DATA-PE: A FRAMEWORK FOR EVALUATING DATA PUBLICATION POLICIES AT SCHOLARLY JOURNALS

N Moles

Faculty of Information, University of Toronto, 140 St. George Street, Toronto, Ontario, Canada, M5S 3G6
Email: n.moles@mail.utoronto.ca

ABSTRACT

With the growing importance of data to the scholarly record and the critical role journals play in facilitating data sharing, the complex landscape of scholarly journal data publication policies has become an obstacle for research. This paper outlines Data-PE, a framework for evaluating these policies. It takes the form of a conceptual foundation, comprising twelve criteria for evaluation, operationalized through an evaluation tool. Its objective is to function as a flexible means for a variety of stakeholders to appraise individual policies. Examples of the use of the framework are provided and means for the validation of the tool are discussed.

Keywords: Data publication, Data sharing, Scholarly journals, Evaluation framework, Editorial policy, Scholarly communication

1 INTRODUCTION

The emergence of data intensive research is making data an increasingly important component of the scholarly record. Accompanying this development, researchers in all fields are recognizing the tremendous potential of sharing and reusing datasets from earlier studies. Journals play a key role in facilitating this ongoing use of data through the publication of supplementary data supporting journal articles. This particular form of data is distinct from the raw data initially collected in research in that it has often undergone significant manipulation. It also relates directly to the conclusions reached in the article it accompanies, giving it added significance for the verification of results and the continuation of a line of inquiry. Research from the PARSE.Insight Project (http://www.parse-insight.eu/) indicates interest on the part of researchers in submitting supplemental data along with their publications as well as a prominent role for journals as a means for finding datasets for secondary analysis (see Smit, 2011 for a concise summary of the project's findings). With recent studies indicating higher citation rates for articles accompanied by supplemental data (Piwowar & Vision, 2013), facilitating data publication in scholarly journals is a necessary component for enabling future research.

Unfortunately, the process of making supplementary data available through publication is not straightforward. The degree of involvement in access, management, and preservation of these data varies widely between journals. Divergent conceptions of what is meant by data publication and what is involved in the process are widespread (RIN, 2008), resulting in a complex landscape that is difficult for stakeholders to navigate. Compounding the issue, researchers and other stakeholders may know little about data curation or have a limited awareness of the issues that need to be addressed for effective data publication. Many disciplines lack discipline-specific data centers to provide expertise, guidance, and leadership. In the event that such data centers are present, a gap often exists between the guidelines and standards developed for data publication and the practices of the communities intended to use them (Kotarski, Reilly, Schrimpf, Smit, & Walshe, 2012). The tension among the competing interests of publishers, funders, and researchers often means that data practices are largely determined by disciplinary norms and cultures. However, even the disciplinary norms themselves are not always a perfect match with the data sharing policies of funders (RIN, 2008). The turmoil resulting from this disjunction creates adverse effects for the development of research and barriers for the re-use of data.

In most cases, the roles and responsibilities a journal is willing to assume are articulated in data publication policies presented with the submission guidelines for authors and integrated into the publication's larger policy framework. These policies function as a means of understanding the actions and commitments of a journal towards data, making them an accurate proxy target for the evaluation of the data curation mechanisms in place. As Sturges et al. (2014)

have pointed out, although these policies vary dramatically, they also exist at a critical point in the research process where there is an immediate incentive for compliance. An evaluation method for scholarly journal data publication policies, informed by insights from relevant literature, can counteract the barriers in the current landscape and become an effective means for navigating the complexity of data publication.

This paper introduces and describes Data-PE, a framework for the evaluation of data publication policies at scholarly journals. The framework is intended to be used to determine the strengths, weaknesses, and appropriateness of data publication policies in relation to specific research situations. As the issues surrounding data publication are highly situational, the framework provides a means for stakeholders in any one of four major user groups to evaluate policies in light of their individual priorities and concerns. It does not attempt to mask over differences in data practices between fields but rather to be broadly applicable enough to be useful for a wide range of disciplines. The following sections of this paper discuss how the Data-PE framework relates to on-going developments in data publishing and the framework's potential users, before detailing the framework itself and its application. A final section will outline a methodology for validating the framework and potential directions for further research.

2 BACKGROUND

The Data-PE framework draws from, and is complementary to, many other initiatives in data publication. A significant amount of research has already been conducted on this topic and related issues. This section provides an overview of closely related projects and how they intersect with the framework.

Effective negotiation of the supplemental data publication process will require means to address and balance the divergent concerns of stakeholders. In 2011, the Opportunities for Data Exchange (ODE) project (http://www.dnb.de/EN/Wir/Projekte/Abgeschlossen/ode.html) produced a report on the integration of data and publications in scholarly communication. Amongst its insights, this report distinguished significant differences between the perspectives of researchers, publishers, and data centers. In exploring these viewpoints, the authors of the report identified generalizable requirements and concerns for stakeholders in each group as well as benefits to be gained from the link between articles and their underlying data. These include making the data more discoverable, guiding interpretation, and facilitating re-use (Reilly, Schallier, Schrimpf, Smit, & Wilkinson, 2011).

Observations from a follow up report produced by the same project support the idea that journals are capable of playing a significant role in data sharing through supplemental data publication. This report identified features that distinguish data from articles and the need for new data publication and citation conventions (Kotarski et al., 2012). It also concluded that publishers should assume responsibility for providing guidance to authors and that few academic libraries assist researchers to integrate data into their publications (Kotarski et al., 2012).

The importance of supplemental data policies at journals and the need for a tool to evaluate them was underscored by the National Information Standards Organization (NISO) in a 2013 report aimed at providing recommendations for publishers and fostering standardization in the treatment of supplemental data. Analyzing current practices, the report's authors identified three types of supplemental data defined by their relationship to the article: Integral Content (which is essential for full understanding by those in the field), Additional Content (data that is relevant, useful, and contextualizing), and Related Content (which is separate content that an author may wish to draw some connection with, but which is not hosted or managed by the journal) (NISO, 2013). The recommendations in the report vary based on the how the supplementary data fits within this classification and because of their separate management; no recommendations are made for Related Content. The submission of datasets in each of these categories will be subject to the journal's data publication policy. Like the second ODE report, NISO places the burden on publishers to communicate their requirements for supplemental data and provide written guidelines (NISO, 2013).

Two high profile projects have made significant contributions to the underlying principles on which data publication rests. The first is a recent report by the CODATA-ICSTI Task Group on Data Citation Standards and Practices, which takes a similar approach to NISO in building consensus through the study of emerging best practices. The centerpiece of this report is the following set of first principles, derived from observation of current behaviour and analysis of the existing literature (CODATA-ICSTI Task Group on Data Citation, 2013):

- status,
- attribution,
- persistence,
- access,
- discovery,
- provenance,
- granularity,
- verifiability,
- metadata standards, and
- flexibility.

Intended to be foundational, holistic, and inclusive of all stages in the publication and citation process, these principles have heavily influenced the development of the Data-PE framework because they integrate curation concerns with the requirements of stakeholders. The open data movement has also made a complementary contribution through the Panton Principles (http://pantonprinciples.org/), a set of four declarations recommended as the basis for open data in science. Created with the assistance the Open Knowledge Foundation (https://okfn.org/), these principles not only provide a strong statement about the desired openness of data but also act as a cornerstone on which more detailed and specific policies can be constructed.

Adding to this literature are two major studies of the attitudes and behaviours of researchers towards data that can be utilized to inform the mechanisms that control data publication. A report commissioned by the Research Information Network (RIN, 2008) contributes significantly to this area by exploring whether researchers do make their data available to others and their challenges in doing so. From the perspective of researchers, it identified ten barriers to data sharing and publication:

- lack of resources,
- limited time to respond to data requests,
- lack of data management expertise,
- concerns about accessibility and usability,
- lack of knowledge about archiving options,
- fear of loss of competitive edge,
- concerns about re-use value,
- limited reward for sharing, legal issues, and
- fear of misuse.

There is some overlap here with the factors that influence researcher data sharing behaviour identified by the PARSE.Insight Project. Although principally focused on preservation throughout the scientific endeavour, one of its deliverables was a survey of data management practices among researchers, publishers, and information professionals. Many of the issues touched on by this report relate directly to data publication. The survey revealed that while most journals accept supplemental data from authors, a large majority do not have preservation policies in place for the data they receive (Kuipers & van der Hoeven, 2009). The commonalities between these reports indicate significant points where the concerns and perspective of researchers and other stakeholders can be utilized.

Efforts to build conventions, develop standards, and provide meaningful guidance for data publishing are on-going. Two of the most active and promising current initiatives are the Australian National Data Service (ANDS) (http://www.ands.org.au/) and the RDA/WDS Publishing Data Interest Group (https://rd-alliance.org/internal-groups/rdawds-publishing-data-ig.html). ANDS is a government funded partnership of stakeholders that works to facilitate all aspects of research data management through outreach, expert guidance, training, and data citation services. The guides published by ANDS cover a wide range of related topics from costs and legal issues to metadata and infrastructure. Research Data Australia (http://researchdata.ands.org.au/) is a data discovery portal operated by ANDS to promote and facilitate re-use of Australian-produced research data. Working in conjunction with Research Data Australia, the Cite My Data (http://www.ands.org.au/services/cite-my-data.html), Identify My Data (http://www.ands.org.au/services/identify-my-data.html), and Register My Data (http://www.ands.org.au/services/register-my-data.html) services provide researchers with an array of options in making their data available. Like ANDS, the Research Data Alliance (RDA) is a partnership of stakeholders, organized to facilitate and enable data sharing in research. Amongst their various interest groups is a joint venture with the World Data System (WDS) focused on data publication. The RDA/WDS Publishing Data Interest Group has four working groups devoted to workflows, bibliometrics, costs, and publication services. Included in the mandate of the first of these working groups is an examination of the role of publishers and journals for the purpose of providing generic workflow models (Parsons, 2013).

Closely related to these initiatives, the Journal Research Data (JoRD) Policy Bank project (http://crc.nottingham.ac.uk/projects/jord.php) looked specifically at the data publication policies of scholarly journals. Its stated objective was to explore the feasibility of developing a sustainable service to aggregate and collate journal policies to assist in the navigation of this complex landscape. Like ANDS and the RDA/WDS Interest Group, the service envisioned by the JoRD Project was intended to provide guidance to researchers. An additional output of this project was the creation of a model for the development of journal policies, based on a review of existing literature and a consultation of stakeholders (Sturges et al., 2014). The model itself drew heavily from an earlier study by Piwowar and Chapman (2008) that reviewed the policies of scientific journals as they relate to the publication of biological gene expression microarrays, in order to identify strengths and link those strengths with actual data sharing practices. The model developed by the JoRD project makes a significant contribution by identifying important areas that policies need to address and providing a roadmap for tailoring these issues to field or discipline specific circumstances. However, this model is too recent to have been adopted widely enough for its effectiveness to be tested, and the policies based on this model lack an independent evaluation mechanism. Despite these limitations, in conjunction with the work of ANDS and the RDA/WDS Interest Group, they comprise a powerful array of tools.

Taken together, this work illustrates a complex landscape of inconsistent practices and conflicting stakeholder needs. This body of research also indicates a demand on the part of multiple stakeholders for tools to help them build policies and best practices. While much progress has been made in developing capabilities for data publication, none of the projects discussed above fulfills the unique function of an evaluation mechanism for data publication policies. In contributing to this situation, the Data-PE framework draws from insights produced by these studies, incorporating the perspective of stakeholders with foundational principles while remaining sufficiently flexible to accommodate diverse data publication scenarios.

3 USES AND USERS

The Data-PE framework has four distinct user groups: researchers, information professionals, publishers, and funding agencies. For researchers, as the primary user group, the Data-PE framework is a tool to assist in the choice of journals in which to publish data-intensive or heavily data-reliant research. In this capacity, it would allow them to meet the requirements of their funding agencies in terms of both public access to the data outputs of research and the long-term curation of research data. Used in conjunction with the online tools generated by the SHERPA Project, such as RoMEO (http://www.sherpa.ac.uk/romeo/), JULIET (http://www.sherpa.ac.uk/juliet/), and FACT (http://www.sherpa.ac.uk/fact/), which assist in navigating copyright and open access policies, it creates a robust platform for publication decision-making. Utilizing the Data-PE framework for this purpose will allow data considerations to exist in parallel with reputation and impact factor as means for navigating the landscape of scholarly communication. This three prong approach to publication will help to embed data considerations in research workflows and promote data as a first class scientific output.

A tool for the evaluation of data publication policies is also of significance to the secondary user groups. By identifying journals that would be most appropriate for specific communities, the Data-PE framework can help librarians in collection development and repository staff in collaborating with academic publishers. On the other side of this relationship, publishers themselves can use the tool to refine future data publication policy and ensure that journals are meeting the needs of their readers. These policies give publishers a chance to display their involvement with issues of concern in scholarly communication and promote their journals as an outlet for data intensive research.

Potential uses for the framework also extend to funding agencies. Access to publically funded research is becoming a more pressing issue. Organizations like the National Institutes of Health (NIH) (http://www.nih.gov/) in the United States and the Engineering and Physical Sciences Research Council (EPSRC) (http://www.epsrc.ac.uk/) in the UK are both placing increasing emphasis on the requirement that data produced from the research they fund be made available without barriers. Funding agencies can use the Data-PE framework to evaluate the outlets by which funded research is made available and the role of individual journals in that process. In this context, the framework can determine which journals meet the requirements set out by the agency for making data accessible and produce additional input for evaluating the compliance of funded research and the development of further policy.

4 DATA-PE FRAMEWORK

The Data-PE framework is structured around a series of concepts articulated as criteria that when combined, capture the role and effectiveness of the journal in data publication. This format allows the framework to draw from the existing literature on data publication and citation, condense curation knowledge, and identify the elements necessary for a journal to be an effective facilitator of data publication. Because the role journals are willing or able to play in data publication will vary, as will their appropriateness for the communities they serve, the framework provides users with tools to make value-judgements specific to their situation. The requirements and concerns of each of the user groups will always be multidimensional and the criteria aim to capture the shared core of those dimensions. As such, it is designed to be flexible enough for users to add new criteria. In this way, it can be modified as needed or tailored to specific situations, users, or the requirements of particular disciplines while retaining the fundamental insights. The framework itself is structured in two components: a conceptual foundation and an evaluation tool.

4.1 Conceptual Foundation

The first half of the Data-PE framework consists of a conceptual foundation made up of twelve interrelated criteria against which policies can be evaluated. There is considerable overlap here with both the CODATA first principles for data citation and the policy model created by the JoRD project. While elements of both projects have been incorporated, Data-PE has a distinct articulation that alters the perspective from general principles and a structure for the development of policies by publishers to an evaluation by stakeholders. The criteria used are not simply an aggregation of the dimensions of data publication found in the literature. Many of the concerns raised by previous research do not have a direct correlation in the conceptual foundations. The aim of this component of the framework is to identify and articulate deeper underlying concepts. Individual aspects of data publication explored in other studies have often been omitted, disassembled, or conflated in order to expose their foundational dynamics. An example can be seen in the concept of *timeliness of availability* used by Piwowar and Chapman (2008) as a factor in determining policy strength. Later research has indicated that the actual concern, from the perspective of researchers, is not the point in the *publication* process at which data is submitted but instead the point in the *research* process at which data is shared (Sturges et al., 2014). As a result, the concerns related to the timeliness of submission are addressed by a range of criteria that include Stewardship, Creator Statement, Pre-Existing Conditions, and Provenance & Attribution. In a similar way, the requirements of other stakeholders such as funding agencies or universities are unlikely to be restricted to a single aspect of data publication and as a result can be more productively addressed through a multi-dimensional approach and are not represented as individual criteria here.

These twelve criteria combine the concerns of researchers with the digital curation concepts necessary for effective data publication. Covering a continuum from managerial to technical, they provide sufficient versatility to examine journals from a wide range of disciplines. Figure 1 illustrates the twelve criteria while the section below defines each of these and operationalizes them through a series of questions users can ask about the data policy to determine its effectiveness.

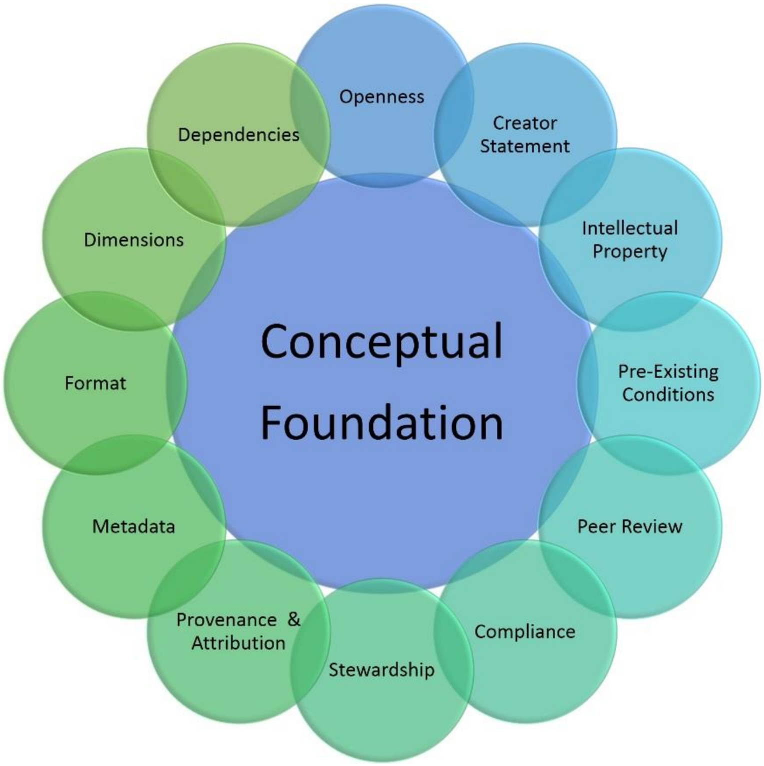

Figure 1. The concepts and criteria of the Data-PE Framework Conceptual Foundation

OPENNESS: This concept refers to the ability of users to obtain access to the dataset with minimal barriers. While desirable for a number of well documented reasons, the optimal degree of openness for a particular dataset is determined by a number of factors such as funder requirements, privacy concerns, economic interests, and the openness of the journal itself. In this capacity, journals act as gatekeepers, and their policies with regard to data are often closely related to those pertaining to articles. Within this aspect of the framework, the following questions are useful: Is the journal itself open access or does it have an open access option? Does the journal install or create any barriers to accessing the data? Is access to the supplemental data linked or mediated by a third party service provider who may have additional charges? Does the policy meet the needs of the researcher's funders?

CREATOR STATEMENT: The Panton Principles have demonstrated the importance of including a clear statement of the data creator's wishes for the dataset along with the data itself in the publication process (Murray-Rust, Neylon, Pollock, & Wilbanks, 2010). Having this element linked to the data will facilitate re-use and help to assuage fears that the data will be misinterpreted or misused. In applying these criteria, users should ask the following questions of a data publication policy: Does the journal provide a means for including an explicit statement from the dataset creator as to their wishes and/or intentions as to what can or should be done with the data? Can the creator declare their obligations to their funders? Does the policy provide guidance for the authors on how to describe their use of the data? Will the creator statement be perpetually linked to both the data and the article? Where will the statement reside in relation to the data and the article?

INTELLECTUAL PROPERTY: Consideration of intellectual property issues is necessary if creators are to receive credit for their work and for the data to be usable by other researchers. The publication policy must strike a balance between the needs of the creator and those of later users while still complying with the norms and expectations of the field. Researchers in particular are rarely clear on intellectual property issues (RIN, 2008) and this criterion seeks clarification. Questions to evaluate this criterion include: Does the journal ask that authors sign over rights to the publisher? Does the journal demand that the data be in the public domain? Does the journal use a recognized licence or waiver (The Open Knowledge Foundation provides a list of conformant licences on their website, available here: http://opendefinition.org/licenses/#Data), as many traditional open access licences such as GPL are not appropriate for data (Murray-Rust, Neylon, Pollock, & Wilbanks, 2010)?

PRE-EXISTING CONDITIONS: As the role journals play in relation to data is limited to publication and management, once initial research has already been completed, the datasets submitted will likely have terms, conditions, and concerns already attached to them by this point in the research process. These can range from specific funder requirements to confidentiality necessitated by data collection methods. This criterion reflects the compatibility of a policy with pre-existing conditions connected to a dataset before its submission. Included under this banner are concerns about how data publication will impact the on-going progress of the research that generated the data and the need for researchers to retain exclusive use of the data until their line of inquiry has been exhausted. Relevant questions under this heading are: At what point in the publication process do the data need to be submitted? Is a data embargo possible? Does publication of the data through this journal compromise the confidentiality promised to study participants? Does publishing data as the policy outlines endanger future research already planned or underway?

PEER REVIEW: If data are to assume the same status as journal articles, it is imperative that they be given similar consideration during peer review. In scholarly publishing, the peer review process is the mechanism for quality control. Unfortunately, in regards to data this process is often poorly defined (Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011). It is to the benefit of authors to know whether data are given a review comparable to articles. Lawrence et al. have provided a generic data review checklist that addresses data quality, metadata quality, and a general review (2011). This criterion does not mandate that this checklist be used, only that the process itself be clearly understood. Questions to ask of a data policy on this point include: Does the dataset get reviewed alongside the article as part of the process? If so, at what depth do the data get examined? Do reviewers have clear instructions from the publisher? Do the reviewers have the software and information they need to correctly interpret the data? Does this process include consideration of quality control and metadata?

COMPLIANCE: Closely related to peer review, the concept of compliance centers on the relationship of authors and data submitters to the policy itself. Sturges et al. have argued that journal research data policies should have clear and prominent consequences for non-compliance to ensure that the policy is consistently followed (2014). This criterion captures all mechanisms for monitoring that authors comply with the data policy as well as all systems of enforcement or penalties for non-compliance. Questions that can be asked to evaluate a policy on this criterion include: What penalties exist for non-compliance? Do the penalties for non-compliance act as a reasonable deterrent? Will compliance with the policy be monitored by the journal? Is there a reporting mechanism for readers to draw attention to non-compliant data submissions? Are some components of the policy more strictly enforced than others? If so, are these priorities clear to authors? Do authors have the opportunity to correct any deficiencies before incurring penalties?

STEWARDSHIP: The term stewardship is an umbrella concept that encompasses all of the roles, responsibilities, and actors needed for the curation of data over the short, medium, and long-term. All of these aspects need not be undertaken by the journal itself but may be done through partnerships with repositories or other organizations. This aspect evaluates the presence of provisions for data curation and links to credible collaborators. Information about these capabilities can be obtained through the following questions: Does the journal have a statement of its responsibilities and capabilities with regards to the curation of data? Is the journal linked with a data center or repository that will have the specialized knowledge necessary for effective data curation? Does the journal curate the datasets itself or recommend a credible third party repository? Does the journal or partner use DOI (Digital Object Identifier) (http://www.doi.org/) or some other system for persistent identification? If stewardship is handled by a third party, do they have a mechanism for bi-directional linking between the article and the data? Are there recommendations from either the journal or a partner repository regarding effective data citation practices?

PROVENANCE & ATTRIBUTION: This aspect of the framework covers two closely related concepts. Internal to a dataset, the concept of provenance refers to the description of where a particular piece of data comes from and the process by which it has come to be in its place (Buneman, Khanna, & Wang-Chiew, 2001). Extending this concept externally, attribution is the process or capability of assigning responsibility for a dataset, and for its current state, to those responsible for it. Both of these complementary factors are necessary for the reproduction of the results of data analysis and for functional data citation. The concept of attribution should also give scholarly credit and interface with citation metric (CODATA-ICSTI Task Group on Data Citation, 2013). To meet these requirements, users should ask: Does the journal require documentation of provenance or the processes by which datasets are created? Does the journal have a means for maintaining this information and making it available to future users of the dataset? Does the journal provide a means for clearly stating the attribution of the dataset to its creator? Are there recommend best practices for citation of the dataset?

METADATA: Here the criterion refers to the metadata that accompany and describe the dataset in question. It includes all metadata linked to the dataset regardless of their creator, semantics, schema, or intended function. In considering this aspect of the framework, users should consider the following questions: What are the metadata requirements for datasets submitted to the journal? Does the journal use established and widely used standards? Are discipline specific standards available and being used? What degree of, or to what extent, is metadata description of the dataset required by the journal? Who will be responsible for meeting these requirements? Will additional work be necessary to prepare the dataset and who will be responsible for it? Does the journal or its partners use established metadata standards?

FORMAT: In general terms, this refers to the arrangement of bits in a digital object that allows for it to be coherently rendered and manipulated. Most disciplines have well-established software formats for research data. However, specific projects or tools may require less known or highly specialized formats. This concept incorporates consideration of data formats into the framework. Questions under this criterion include: Which formats does the journal accept? Does the journal accept the most widely used formats in the field? Will the working dataset need to be transformed to meet the journal's requirements? Do they insist on proprietary formats or accept open formats? Does the journal accept raw data, processed data, or both?

DIMENSIONS: In this context, the term dimensions refers to the characteristics of the dataset as a digital object, independent of its semantics. Journal data policies may have set limitations on the characteristics of the datasets it will accept. These limitations can regard size, number of attributes, granularity of the data, or the complexity of the interrelationships of its components. Useful questions in this area include: Are there limitations on the size of the data the journal will accept? Is the journal capable of handling the full complexity of the interrelationships necessary for correct interpretation of the data? Will the journal or its readers be able to work with the data at the granularity described in the article?

DEPENDENCIES: All digital objects require a combination of hardware and software in order to be rendered, processed, and manipulated. This component of the framework captures those requirements and includes any computational or analytical tools on which the analysis is based. Considering the dependencies of a dataset, the user can ask: Does the journal or its partners have the necessary software or hardware to render, process, or manipulate the dataset? Has custom code been written for the processing or analysis of the data that produced the results? If so, does the journal's data policy have a mechanism for addressing this limitation and reproducing the analysis? Is the researcher willing or able to submit software or algorithms along with the data?

4.2 Evaluation Tool

The second half of the Data-PE framework is an evaluation tool that embodies the twelve criteria from the conceptual foundation in a grid-based schema for comparing multiple policies or evaluating them individually. Echoing the evaluations of none, weak, and strong used by Piwowar and Chapman (2008), each criterion is given a score of 0, 1, or 2 in response to the questions above. These values are associated with the following evaluations: (0) Absent, (1) Obscure/Insufficient, (2) Clear/Appropriate. The condensation of multiple questions into a single score will provide room for individual judgement that allows the tool to reflect the unique circumstances of individual research situations. Assigning a score to each criterion will allow the identification of specific strengths and weakness within each policy. By adding together the total score, policies can be given an overall ranking and multiple policies can be compared.

In May, 2014 an evaluation and comparison of the data publication policies at four journals in the field of genetics was conducted using the Data-PE evaluation tool. Table 1 illustrates the use of the tool in evaluating policies from the following four journals: PLoS Genetics (http://www.plosgenetics.org/), Genetics (http://www.genetics.org/content/current) (Genetics Society of America), Genetics Research (http://journals.cambridge.org/action/displayJournal?jid=GRH), and European Journal of Medical Genetics (http://www.journals.elsevier.com/european-journal-of-medical-genetics/). Unfortunately, detached from any specific circumstance of research, this evaluation was done outside of the context of data sharing incentives and constraints. Conducted without input from actual research considerations, this evaluation does not reflect issues such as commercial interests and the desire to protect the ability to make future publications from a particular datasets, which studies have shown influences data sharing practices in the field (Campbell, Clarridge, Gokhale, et al., 2002). It is presented here for demonstration purposes only. This scoring also favours open access to data and assumes a well-curated dataset prior to publication, neither of which is universally true for the field of genetics. The evaluation tool with full scores is below:

Table 1. Comparison and evaluation of data publication policies at four journals in the field of genetics using the Data-PE evaluation tool

	Openness	Creator Statement	Intellectual Property	Pre-Existing Conditions	Peer Review	Compliance	Stewardship	Provenance & Attribution	Metadata	Format	Dimensions	Dependencies	Total Score
PLoS-G	2	0	2	1	1	2	2	0	1	2	2	2	17
Genetics	2	1	1	1	1	2	2	0	1	1	1	0	13
GR	1	2	1	0	1	1	2	0	1	0	0	0	9
EJMG	1	0	1	0	0	0	1	0	1	1	1	1	7

None of the data publication policies at these journals adequately addressed all of the criteria in the Data-PE framework. Particular weak points were Provenance & Attribution, which no policy acknowledged directly, Pre-Existing Conditions, and Metadata, which appeared to be intentionally vague in some cases. Somewhat surprisingly, the strengths of these policies were in Openness, Intellectual Property, and Stewardship. Likely as a result of legal concerns and pressure from open access advocates, all journals provided statements about intellectual property and some degree of openness. In the case of the closed publications, European Journal of Medical Genetics, Genetics Research, and Genetics, the article abstract and supplementary data are open, or the option of providing open access is available for an additional fee. Each of the journals was also willing to provide information about the stewardship of data submitted although in one case this was not deemed satisfactory. In the final comparison, PLoS Genetics provided the most robust and comprehensive policy, despite some significant shortcomings.

5 VALIDATING THE DATA-PE FRAMEWORK AND FURTHER RESEARCH

In addition to actual use, the outline of the Data-PE framework above holds great potential for further research in a number of different directions. Most prominent amongst these are validation, further development, and an exploration of the potential symbiotic relationship with the JoRD policy model. The most effective means for testing and validating the Data-PE framework is through documentation of its adoption and use in real-world scenarios. By reaching out to researchers, librarians, publishers, and funding agencies from a wide range of disciplines to test the framework, the uniqueness of considerations in each group can be incorporated. The larger the number of test instances, the more robust the validation of the framework will be. Real-life evaluations will inevitably differ from the examples provided above and all four user groups hold great potential for generating input into the evaluation of individual policies and the continuing evolution of Data-PE. Variations in the weighing of criteria will test the flexibility of the framework and potentially add new insights.

The key to successful validation of the framework lies partly in achieving the widest extent of use possible but also in the opportunity this use presents to gather feedback and harness it as a resource for further development. Capturing the scoring and feedback from diverse users and aggregating results of the evaluations and the rationale for ratings will create a resource that future users can consult in their own decision-making processes and use as a shorthand to save the time and effort necessary for a full evaluation. As the volume of evaluations and supporting rationales grows, this dataset can itself be subjected to analysis to determine the consensus of the research community regarding the evaluation of certain data publication policies and to note changes over time or responses to policy developments. Input from each of these user groups is a necessary pre-requisite for progress in the maturation of the framework. It is only once this information has been gathered and consensus has been reached, that effective benchmarks can be established and prescriptive guidelines for not just policy, but the larger role of journals in data publication, can be created.

The potential for cyclical and mutually re-enforcing development of data publication policy derived from on-going evaluation is captured by the relationship of Data-PE with the JoRD policy model. Data-PE provides a means for introducing input from stakeholders linked directly with the particular data publication situation and which complements the broader empirical research on which the model is based. Perhaps more significantly, in this context, the Data-PE framework can function as an evaluation mechanism for policies developed using the model while the JoRD model can serve to validate the framework. As projects with similar objectives, these tools work together to build the base knowledge of data publication, push the emergence of best practices, and promote awareness of the issues surrounding data publication while encouraging wider use of the framework.

6 CONCLUSIONS

The Data-PE framework extends existing research into data publication and fills a crucial gap in the current landscape by directly addressing capabilities at a key point in which data are accessed and discovered. By providing a range of key stakeholders with a means of evaluating data publication policies, which is adaptable to unique research situations, the Data-PE framework facilitates effective navigation of this complex landscape and enables successful data sharing through the publication process. As the policies function as proxies or representations of the data curation capabilities of scholarly journals and their partners, the framework provides an opportunity for evaluators to address curation concerns more broadly. Adoption of this framework by the user groups detailed above will connect data publication with other elements in the curation of data for data-intensive research, such as citation and preservation. Its use will also promote uniform data practices by pressuring journals to develop more sophisticated, nuanced, and applicable data publication policies. Although Data-PE is an important step forward, it is not a complete solution to current barriers in data publication. As abstracting and indexing services have only recently begun to track data (The Thomson Reuters Data Citation Index was launched in October, 2012: (http://thomsonreuters.com/data-citation-index/) or indicate if supplemental data are available with an article, major issues associated with discoverability remain. The framework has no mechanism for correcting discoverability at this larger scale nor does it address the issue of whether the journal policies are actually implemented and followed. Both of these challenges are outside of the scope of the framework and are larger issues than can be dealt with here. Despite these limitations, the Data-PE framework makes a significant contribution to the on-going development of data publication and in doing so helps to overcome obstacles to data intensive research.

7 ACKNOWLEDGEMENTS

The author would like to thank Dr. Seamus Ross and Dr. Juan Ilerbaig of the Faculty of Information, University of Toronto for providing feedback and assistance in the preparation of this manuscript.

8 REFERENCES

Buneman, P., Khanna, S., & Wang-Chiew, T. (2001) Why and Where: A Characterization of Data Provenance. In den Bussche, J.V. & Vianu, V. (Eds.), *Database Theory — ICDT 2001*, Berlin: Springer

Campbell, E. G., Clarridge, B. R., Gokhale, M., et al. (2002) Data withholding in academic genetics: Evidence from a national survey. *JAMA 287*(4), 473–480.

CODATA-ICSTI Task Group on Data Citation. (2013) Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal 12*. Retrieved October 29, 2013 from the World Wide Web: https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article

Kotarski, R., Reilly, S., Schrimpf, S., Smit, E., & Walshe, K. (2012) Report on best practices for citability of data and on evolving roles in scholarly communication. *Opportunities for Data Exchange (ODE)*. Retrieved April 21, 2014 from the World Wide Web: http://www.stm-assoc.org/2012_07_10_STM_Research_Data_Group_Data_Citation_and_Evolving_Roles_ODE_Report.pdf

Kuipers, T. & van der Hoeven, J. (2009) PARSE.Insight: Survey Report (Deliverable No. D3.4). *PARSE.Insight Project*. Retrieved April 21, 2014 from the World Wide Web: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011) Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation* 6(2), pp 4–37.

Murray-Rust, P., Neylon, C., Pollock, R., & Wilbanks, J. (2010) Panton Principles, Principles for open data in science. Retrieved April 20, 2014 from the World Wide Web: http://pantonprinciples.org/

National Information Standards Organization (NISO) (2013) Recommended Practices for Online Supplemental Journal Article Materials (NISO Recommended Practices No. NISO RP-15-2013). Baltimore, MD. Retrieved April 21, 2014 from the World Wide Web: http://www.niso.org/apps/group_public/download.php/10055/RP-15-2013_Supplemental_Materials.pdf

Parsons, M.A. (2013) Publishing Data Interest Group Charter. *Research Data Alliance (RDA)*. Retrieved May 6, 2014 from the World Wide Web: https://www.rd-alliance.org/filedepot?cid=132&fid=129

Piwowar, H.A. & Vision, T. J. (2013) Data reuse and the open data citation advantage. *PeerJ* 1.

Piwowar, H.A. & Chapman, W.W. (2008) A Review of Journal Policies for Sharing Research Data. *Proceedings of ELPUB 2008 Conference on Electronic Publishing*, Toronto, Canada.

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011) Report on Integration of Data and Publications. *Opportunities for Data Exchange (ODE)*. Retrieved May 6, 2014 from the World Wide Web: http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/ODE-ReportOnIntegrationOfDataAndPublications.pdf

Research Information Network (RIN). (2008) To share or not to share: research data outputs. UK. Retrieved April 21, 2014 from the World Wide Web: http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs

Smit, E. (2011) Abelard and Héloise: Why Data and Publications Belong Together. *D-Lib Magazine 17* (1/2). Retrieved April 8, 2014 from the World Wide Web: http://www.dlib.org/dlib/january11/smit/01smit.html

Sturges, R.P., Bamkin, M., Anders, J.H.S., Hubbard, B., Hussain, A., & Heeley, M. (2014) Research Data Sharing: Developing a Stakeholder-Driven Model for Journal Policies. *Journal of the Association for Information Science and Technology* (Forthcoming). Retrieved June 5, 2014 from the World Wide Web: http://eprints.nottingham.ac.uk/3185/1/Research_Data_Sharing_Jord_article_with_table.pdf

(Article history: Received 29 September 2014, Accepted 04 January 2015, Available online 19 January 2015)