Increasing Openness and Transparency in Research Full Proposal Narrative

Identifying Information

Project Title: Mutual Benefit of Text and Data Integration in Open Access Platforms
Full Proposal I.D.: 52040
Applicant Name: Suzanne Wones and Mercè Crosas
Legal Name of Applicant Organization: President and Fellows of Harvard College

Full Proposal Narrative

INTRODUCTION

Harvard University requests $70,000 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.[1] 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 are not as widespread 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.[2,3] Similarly, there is evidence that sharing research data is also associated with increased impact.[4,5]

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 *Methodology* section below with success. We seek to build on this trial by expanding the integration to all articles and datasets that can be connected.

In addition to answering the primary question of mutual benefit, 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.

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 releasing data OA simultaneously with publication of findings, and result in more rigorous research by providing a mechanism for testing reproducibility.

This study expects to directly fulfill four goals articulated by RWJF in the call for proposals. They are listed below in bold followed by an explanation of how the study expects to meet each.

"…increasing transparency and accountability throughout the research life cycle, with a special emphasis on open access, to increase the speed of the flow of information from funded research to action."

This project focuses on "green" open access; i.e. open access through repositories as opposed to open access through journals.

If a mutual benefit is shown, that will build evidence that releasing data alongside publications benefits the discipline as well as the individual researcher. While additional study will need to be done, potential downstream effects of the proposed text/data integration (discussed in *Methodology* below) could increase the speed of information from research to practice.

"…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 Dataverse software proposed for use is open source and has 21 installations worldwide. Additionally, nearly all OA repository software possesses the ability to change metadata attributes of a deposited article; adding even simple HTML links to the list of attributes would be trivial for most platforms once candidates for integration are identified.[6] The same is true for data repositories that provide metadata descriptions for datasets. A third category of repository allows for multiple objects to uploaded together under a single umbrella object (e.g. a project); repository software in this category would simply need to host article and data files to receive any mutual benefit. The ability to implement integration as proposed in this study is accessible to nearly all OA repositories.

As described in *Dissemination Plan* below, we expect to make publically available any source code and/or workflows that automate the matching and integration processes between DASH and Harvard Dataverse. This will enhance the ease of implementation of text/data integrations should any organization choose to do so. With minimal configuration and tweaking, the code could be used by any organization to perform basic text/data integration between their article and data repositories. For those organizations that cannot or choose not to make use of any source code, the published workflows will still provide a useful starting point for their own integration plans.

"Investigate 'nudge policies'…that build the supply of and demand for more open and accessible research…"

Investigating evidence of increased demand through text/data integration is the primary goal of this project. If a benefit is found, this project can provide evidence for slight changes to open access policies across research institutions to supplement current opt-out open access article policies with similar opt-out open research data policies. Even if no benefit is found, this project will provide methods to the public to increase the supply of open and accessible research by investigating an innovative idea in green OA and making the tools and workflows of the project available.

We consider that encouraging researchers to make their data open concurrently with article publication to be a 'nudge policy' that extends current open access article policies. All federal

agencies granting more than $100,000,000 in research and development funds have data publication requirements for grantees.[7] Many non-profit organizations have similar data accessibility requirements.[8] Given this environment, many researchers are already familiar with requirements to make data publically available. This study would build on those requirements by encouraging repository managers to integrate text and data, as well as build the case that adopting opt-out open research data policies is beneficial to researchers individually and research disciplines as a whole.
“Propose innovations leading to greater efficiency in the peer review process, leading to increased speed, collaboration, transparency, and ultimately more rigorous research…”
As multiple science disciplines struggle with a replication crisis, integrating text and data can provide one of many paths toward improving on the current state of science.[9] This project will create tools and ideas that lead to more rigorous research, regardless of whether there is an observed change in the impact of open access research. Text/data integration increases transparency in research by allowing peer reviewers to examine raw data, replicate studies more accurately, and review proposed publications with more context than before. While we expect that there will be a positive effect on the impact of the research, this project will produce useful outcomes and tools for increasing transparency in research even if the effect is not found. If a positive benefit is found, then it is reasonable to conclude that more people (professional researchers, amateur researchers, and/or the general public) are reviewing the literature and data after publication. This, then, is a significant data point that research is being consumed, put to use, and/or scrutinized by more individuals than it would have otherwise.

METHODOLOGY

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 project will move in four phases:

Phase 1: Article and Data Discovery (~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. Based on available information, we will first attempt to match DOIs of articles with a metadata field for the associated publication in Dataverse. In these cases, the integration has already been done one way (Dataverse has a link to the article), but not the other. If this method does not yield a suitably large list of article/data pairs, we will attempt to match on other metadata attributes such as object name, author name, etc.

Deliverables from this phase potentially include code and/or workflows for matching datasets and articles that provide bases for future researchers and practitioners to perform their own text/data integration.

Phase 2: Integration (~4 months)

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. Adding links in these three ways will provide necessary differentiation to distinguish whether one type of linking is more effective than another and whether the binary linking provides an advantage over unary. In addition, we will designate a control group of articles and datasets that will have been identified as related pairs but left without any integration between them to serve as a control group. This group will serve as a barometer for changes in impact due to circumstances outside of the integrations. Among each group we will attempt to choose articles/datasets with varying degrees of popularity before integration – again in an attempt to avoid biasing one type of integration with an unfair concentration of already-popular articles or datasets.

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.

Deliverables from this phase include the list of all candidate text/data pairs and the integration group they were added to, as well as workflows and/or code for performing integrations in DASH and Dataverse. For those institutions using Dataverse as their research data repository, this code will provide a basis for integrations using Dataverse. Even for those using other repository systems, the delivered workflows will provide a useful starting point.

Phase 3: Integration Evaluation and Solution Proposals (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. Further discussion of issues with choosing appropriate metrics for measuring impact is included in *Potential Challenges* below.

During this phase we will also complete work on the three other remaining issues raised in the *Problem Addressed* section above:

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.

The primary deliverable from this phase will be raw data of the metrics discussed above over the course of a 12-month period. Additional deliverables include text and diagrams of proposed workflows for deeper text and data integrations and potential workflows for integration between multiple data and text repository solutions. Looking beyond this study, we hope to use the text/data integration candidates as a starting point to recruit researchers to work on a longer-term study to investigate the effect of text/data integrations across the publication lifecycle rather than on those works that have already been published for some time. We expect to deliver to RWJF a plan for follow-on study of text/data integration in multiple ways.

Phase 4: Analysis and Publication of Findings (~2 Months)

This phase will include analyzing the data collected in the previous 12 months as well as publication of findings.

Key deliverables from this phase include formal findings and discussion of the study in the form of a journal article. Additionally, all deliverables will be disseminated as discussed in the *Dissemination Plan* below. This phase will also close the grant and provide any additional deliverables or reports as requested by RWJF.

DISSEMINATION PLAN

In accordance with the spirit of this grant and the convictions of the applicants, we commit to releasing nearly every deliverable of this grant to be freely and openly available to the public. The primary deliverable, a published article of the findings of the study, will be published in an OA journal, as well as uploaded into DASH. The article version in DASH will contain links to all other public deliverables. Any datasets created as a part of this project will be uploaded onto the Harvard Dataverse and (of course) linked to the DASH record of the associated article. Any code created as a part of this project will be uploaded into the Harvard Dataverse and hosted on a Harvard-managed GitHub account. All code will be licensed under a suitable OA/public license. Workflows created as part of this project will be hosted in Dataverse. The final substantial deliverable is a list of potential volunteers for a follow-on study to determine the effects of text-data integration throughout the publication lifecycle. Due to privacy concerns and the right of researchers to not participate in any studies at a later date, this list will not be shared with anyone except internally within Harvard as appropriate and with RWJF.

Additional dissemination may occur at regional and national conferences. Librarians and other repository managers will be encouraged to integrate text and data as long as the study does not show a negative relationship on impact (highly unlikely). These text/data integrations improve the transparency and quality of research, and are thus worth pursuing even without a substantial increase in impact.

POTENTIAL CHALLENGES

This study poses a number of moderate challenges, all of which we foresee as surmountable.

First, the Research Assistant position is not currently filled, and will have to be filled before the start of the project. If this project is funded, we commit to ensuring that the RA has appropriate access to all tools and platforms necessary to complete this project as described above. Suzanne Wones directs Director of Library Digital Strategies and Innovations and Mercè Crosas directs the development of the Harvard Dataverse. As Co-PIs on this project, they will ensure that the RA is given appropriate access DASH and Dataverse. We have additionally obtained approval and verbal offer of help from development teams from the Harvard Institute for Quantitative Social Science (IQSS), which oversees the Dataverse project.

Second, as elaborated upon in the Statement of Roles document, Suzanne will officially serve in an advisory capacity only. She will have as much involvement as she deems appropriate. She has agreed to serve as PI in order to ensure that RA has appropriate access to programs and platforms at Harvard to complete the work. Mercè will have as much involvement as she deems appropriate, outside of ensuring that RA has appropriate access to the Harvard Dataverse to complete the work.

Third, we do not have a good idea of how much work will be required to create a list of potential article/dataset pairs for integration and what the quality of that list will be. As discussed in *Methodology* above, we will attempt to create a pool of candidate pairs using DOI matching, object name matching, and author name matching. Based on a previous small-scale pilot of this integration work, we expect that a combination of these three metadata attributes will yield a pool large enough for study. If that does not yield a large enough pool, we will consider methods to manually generate candidate integration pairs. Currently, DASH has over 30,000 text objects while Dataverse has 60,000 dataset objects. This presents a very large raw pool from which to generate integration pairs; we fully expect to be able to generate a list of appropriate size to study.

Fourth, we may have trouble recruiting authors willing to participate in follow-on studies of the effects of text/data integration throughout the publication lifecycle. While we believe that we would have the right to use any OA article or dataset in any partner repository, finding authors willing to alert us to an article and dataset in a repository at the moment of deposit will be critical to planned follow-on studies. We expect to find a small contingency of authors at Harvard and will work with partner organizations to recruit more authors as appropriate. This deliverable can be used as the basis for follow-on studies.

Finally, we recognize that measuring page views and downloads as a measure of impact is a less-definitive measure than using citations. 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. Alternative metrics ("Altmetrics") such as downloads, page views, social media mentions, etc. are known to moderately correlate with scientific citations of articles, though there is some question of whether altmetrics measure scientific impact or other types of impact.[10] Regardless of whether altmetrics measure scientific or other types of impact, they still adequately measure if research is being engaged with in any way, and thus are measures of "demand for open and accessible research…" We believe that in the timeline presented here page views and downloads represent strong indicators of an object's impact both inside and outside academia and begin to show differentiation from the norm faster than citations. Citations of the version of record for articles and of datasets will also be measured, though we expect them to lag behind page views and downloads.

Overall, this project presents few potentially catastrophic challenges. The challenges presented above have all been evaluated by the applicants and we believe satisfactory solutions or mitigation strategies have been found for each. We welcome the opportunity to address any questions, comments, or additional concerns the grant committee may have.

PROJECT IMPACT

This project will have a positive impact on OA and research more generally regardless of whether the primary study shows any benefit to text/data integration.

First, it will generate tools and workflows for undertaking text/data integration in DASH and Dataverse, as well as across other popular repository software solutions. These tools and workflows will be publically available, providing immediate, tangible benefit to researchers and the OA community.

Second, it will generate proposals for how to integrate text and data more deeply than is proposed here, providing a basis for further research. These proposals, in combination with the volunteer list deliverable, will provide a substantial basis for additional studies of the effect of data sharing on research across multiple disciplines.

Finally, it will influence both new and pre-existing OA policies at research institutions and funding organizations by exploring the efficacy of a change in data sharing policy. Again, even if this study finds no benefit to text/data integration, it will continue to make the argument that such integrations provide a benefit to researchers by making their work more transparent. If, as we expect, there is a benefit to integration, this study will show that such integrations and other data sharing endeavors provide direct benefit to individual researchers in the form of increased impact and to research in general by making finding and access easier. In that case, this study can be used to justify changing or creating new data sharing policies in academia that highlight the need for data sharing and integration with their associated publications.

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, while also providing immediate benefit to the OA community through the creating of integration tools and workflows.

We welcome any further requests for information from RWJF and look forward to working with RWJF over the course of this project.