Increasing Openness and Transparency in Research
Letter of Intent

Identifying Information
Project Title: Mutual Benefit of Text and Data Integration in Open Access Platforms
Brief Proposal I.D.: 50650
Applicant Name: John O'Connor
Legal Name of Applicant Organization: Harvard University

Brief Proposal Narrative
Introduction
Harvard University requests $65,000 in order to study the effects of minimal text and data integration on impact of Open Access research.

Problem Addressed
Hundreds of universities across the world have created and implemented "opt-out" Open Access (OA) policies, wherein OA for publication of research outcomes is the default. This model of OA policy has been relatively successful, with more universities creating similar policies every year. However, similar policies for research data are few and far between. Open research data is mandated by federal funders, a small number of private funders, and increasingly by OA journals. Nevertheless, open research data policies at research institutions have not been as successful as OA policies for publications.

Multiple studies have shown that OA publications have broader reach and higher impact than publications that are toll-access alone.[1] Similarly, there is evidence that sharing research data is also associated with increased impact.[2]

This study will investigate whether there are positive effects on the impact of OA research with a minimal type of integration between text in OA repositories with their underlying datasets in open data repositories. We believe that there will be a mutual benefit to minimal text/data integration that will speed the flow of research to action.

This type of integration between text and data has been minimally piloted between the platforms described in the "Methods" section below with success.

In addition to answering the primary question, this project will address four other issues:
1. Creating a generalizable workflow for performing simple text/data integrations to share with the OA community, including any code for implementation of those workflows;
2. Create proposals for how to do deeper text/data integrations than those in this study;
3. Propose solutions to text/data integration barriers between different OA and open data platforms; and
4. Recruit authors to integrate text and data from the beginning of the OA publication lifecycle to perform follow-up analysis to determine if there is a different effect on new publications compared to publications already released in OA repositories.

[1] Atchison, Amy, and Jonathan Bull. "Will OA Get Me Cited? An Analysis of the Efficacy of Open Access Publishing in Political Science." PS: Political Science & Politics 48, no. 1 (2015). doi:http://dx.doi.org/10.1017/S1049096514001668.
[2] Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. "Sharing Detailed Research Data Is Associated with Increased Citation Rate." PLOS One 2, no. 3 (March 21, 2007). doi:http://dx.doi.org/10.1371/journal.pone.0000308.

Methods Used
This project will take place over 19 months in coordination between Harvard University's Digital Access to Scholarship at Harvard (DASH) and Harvard Dataverse repository. DASH is an OA publication platform for all Harvard faculty and affiliates. It currently has over 30,000 articles posted with 8.5 million downloads across the platform. Harvard Dataverse is a public repository for sharing research data. It currently has over 60,000 datasets and nearly 2 million downloads.

The study will move in three phases:
Phase 1: Discovery of Articles and Data (~1 month)
In this phase, we will determine which articles in DASH have corresponding datasets in Harvard Dataverse and create a list of candidates for integration.

Phase 2: Author Permission and Integration (~4 months)
Although depositing publications and data in OA platforms would imply permission to link them together, it is the intent of the researchers to gain permission from each researcher whose work product we plan to use in the study.

We plan to use multiple different types of minimal text/data integrations in this study. All of the integrations will be HTML links back to the corresponding object in each repository. This means that we will add HTML links to objects in DASH and Dataverse in three ways: data to publication only, publication to data only, and binary linking between both. HTML links will be added to object pages in their respective platform. If there is a reasonable method for adding links into the objects themselves (e.g. a page in a PDF or a line in a CSV), we will attempt to add links in that way as well. Integrating text and data in these different ways will provide the clearest picture of whether binary integration is the driver of increased impact, or whether other types of linking are just as successful and potentially more efficient.

Phase 3: Metric Measurement and Integration Evaluation (12 Months)
In order to measure the effect of the integration, we will measure changes in three different metrics: page views for the object, downloads of the object, and citations of the object. These metrics (in order of weakest to strongest indicator) provide direct evidence of impact of the research. While we would ideally measure citations alone as the most direct evidence of impact, tracking citations continues to be a difficult problem and the timeline of the study as proposed does not allow for a significant number of citations to be produced.

During this phase we will complete work on the four other issues raised in the "Problem Addressed" section above.

Phase 4: Analysis and Publication of Findings (~2 Months)
This phase will incude analyzing the data collected in the previous 12 months as well as publication of findings. All publications and data will be released OA.

Potential Impact
If a mutual benefit is found, OA advocates will have evidence that even minimal text/data integration can speed the flow of information, build the demand for OA data and publications together, and result in more rigorous research.

In particular, this study will advance four goals of the grant:

RWJF Goal	Project Benefit
"…increasing transparency and accountability	If the integration increases the impact of

throughout the research life cycle…to increase the speed of the flow of information from funded research to action."	research, then it should follow that the research will be more quickly validated in peer review and put into action.
"…demonstrate the feasibility of practical application of an already established approach. Preference will be given to projects designed to produce implementable changes."	This study builds on established OA policy and platforms for data and publications. The integrations studied in this project will be readily implementable by any other OA platform.
"Investigate 'nudge policies'…that build the supply of and demand for more open and accessible research…"	Encouraging text/data integrations is a "nudge policy" for nearly all OA policies across the world. Providing evidence of increased demand through integration is the primary goal of this project.
"Propose innovations leading to greater efficiency in the peer review process, leading to increased speed, collaboration, transparency, and ultimately more rigorous research;"	Regardless of observed change in demand, text/data integration increases transparency in research by allowing reviewers to examine raw data and replicate studies more accurately, which leads to more rigorous research.

Statement of Interest in Openness

Peter Suber is the Director of the Harvard Office for Scholarly Communication, Director of the Harvard Open Access Project, Senior Researcher at the Berkman Klein Center for Internet and Society, and Senior Researcher at the Scholarly Publishing and Academic Resources Coalition. He has written extensively on Open Access, including the seminal overview of the subject *Open Access* (MIT Press, 2012). Peter has provisionally agreed to be Co-PI on this grant, contingent on approval from his current funding organization. He will provide supervision of the research process completed by John O'Connor.

Mercè Crosas is the Chief Data Science and Technology Officer at the Institute for Quantitative Social Science (IQSS) at Harvard University. She leads the vision and strategic direction of all software projects at IQSS, including the Dataverse project for data sharing and archiving, the Zelig project for statistical analysis, and the Consilience project for text analysis. Mercè has agreed to be the Co-PI on this grant.

John O'Connor is a recent graduate of the MS, Information Science and Master of Public Administration programs at the University of North Carolina at Chapel Hill. He was awarded the Educating Stewards of the Public Information Infrastructure (ESOPI^2) fellowship to help improve government stewardship of citizen information. Currently, he is creating OpenAcademicData.org, to house a vision and standards for creating open data portals in academia. John will perform the research for the grant, supervised by Peter and Mercè.

Conclusion

We believe that this project provides a substantial benefit for the OA community and advances RWJF's goals in OA. This project, studying the effects of text and data integration on research impact, has the potential to influence the structure of OA policies and portals into the future. We welcome any further requests for information from RWJF.

With Warmest Regards,
Peter Suber
Mercè Crosas
John O'Connor