Rewarding data sharing and reuse: Initial results of an interview study

Kathleen Gregory[*], Anton Boudreau Ninkov[**], Emma Roblin[***], Sarah Lawrance[****], Isabella Peters[*****], Stefanie Haustein [******]

[*]*kathleen.gregory@unive.ac.at*
https://orcid.org/0000-0001-5475-8632
Faculty of Computer Science, University of Vienna, Austria
School of Information Studies, Scholarly Communications Lab, University of Ottawa, Canada

[**] *anton.boudreau.ninkov@umontreal.ca*
https://orcid.org/0000-0002-8276-7656
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada

[***]*erubl066@uottawa.ca*
https://orcid.org/0000-0002-9179-0620
School of Information Studies, Scholarly Communications Lab, University of Ottawa, Canada

[****]*sarah.lawrance@uottawa.ca*
School of Information Studies, Scholarly Communications Lab, University of Ottawa, Canada

[*****]*i.peters@zbw.eu*
https://orcid.org/0000-0001-5840-0806
ZBW-Leibniz Information Center for Economics & Kiel University, Germany

[******]*stefanie.haustein@uottawa.ca*
https://orcid.org/0000-0003-0157-1430
School of Information Studies, Scholarly Communications Lab, University of Ottawa, Canada
Observatoire des Sciences et des Technologies (OST), Centre
Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal

Abstract
How do researchers want to be recognized and rewarded for sharing and reusing data? Which metrics are of interest? This research-in-progress paper presents initial findings from a semi-structured interview study investigating data practices of researchers across disciplines who both reuse or do not reuse data. We report the initial findings of a thematic analysis of questions related to recognizing and rewarding 'data work' and examine the role of citations as mechanisms for reward. We conclude by discussing next steps in our ongoing analysis and outline future directions.

1. Introduction

Data sharing and reuse are recognized as important pillars in conducting and enabling open science. These practices are shaped by existing infrastructures, policies and disciplinary norms (Borgman, 2015), as well as by barriers ranging from a lack of time to concerns about protecting future research directions (Tenopir et al., 2020). Another potential barrier is that the work involved in managing and sharing data is often not currently recognized in systems of academic reward (Alperin et al., 2020).

Data citations and metrics based on views or downloads have been suggested as mechanisms for rewarding and incentivising data sharing and reuse (Lowenberg et al., 2019). But do researchers view data citations as 'rewards' in academia? Would researchers like their data work to be rewarded and assessed? Which metrics, if any, would be of interest?

This **research-in-progress paper** presents initial results from an interview study addressing these questions with both reusers and 'non-reusers' of research data across disciplines. The interviews build on and explore in depth the findings of a recent survey on practices of data citation and reuse (Gregory, Ninkov et al., 2023). We begin by briefly situating the interview study within the context of the broader literature as well as the relevant results from this survey. We then report initial findings from the interviews related to questions of academic recognition and reward for data work (e.g. data management, sharing and reuse). We conclude by discussing next steps in our analysis and in this ongoing research project.

2. Background

The concepts of incentives, rewards, and academic recognition are centrally linked to narratives about encouraging data sharing, reuse, and citation. Citations, framed by Merton as 'pellets of peer recognition' (1988, p. 621) in the context of literature citations, are an oft-emphasized motivation for sharing and crediting the use of research data.

Several recent surveys have empirically confirmed the role of data citations in encouraging and rewarding data sharing. In a recent survey of researchers in the physical, life, and computing sciences, 92% of respondents indicated that data citations provide an important means of crediting data creators (Tenopir et al., 2020). A series of surveys, conducted by Digital Science since 2016, support this finding. According to the most recent report, approximately two-thirds of respondents across disciplines see citations to articles about the data as a motivation for sharing data, with 54% selecting data citations. (Digital Science et al., 2022).

The majority of respondents to a cross-disciplinary survey by Khan et al. (2023) also selected data citations more frequently than other options as a means of recognizing data sharing. Additionally, respondents suggested other recognition mechanisms including data badges and financial rewards. Publishers and open science advocates have proposed similar 'alternative' incentives (e.g., Kidwell et al., 2016). Others have also suggested altmetrics or usage metrics, including views and downloads, as possible ways of incentivising data sharing (Konkiel, 2020; Lowenberg et al., 2019). However, such measures are not yet widely implemented or fit for practice.

In our own survey, we asked specific questions about recognizing data as standalone research outputs and ways of rewarding data work, including preferred metrics. (Gregory, Ninkov et al., 2023; Ninkov et al., 2023). In line with the findings of previous surveys, 82% of 2,492 respondents indicated that they would find it important or extremely important to know the number of citations which their data receive. However, 70% of respondents also indicated that they would be interested in detailed 'data narratives' describing the context of how their data had been reused (Ninkov et al., 2023).

We hypothesized that there may be differences between the attitudes of 're-users' and 'non-reusers' of data regarding the recognition and reward of data work. Surprisingly, not many statistically significant differences between the two groups were identified. Across the population, 78% of survey respondents found it to be important to reward the creation of good

data documentation and workflows and 60% indicated it was important to have data recognized as standalone research outputs (Ninkov et al., 2023).

Mirroring the desire among our respondents to reward data workflows, science policy is increasingly moving away from one-dimensional indicator-based assessment of research outputs. Awareness has grown that research assessment needs to be reformed, and that it should recognize research as a *process* rather than a series of outputs, e.g. in the Open Science Career Assessment Matrix (Directorate-General for Research and Innovation (European Commission) et al., 2017). Accordingly, we have conceptualized the idea of 'data work' for our interview study: data work not only comprises data as an output, e.g., a shared dataset, but also the processes of collecting, cleaning, working with, documenting, and sharing data.

3. Interview Methods and Data

3.1. Interviews

We conducted 20 interviews with researchers across disciplines who self-identified as reusing or not reusing data. Interviews lasted approximately 60 minutes, and were conducted between September 2022-March 2023 via Zoom.

3.2. Participant recruitment and description

Seven participants were recruited via convenience sampling as part of planned pilot testing, and 13 participants were recruited from respondents to our survey who indicated that they would be willing to participate in future research. We aimed to speak with participants across a variety of research domains who either reuse data or do not (Table 1). A secondary inclusion criteria was to have a diversity of research methodologies present in the interview sample.

Table 1. Participant description

Participant	**Discipline**	**Reuses data**	**Career age (years)**	**Primary methods**
P01	Natural Sciences	Yes	31+	Mostly quantitative
P02	Humanities	Yes	6-15	Mixed methods
P03	Social sciences	No	6-15	Mostly quantitative
P04	Medical and health	Yes	6-15	Mixed methods
P05	Natural sciences	No	0-5	Mixed methods
P06	Natural sciences	No	6-15	Mostly quantitative
P07	Humanities	Yes	0-5	Mostly qualitative
P08	Social sciences	Yes	0-5	Mostly quantitative
P09	Humanities	Yes	0-5	Mixed methods
P10	Natural sciences	Yes	0-5	Mostly quantitative
P11	Natural sciences	No	0-5	Mostly quantitative
P12	Medical and health	No	31+	Mostly quantitative

P13	Humanities	No	31+	Mixed methods
P14	Humanities	No	6-15	Mixed methods
P15	Natural sciences	Yes	16-30	Mostly quantitative
P16	Medical and health	Yes	6-15	Mostly qualitative
P17	Humanities	Yes	31+	Mostly qualitative
P18	Social sciences	No	16-30	Mostly qualitative
P19	Humanities	Yes	16-30	Mostly qualitative
P20	Social sciences	No	16-30	Mostly qualitative

3.3. Interview protocol

Using an iterative process, we developed a semi-structured interview protocol for researchers who reused data, which was slightly modified for participants who do not (Gregory, Roblin et al., 2023). After the first two pilot interviews, slight changes to the question wording and ordering were made. As the meaning of the protocol and responses remained the same, all pilot interviews are included in our results. Questions centered around i) participant's own data citation and reuse practices; ii) preferences for their own data; and iii) recognition and reward of data work. This paper reports initial findings from the third set of questions.

3.4. Data and Initial Analysis

Interviews were recorded and transcribed using otter.ai, followed by manual correction and anonymization. We used a combination of deductive and inductive coding in this initial thematic analysis. Deductive codes mirrored the structure of the protocol. The authors individually read a subset of transcripts to develop an initial set of inductive codes, which were collaboratively discussed and defined. The transcripts were read and re-read; special attention was paid to questions related to reward and recognition, but also to other pertinent sections. Relevant codes were applied and analyzed as a first step in our thematic analysis. Emerging themes are indicated in bold in Section 4.

4. Initial Findings

4.1. Rewarding data and data work

Some participants believe that data should be seen as equal to other research outputs, i.e. publications, in part because of the vast amount of work required to make data available and understandable (P1, P2, P13, P16). At the same time there may be differences in this perspective even within disciplines.

> The amount of work and research that goes into carefully curating these datasets to be able to analyze in this way is 75% of the project. But my perception, again, could be wrong, is that folks in my field [..], the musicology side of the field, just don't see it as research and so they don't regard it on the same level. Whereas in Digital Humanities, they totally get how much work it is [...]. I would like the creation of and production of datasets to be treated at the same level as, you know, the publication of a peer-reviewed article. (P2)

The desire to have **data recognized as standalone research outputs** often does not match the reality some researchers face (P3). This is in part driven by a product-based mentality, e.g. in sociology, where articles and books are seen as typical products, but creating publicly available data and software is "at best, [..] looked at positively. But [..] in probably virtually no case is it evaluated as equally as, like, an article" (P8).

If data were to be considered as separate outputs in academic assessments, more education would be needed about how to evaluate the data themselves, as evaluators may not have the expertise, and because shared data have already been trimmed and cleaned (P3).

It is not always possible for participants to **separate 'their data' from other research outputs and activities**. This was clear, e.g., for a philosophy scholar whose data are the literature they use to build arguments, which makes it impossible to evaluate data independently from their publications (P7).

For another participant, synthesis, analysis, and writing are part of a broader process that cannot be meaningfully separated from the data themselves. Solely making data available - without any analysis - may be too reductive of an activity.

> To me, it's part of the whole process. But if [..] colleagues of mine would only have data as their outputs it's too easy. The part of the challenge in every field is to synthesize something from the data [..] and to come up with original ideas [..] hypotheses that fit the data, etc. [..] Just outputting data is not a merit, I think; it's part of it. (P19)

Data are also not seen as valued outputs because academic assessments focus on publications and associated citations. Assessment practices therefore pre-determine the value of both data and data citations.

> Interviewer: If you produce a dataset and you produce an article is that an equal contribution?
>
> P10: I guess I must not see it that way because I don't really care about being cited for data, but I would want someone to cite my paper [..] maybe it's because I don't really think that a hiring committee would care as much that I had a cited dataset.

4.2. Citations and metrics as mechanisms for reward

Surprisingly, participants do **not tend to conceptualize citations - to literature or to data - as being a 'reward'** in and of themselves. Rather, citations are viewed more as status symbols within a community, and data citations more as indicators of data reliability and trustworthiness (P5).

In the context of academic assessments, (literature) citations, as well as grant procurement, are seen as checkboxes or evaluation criteria to be met. According to a senior researcher, just because something is a criterion, does not mean it is an incentive. The mere fact of having something listed as an evaluation criterion may decrease researchers' internal motivations (P20).

Participants were often hard-pressed to identify potential metrics for data without prompts. One participant reflected that any data metrics were likely to lead to more work for researchers,

as they would need to be collected for application material in the context of tenure or grant proposals (P4).

While some participants saw data metrics as a potential, albeit not quite fully developed, first step to incentivize and reward data sharing, there was an **overall skepticism regarding the use of metrics and citations** in evaluations, due to embedded biases and the potential gaming of metrics (P3). Some participants proposed alternative means for recognizing and rewarding data work. Rather than adhering to the view that data citations could be a form of academic currency, they advocated for the use of monetary currency as a form of reward (P13, P15).

> I think a cash bonus from the National Science Foundation would be nice. I'm like semiserious there. It's a [..] ton of work to package this all up. And so getting a little boost would be nice or, you know, a prize would be nice. So it's not going to be citations, but [..] something that would enhance your reputation a little better. (P15)

When prompted, only few participants saw expanded forms of data documentation, e.g. descriptions in lab notebooks, as a possible avenue for assessing data. In order to be effective, evaluators of these descriptions would need both disciplinary and data management expertise (P5).

An emergent theme in the analysis is that the **quality of potential metrics** or rewards is possibly more important to participants than details about which metrics are used. We are also beginning to see that data recognition and reward are embedded within participants' communities, often **very small research communities, rather than broader disciplines**. The value of data work lies in making something that is useful to these communities (P1, P19, P2, P3).

5. Discussion

Comparing our initial findings to the literature, we see that citations may act as 'status symbols', rather than as rewards per se. We also see that for some participants, it is not possible to separate data from other research outputs or from other research and analysis activities. In line with Khan et. al (2023), our results suggest that financial rewards and prizes may offer incentives that citations cannot. Questions remain about who can best evaluate data work and for which purposes.

Our continuing analysis will examine potential differences (and similarities) between data re-users and non-reusers. We will also further develop emerging themes and relationships with research methodologies. We plan to combine the thematic analysis of the interview transcripts with a complementary analysis of the open-ended questions from our survey.

Thematic analysis is an iterative process. Presenting a step in this analysis as a research-in-progress paper allows the process itself to be more visible and open. We expect that our codes and themes will continue to evolve as our analysis continues.

Open science practices

The presented research is part of the Meaningful Data Counts-project, which aims to provide empirical evidence on data reuse and data citation practices, and to improve the understanding of the role that data play in scholarly communication. The research itself therefore contributes to a better empirical understanding of the effects of open science and its related practices.

Furthermore, the project team actively engages in open science practices, especially via sharing pre- and post-prints of published research articles, data, data analysis protocols (such as Jupyter notebooks), interview protocols, and (anonymized) survey and interview material. A research data management plan (RDMP) serves as a living document to guide our own data practices (Ninkov et al., 2020). The original and updated versions of the RDMP document and make visible the evolution of our data handling. All research outputs and the interview protocol for this study are openly published on Zenodo under https://zenodo.org/communities/meaningfuldatacounts/. Interview transcripts from this study will not be shared to protect participants. We are currently looking into ways of sharing our codebook and coding to make the thematic analysis more transparent without jeopardizing participants' anonymity.

Acknowledgments
We thank all of our interview partners in this study, as well as to our colleagues at the Scholarly Communications Lab, whose work supports and stimulates our own. We are grateful to the Alfred P. Sloan Foundation for funding this work.

Author contributions
According to the CrediT taxonomy the authors contributed to the research and paper as follows: conceptualization (KG, ABN, ER, IP, SH), data curation (SL, ER), formal analysis (KG), funding acquisition (SH, IP), investigation (KG, ER, ABN), methodology (KG, ABN, IP, SH), project administration (SL, SH, IP), resources (SH), supervision (SH, IP), validation (KG, ABN, ER, SL, IP, SH), writing – original draft (KG, IP), writing – review & editing (KG, ABN, SH, IP).

Competing interests
We have no competing interests to declare.

Funding information
The Alfred P. Sloan Foundation (Sloan Grant G-2020-12670) supported this work.

References

Alperin, J. P., Schimanski, L. A., La, M., Niles, M. T., & McKiernan, E. C. (2020). The value of data and other non-traditional scholarly outputs in academic review, promotion, and tenure in Canada and the United States. In A. Berez-Kroeker, B. McDonnell, E. Koller, & L. Collister, *Open Handbook of Linguistic Data Management*. MIT Press.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT press.

Digital Science, Goodey, G., Hahnel, M., Zhou, Y., Jiang, L., Chandramouliswaran, I., Hafez, A., Paine, T., Gregurick, S., Simango, S., Palma Peña, J. M., Murray, H., Cannon, M., Grant, R., McKellar, K., & Day, L. (2022). *The State of Open Data 2022* [Report]. Digital Science. https://doi.org/10.6084/m9.figshare.21276984.v5

Directorate-General for Research and Innovation (European Commission), Cabello Valdes, C., Rentier, B., Kaunismaa, E., Metcalfe, J., Esposito, F., McAllister, D., Maas, K.,

Vandevelde, K., & O'Carroll, C. (2017). *Evaluation of research careers fully acknowledging Open Science practices: Rewards, incentives and/or recognition for researchers practicing Open Science*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/75255

Gregory, K., Ninkov, A. B., Ripp, C., Roblin, E., Peters, I., & Haustein, S. (2023). *Tracing data: A survey investigating disciplinary differences in data citation* (preprint). Zenodo. https://doi.org/10.5281/ZENODO.7555266

Gregory, K., Roblin, E., Ninkov, A., Ripp, C., Lawrance, S., Peters, I., & Haustein, S. (2023). *Meaningful Data Counts Interview Protocol*. https://doi.org/10.5281/zenodo.7824837

Khan, N., Thelwall, M., & Kousha, K. (2023). Data sharing and reuse practices: Disciplinary differences and improvements needed. *Online Information Review*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/OIR-08-2021-0423

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Konkiel, S. (2020). Assessing the Impact and Quality of Research Data Using Altmetrics and Other Indicators. *Scholarly Assessment Reports*, *2*(1), Article 1. https://doi.org/10.29024/sar.13

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). *Open Data Metrics: Lighting the Fire*. Zenodo. https://doi.org/10.5281/zenodo.3525349

Merton, R. K. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis*, *79*(4), 606–623. https://doi.org/10.1086/354848

Ninkov, A. B., Gregory, K., Ripp, C., Roblin, E., Peters, I., & Haustein, S. (2023). *A survey of researchers on rewarding data citation and reuse*. (preprint) Zenodo. https://doi.org/10.5281/zenodo.7823626

Ninkov, A., Gregory, K., Ripp, C., Morissette, E., Harper, L., Peters, I., Tayler, F., & Haustein, S. (2020). *Research Data Management Plan for the Meaningful Data Counts Project*. https://doi.org/10.5281/zenodo.6473351

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE*, *15*(3), e0229003. https://doi.org/10.1371/journal.pone.0229003