A review of journal policies for sharing research data

Heather A. Piwowar and Wendy W. Chapman
Department of Biomedical Informatics, University of Pittsburgh School of Medicine

Accepted to ELPUB2008 (International Conference on Electronic Publishing):
Open Scholarship: Authority, Community and Sustainability in the Age of Web 2.0
This extended abstract has been archived at Nature Precedings, March 2008.
Paper in development.
Data from this study will be shared on our website.
For more information on our data sharing research please email or visit!

Keywords:
- Data Sharing
- Systematic Review
- Instructions for Authors
- Editorial Policies
- Periodicals as Topic
- Publishing/standards
- Bibliometrics
- Bioinformatics
- Gene Expression Microarrays

Abstract
Background: Sharing data is a tenet of science, yet commonplace in only a few subdisciplines. Recognizing that a data sharing culture is unlikely to be achieved without policy guidance, some funders and journals have begun to request and require that investigators share their primary datasets with other researchers. The purpose of this study is to understand the current state of data sharing policies within journals, the features of journals which are associated with the strength of their data sharing policies, and whether the strength of data sharing policies impact the observed prevalence of data sharing.
Methods: We investigated these relationships with respect to gene expression microarray data in the journals that most often publish studies about this type of data. We measured data sharing prevalence as the proportion of papers with submission links from NCBI's Gene Expression Omnibus (GEO) database. We conducted univariate and linear multivariate regressions to understand the relationship between the strength of data sharing policy and journal impact factor, journal subdiscipline, journal publisher (academic societies vs. commercial), and publishing model (open vs. closed access).
Results: Of the 70 journal policies, 18 (26%) made no mention of sharing publication-related data within their Instruction to Author statements. Of the 42 (60%) policies with a data sharing policy applicable to microarrays, we classified 18 (26% of 70) as moderately strong and 24 (34% of 70) as strong. Existence of a data sharing policy was associated with the type of journal publisher: half of all commercial publishers had a policy compared to 82% of journals published by academic society. All four of the open-access journals had a data sharing policy. Policy strength was associated with impact factor: the journals with no data sharing policy, a weak policy, and a strong policy had respective median impact factors of 3.6, 4.5, and 6.0. Policy strength was positively associated with measured data sharing submission into the GEO database: the journals with no data sharing policy, a weak policy, and a strong policy had median data sharing prevalence of 11%, 19%, and 29% respectively.
Conclusion: This review and analysis begins to quantify the relationship between journal policies and data sharing outcomes and thereby contributes to assessing the incentives and initiatives designed to facilitate widespread, responsible, effective data sharing.

Background

Sharing data is a tenet of science, yet commonplace in only a few subdisciplines. Recognizing that a data sharing culture is unlikely to be achieved without policy guidance, some funders and journals have begun to request and require that investigators share their primary datasets with other researchers. The purpose of this study is to understand the current state of data sharing policies within journals, the features of journals which are associated with the strength of their data sharing policies, and whether the strength of data sharing policies impact the observed prevalence of data sharing.

Gene expression microarray data provides a useful environment for investigating these relationships, for several reasons. Despite being valuable for reuse and costly to collect, microarray data is often but not yet universally shared. An active Microarray and Gene Expression Data (MGED) Society has developed formatting and minimum-inclusion reporting standards and actively called for journal data sharing

policies. A few centralized databases have emerged as best-practice repositories: the Gene Expression Omnibus (GEO) and ArrayExpress. Finally, the National Center for Biotechnology Information's (NCBI) Entrez website makes it easy to identify journal articles with associated datasets within GEO, allowing us to study the association between journal policies and observed data sharing practice.

Methodology
We used Thomson's Journal Citation Reports to identify journals with more than 15 articles published on "gene expression profiling" in 2006. We extracted the journal impact factors, subdiscipline categories, and publishing organizations. We consulted the Directory of Open Access Journals to identify which journals use an open-access publishing model. We downloaded the Instructions for Authors for each of these journals and reviewed them for any mention of data sharing. We classified the policies into three categories: no mention of sharing microarray data, a weak suggestion or requirement for sharing microarray data, and a strong and well-described requirement for sharing microarray data. We conducted univariate and linear multivariate regressions to understand the relationship between the strength of data sharing policy and journal impact factor, journal subdiscipline, journal publisher (academic societies vs. commercial), and publishing model (open vs. closed access).

Next, we developed a PubMed query for identifying journal articles which are likely to have generated gene expression microarray data in 2006-2007. Using the NCBI's Entrez website, we collected the total number of such articles for each journal in our cohort and the percentage of articles with links to the GEO data repository. Finally, we conducted univariate and linear multivariate regressions over the journal data-sharing prevalence percentages to understand if strength of data sharing policy was associated with observed data sharing prevalence, independently of journal impact factor, subdiscipline, publisher type, and publishing model.

Data and interpretation
Seventy journals met the selection criteria, spanning a wide range of impact factors (0.9 to 30.0, median: 4.5). A minority were published by academic societies (22/70=31%). Only 4 (6%) use an open-access publishing model. Thomson's Journal Citation Reports identified 27 subdisciplines covered by these journals; we retained the categories with more than five members: Biochemistry and Molecular Biology (n=19), Biotechnology and Applied Microbiology (11), Cell Biology (11), Genetics and Heredity (11), Oncology (19), and Plant Sciences (7). We also retained Multidisciplinary Sciences (n=4).

Of the 70 journal policies, 18 (26%) made no mention of sharing publication-related data within their Instruction to Author statements. Another 11 policies (16%) included requests or requirements for sharing non-microarray types of data (usually DNA and protein sequences), but no statement covering data in general or microarray data in particular. Of the 42 journals (60%) with a data sharing policy applicable to microarrays, 24 (34% of 70) had a general statement about data sharing and 38 (54% of 70) covered microarrays explicitly. We classified 18 (26% of 70) of these 42 policies as moderate and 24 (34% of 70) as strong.

Data sharing policy strength was associated with impact factor: the journals with no data sharing policy, a weak policy, and a strong policy had respective median impact factors of 3.6, 4.5, and 6.0. Data sharing policy was also associated with journal publisher: half of all commercial publishers had data sharing policy, whereas 82% of journals published by academic society had a data sharing policy. All four of the open-access journals had a data sharing policy. In multivariate analysis, impact factor and open access (positively) and Oncology (negatively) were statistically significantly associated with the existence of a microarray data sharing policy, with non-significant positive trends from academic society publishers and the Biochemistry&Molecular Biology subdiscipline. No multivariate associations were found to differentiate between a weak and a strong data sharing policy.

Data sharing policy strength was positively associated with data sharing prevalence in GEO: the journals with no data sharing policy, a weak policy, and a strong policy had median data sharing prevalences of 11%, 19%, and 29% respectively. Impact factor and academic society publishers were also associated with observed data sharing prevalence. We found that publisher type and policy strength interacted in multivariate analysis: policy strength increased with data sharing prevalence for journals published by commercial publishers, but not for academic society journals.

Conclusions
Our review found policies covering data sharing requirements vary widely between scientific journals, even for a specific data type with well-defined reporting standards and centralized repositories. A surprising number of journals had a policy for microarray data but not data in general, suggesting that our results may differ for datatypes for which the sharing infrastructure is not as mature. Data sharing prevalence was quite low, even for journals with very strict sharing requirements; further investigation is

needed to understand this finding. Exploring the relationship between publisher type, subdiscipline, and data sharing prevalence is also worthy of additional work, as it may help to illuminate the features of research culture which affect data sharing principles independently of mandates.

This study has several limitations: it explores a limited number of journal policies for only one type of data, the measured data sharing behavior predates the policy downloads and so may not be reflective of current behavior, the method of measuring data sharing behavior captures most but not all articles that shared data, and the policy classifications were performed by only one investigator. We also note that the reported associations do not imply causation: future study is required to determine the degree to which a change in a journal's data sharing policy will change the behavior of their authors.

Nonetheless, we believe this review and analysis is an important step in understanding the relationship between journal policies and data sharing outcomes. Policies are implemented with the hopes of affecting change, but it is often said, "You cannot manage what you do not measure." We need to understand and quantify various incentives and initiatives if we hope to unleash the benefits of widespread data sharing.