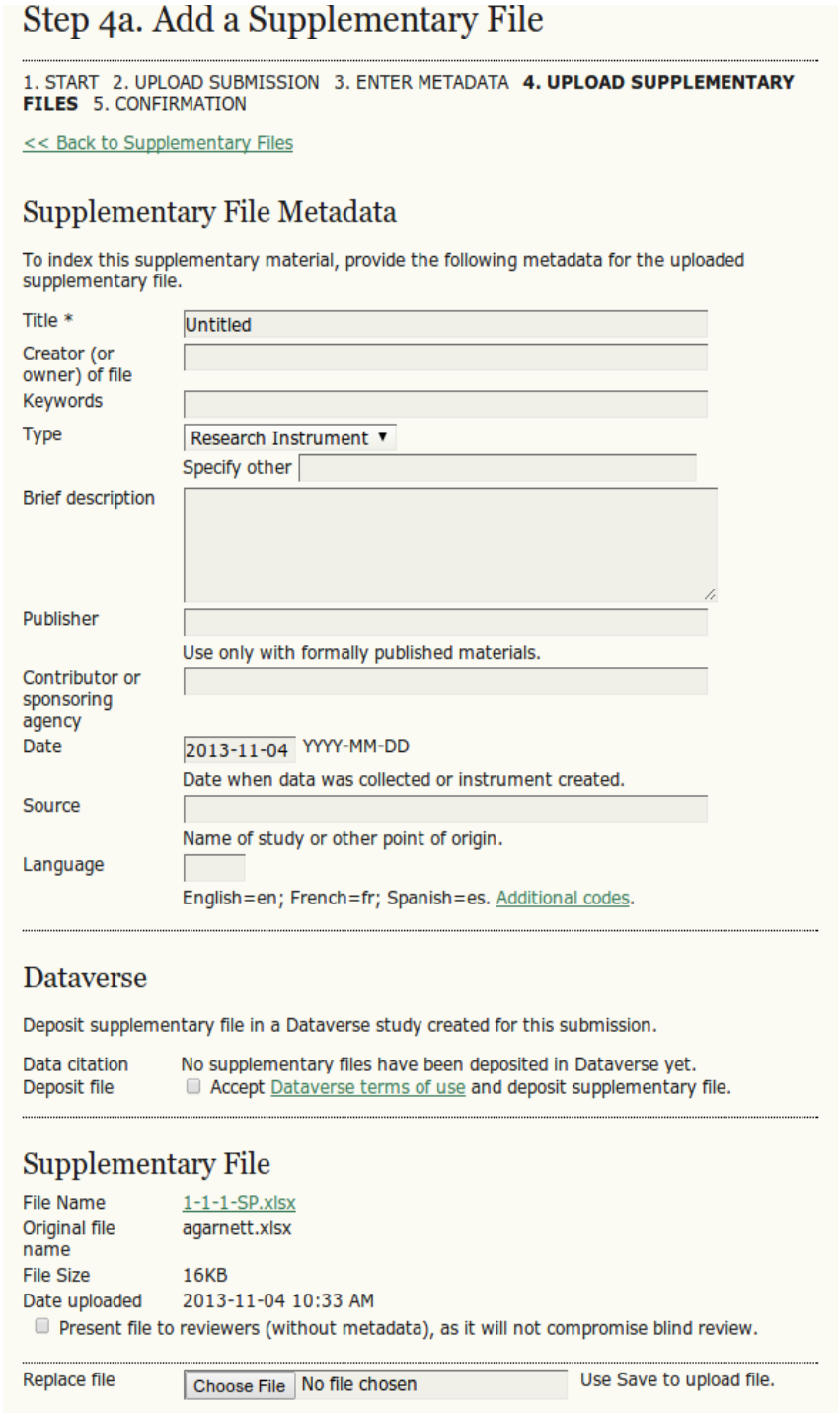

Figure 4. OJS' new supplemental file deposit form.

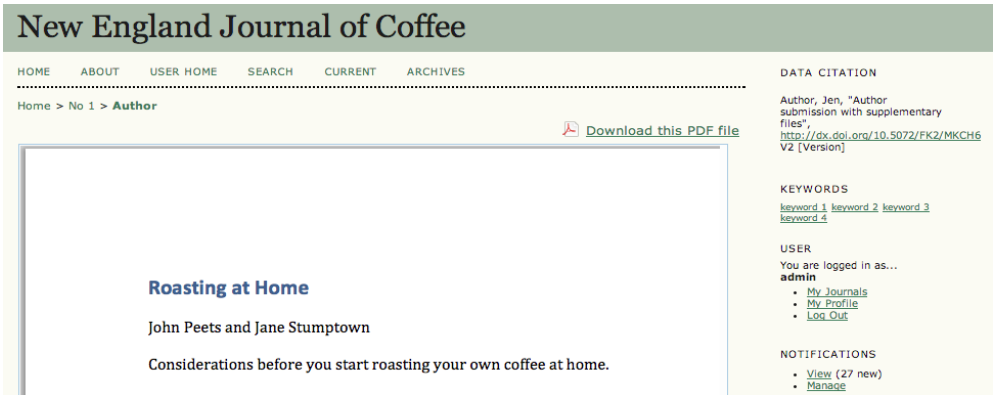

Figure 5. Article in OJS with a data citation at the top-right of the page.

Recent Survey Results

At the time of writing, our first round of plugin testing is nearly complete. Feedback has been largely positive so far, with testers (all of whom are relatively experienced OJS users) praising the integration with the existing OJS workflow. So far, the most commonly requested feature is the ability to customize which metadata fields are available as part of the supplemental file deposit form, which should be implemented in a future version. Users have also expressed some anxiety about the visual design of the plugin interface, particularly as it will relate to OJS 3, but these are not unexpected issues, and we look forward to addressing them in a second revision. Given the choice between an option to publish data within Dataverse immediately upon acceptance of the accompanying article in OJS, or to wait until the article is published, respondents are so far more or less equally divided. This is a good sign to us, as this option is user-configurable (and has interesting implications for time-to-publication statistics collected at a later date). Somewhat surprisingly, another feature that allows supplemental data from rejected submissions to be retained in the database for export to another Dataverse so that it may be published elsewhere without requiring additional effort by authors has been universally rejected, with respondents claiming that it should be authors' responsibility, rather than editors', to handle any issues related to resubmission. This feature will ship disabled by default.

Conclusion

Next Steps

Based on the testers' feedback, use cases for data publication will be refined. Concurrently, any reported bugs and feature requests will be scheduled for upcoming phases of development to improve on the API and plugin. This will be followed by larger scale outreach and support for OJS journals to try out the new Dataverse plugin. Recommendations for data publishing best practices (on reviewing, citing and sharing research data), along with any additional documentation will also be published on the project website[14]. Please note that since this is a reference implementation, any interested users or developers are encouraged to provide feedback on the open source plugin code and API at any time.

14 PKP Dataverse Integration Project: http://projects.iq.harvard.edu/ojs-dvn

Acknowledgements

Firstly, we would like to thank the Sloan Foundation for their financial support of the project. Secondly, Phil Durbin, Dataverse Software Developer, IQSS, Mercè Crosas, Director of Data Science, IQSS, and Cris Rothfuss, Executive Director, IQSS, provided valuable comments to the writing of the project summarized here. Additionally, special thanks are extended to the journals and publishers who partnered with us, and the entire PKP-Dataverse Integration Project team for their ongoing support of the project.

References

Crosas M. (2013). A data sharing story. *Journal of eScience Librarianship*, *1*(3), 173-179. doi:10.7191/jeslib.2012.1020

Gherghina, S., & Katsanidou, A. (2013). Data availability in political science journals. *European Political Science*. doi:10.1057/eps.2013.8

Institute for Quantitative Social Science. (2012). IQSS and PKP to develop data sharing system for journals [Press release]. Retrieved from http://www.iq.harvard.edu/

King, G. (2014). Restructuring the social sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science and Politics 47*(1), 165-172. Retrieved from http://j.mp/17Cobeu

King, G. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research*, *32*(2), 173–199.

Lewis, S., de Castro, P., & Jones, R. (2012). SWORD: Facilitating deposit scenarios. *D-Lib Magazine, 18*(1/2). doi:10.1045/january2012-lewis

MacGregor, J., Stranack, K., & Willinsky, J. (2014). The Public Knowledge Project: Open source tools for open access to scholarly communication. In S. Bartling & S. Friesike (Eds.), *Opening science: The evolving guide on how the Internet is changing research, collaboration and scholarly publishing* (pp. 165–175). Cham, Germany: Springer. doi:10.1007/978-3-319-00026-8_11

Vines, T.H., Albert, A.Y.K., Andrew, R.L., Débarre, F., Bock, D.G., Franklin, M.T., … Rennison, D.J. (2013). The availability of research data declines rapidly with article age. *Current Biology, 24*(1), 94–97. doi:10.1016/j.cub.2013.11.014

Vlaeminck, S. (2013, April 2). JoRD project presents results [Web log post]. Retrieved from http://www.edawax.de/2013/04/jord-project-presents-results/

Willinsky, J. (2005). Open Journal Systems: An example of open source software for journal management and publishing. *Library Hi-Tech, 23*(4), 504–519.

Wipperman, A. (2013, June 17). G8 science ministers' recommendations on access to research [Web log post]. Retrieved from Biomed Central blog: http://blogs.biomedcentral.com/bmcblog/2013/06/17/g8-science-ministers-recommendations-on-access-to-research/