Next steps for digital data from federally funded research

Todd Vision,
Associate Professor of Biology
University of North Carolina at Chapel Hill

Heather Piwowar
Postdoctoral Research Associate
DataONE and Duke University

Submitted to the U.S. White House Office of Science and Technology Policy in response to the Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research, and available for redistribution and reuse under the terms of a CC-BY 3.0 License.

Citation: Vision TJ, Piwowar HA (2011) Next steps for digital data from federally funded research. http://bit.ly/sPyZrz

We would first like to commend the recommendations and broad strategic vision outlined by the Interagency Working Group on Digital Data in "Harnessing the Power of Digital Data for Science and Society" (IWGDD 2009). While there are many challenges in realizing the vision articulated in that document, **we think the greatest risk is inaction**. We currently tolerate a high level of wasted investment in data from Federally Funded Scientific Research (FFSR) that cannot be verified or reused, and are paying tangible opportunity costs as a result. Leaving behind us what has been called the "digital dark age" (Kuny 1998) as it applies to research outputs should be one of the top priorities for the US science policy.

We are responding as individual scientists, though we are affiliated with data archiving initiatives in the biosciences, namely the Dryad Digital Repository (http://datadryad.org) and the Data Observation Network for Earth (DataONE, http://dataone.org), and have been active for a number of years in both research and implementation aspects of data archiving. Our response will be primarily concerned with basic research data, which is where we have relevant experience.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Data archiving mandate. Our first recommendation is based on the experience of Dryad, which grew out of a grassroots initiative among a number of biology journal editors to craft a Joint Data Archiving Policy (JDAP) and to ensure that a suitable repository infrastructure existed to support the specifics of the policy before it came into effect (Moore et al. 2010). Thus, a policy mandate acceptable to the community came first and the repository infrastructure to support that specific policy was developed as a result. We recommend a similar approach at the federal agency level, first applying a strong and common data archiving policy for data arising from

basic research funding investments across all agencies, and promoting the development of technical solutions to support the policy as a second step. We offer a template for such a policy as it pertains to data associated with publications:

"This agency requires, as a condition for funding, that data supporting the results in research publications must be archived in an approved public repository. Grantees may elect to have data publicly available either prior to or at the time of publication, or may opt to embargo access to the data for a period up to a year after publication. Exceptions to this policy may be granted if justified in the Data Management Plan for data that meet certain exemption criteria [to be enumerated for each agency or program]."

We feel that a policy of this nature is broadly applicable across agencies and disciplines, which are free to make specific guidelines regarding suitable repositories, what qualifies as an exemption criterion, and the presence/length of the embargo period.

Timely Archiving. It is important that data be archived in a trusted repository at the time the research concludes (in the case of JDAP, at the time of publication), rather than shared upon request after the fact. Multiple studies have shown that disseminating data upon request does not work: researchers or data can't be found, investigators share data selectively with certain colleagues, impose unreasonable conditions on reuse, and are more likely to decline requests when there are quality issues with the data or analysis in question (Campbell 2000, Wicherts et al. 2011). Repositories can offer limited-term embargoes on data release (as discussed above) in order to protect researchers from competitive pressure where this is deemed appropriate. It is important that embargoes not be longer than necessary. We have observed that investigators publish almost all associated papers within two years of archiving their data; in contrast, published reuses of data by third party investigators continue to accumulate for years beyond that timeframe (Piwowar 2011c).

Repository Oversight. To ensure the responsible stewardship of public assets, federal agencies should coordinate policy regarding certification of trusted repositories. This would help ensure that repositories meet agency expectations for preservation processes, metadata standards, governance, financial sustainability, and so on. One lightweight model for such certification is the Data Seal of Approval (http://www.datasealofapproval.org/).

Peer Review for Data. For data associated with publications, an increasing number of journals require that data be made available at the time of peer review (for instance, those published by the Public Library of Science). This is a useful model for funders to promote, in that enlists expert reviewers and editors in ensuring data availability and re-usability. Funders (and research institutions) pay considerable sums for the service of peer review provided by publishers, and so have the right to expect a high level of service from it. The capacity to support secure, anonymous access of peer reviewers to data may be included among the expectations for trusted repositories of publication-related data.

Recognizing the Scientific Impact of Data. The research community need to be confident that publicly archived datasets are valued as first-class scholarly objects by funders and grant reviewers. Specifically, producing a highly valued dataset should contribute more to success in obtaining future funding than producing an insignificant article. We recommend the following:

1. Federal agencies should explicitly encourage the inclusion of publicly archived datasets in the credentials of grant applicants. As an example of current practice, instructions for NSF biosketches mention only that that "patents, copyrights, and software systems developed may be substituted for publications" and that Synergistic Activities may include the "development of databases to support research and education" [GPG Chapter II]. These guidelines inadvertantly imply that datasets are not scholarly products of value.
2. Agencies should systematically collect information on the datasets that have been produced by each grant through the annual and final report mechanisms.
3. Funding agencies should work to promote the infrastructure needed to support impact tracking of datasets (see Question 7). For instance, funders may require the assignment of DataCite IDs (http://datacite.org) as part of the certification criteria for a trusted data repository.

Take Data Management Plans Seriously. We recommend that all federal agencies ensure that data management plans are rigorously reviewed during evaluation grant proposals, and ensure that grant budgets include funds for the execution of the plan. Following the lead of funders such as the Wellcome Trust and the UK Research Councils, US federal agencies should issue a common statement that the costs for curation, preservation and access of research data are integral to the costs of doing research, and thus must be explicitly budgeted.

Filling Repository Gaps. Many disciplines lack appropriate repositories for data, code, mathematical models and other digital research outputs. Research funders should provide seed funding for such infrastructure. Funders should ensure that new infrastructure efforts are not chosen on the basis of technical innovation alone, but will have the capacity to be trustworthy stewards of public assets.

Research For More Effective Research. The effectiveness of data policy and infrastructure must be systematically monitored so that future decisions may be informed by evidence. Federal agencies should issue specific solicitations to researchers to collect the relevant, actionable evidence they need to make such decisions.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

We agree with the response from Cameron Neylon to the "sister" Publications RFI that intellectual property is to be treated as of as a means of incentivizing investments in research rather than as an end in itself (Neylon 2012). FFSR may at times require access to data that is confidential for legitimate commercial reasons (being trade secrets or of relevance to an undisclosed patent), but agencies supporting basic science should not fund the original

acquisition of such data, except - under current law - as it relates to an invention under the terms of the Bayh-Dole Act. Other data are kept confidential for reasons of national security, protection of personal privacy, or protection of sensitive assets (endangered species, cultural artifacts) and may legitimately be produced with federal funds. Protecting the confidentiality of data for commercial exploitation should require significant value-added investment. This is consistent with the position of the International Association of Science, Technical and Medical Publishers in the so-called Brussels Declaration, which states that "raw research data should be made freely available to all researchers" (STM 2007).

In the absence of the above reasons for confidentiality, intellectual property policy should protect the driving incentive for ongoing research, which is the availability of public funds for the conduct of science. **Researchers and universities do not require further IP as incentive to conduct FFSR, as it is already the nature of what they do.** Rather, the continuation of generous public support for FFSR is endangered by policies that allow researchers, universities, publishers or others to place unnecessary restrictions on the exploitation of outputs from public investments in research. Thus, it is in the interests of maintaining a healthly FFSR enterprise, and the corresponding commercial innovation sector that it spawns, that federal agencies ensure restrictions not be imposed where they serve no legitimate public purpose.

Furthermore, since most scientific data, being facts and not creative works, are generally not subject to copyright, a Creative Commons Zero waiver is the most suitable instrument for providing clear and nonrestrictive terms of reuse for data. (See Question 9 for a discussion of rewarding credit to data authors). **Funders should not permit restrictions on commercial use or derivative works for the outputs of FFSR**, as such restrictions stifle innovation without providing incentive for research investment.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Every discipline perceives itself to be unique. However, it is appropriate for federal agencies to articulate strong *general* principles and policies with regard to the management and dissemination of research data, while allowing for discipline-specific implementations that are sensitive to inherent differences such as data volume, machine format, complexity of human curation, long-term value, the applicability of particular metadata standards, etc. In truth, **many of the sociotechnical challenges in data management, standardization and dissemination are shared across disciplines,** particularly for the high-value portion comprising the “long-tail” (Heidorn 2008) of “small science” (Onsrud and Campbell 2007) data associated with publications.

A strong interdisciplinary ‘information community’ (in the parlance of the IDWGG) of data librarians, data scientists and educators should be cultivated. Development of such a workforce should be modeled on exemplar efforts such as the NSF DataNets, the Digital Curation Center in the UK, and the Australian National Data Service. This community is needed to help shape and support general policy and infrastructure within and among agencies, and to help spread data expertise into the educational and research communities.

At the same time, grass-roots 'communities of practice', *sensu* IDWGG, must engage disciplinary scientists in order to determine how to implement general agency policies. Such communities would be in the best position to develop the discipline-specific standards that govern the reporting of data, as well as other research products (*e.g.*, software code).

Individual disciplines and communities may wish to opt-out of general policies (*e.g.*, data archiving). This should be permitted only where the community makes a strong public case that the principles and goals are not applicable to their area, or that the same goals may be effectively achieved in a different way. **Funding agencies are the only stakeholder that can be relied upon to speak for the public interest** in the dissemination of data from FFSR when it is in conflict with the short-term competitive interests of other stakeholders in the research enterprise, and taxpayers expect their government to exercise that responsibility.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

It is frequently impossible to accurately determine the reuse value of a dataset at the time of initial reporting. Many reuses -- indeed, perhaps the most valuable ones -- are for unanticipated applications. Furthermore, we have seen in biomedical data archives that data reuses are not confined to just a few "hot" datasets but spread broadly among them (Piwowar et al. 2011).

Best-practice data archiving is less expensive than many assume. On the basis of our projections for Dryad, the marginal cost of data publication is only a small fraction (< 2%) of the cost of scientific article publication (Beagrie *et al.* 2010a, Vision 2010). For Dryad, it turns out to be much less expensive to accept all the data deposited, and to hold it indefinitely, than to make decisions regarding what to ingest or remove. By comparing the number of published articles generated by a typical grant with that enabled by typical patterns of data reuse, we have found that the modest amount of funding needed to maintain a repository like Dryad is almost certain to generate a **comparatively large scientific return on investment** (Piwowar *et al.* 2011b).

Curation at the time of ingest is a much more significant expense for many repositories than long term storage (Beagrie *et al.* 2010a) and much of the most valuable data (*e.g.,* that associated with publications) is relatively small. For example, the average dataset in Dryad is less than 5MB in size. Furthermore, cost-effective models for the publication of very large datasets are emerging, such as the recently launched BMC journal *GigaScience* (http://www.gigasciencejournal.com).

Finally, the burden of archiving on individual investigators should not be overestimated. Although new practices invariably generate anxiety, Whitlock (2010) and others have demonstrated that **basic guidelines for good data archiving and reuse can be made simple and intuitive**.

(5) How can stakeholders (*e.g.*, research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

While many stakeholders must play a role, we wish to emphasize the crucial role of funders in monitoring the effectiveness of agency policy and individual adherence to data management policy..

Funders should recognize the depth of the need to raise awareness about expectations regarding data management. As part of an ongoing study (Piwowar 2010b), we asked corresponding authors of biology articles about their funders' policies on data archiving: **27% of the investigators responded that they didn't know if their funder had a data archiving policy** (*n*=1500; 39% said their funder had no policy, 10% said their funder required online public archiving). At the same time, consistent with other studies, respondents overwhelmingly believed that mandatory public online data archiving is the "right thing to do." It appears that funders are missing an opportunity to reinforce the best instincts of their funded researchers.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

A variety of funding mechanisms will be needed to provision support for data services (curation, dissemination, migration, replication, *etc.*), given the heterogeneity of all FFSR data. The desired model specifically for long-tail small-science data will (a) provide for some direct investments in repository research and development, (b) scale with the volume of service provided, (c) facilitate the operation of an efficient market for data services, and (d) enable investments in shared international infrastructure.

Investment in repository infrastructure. There will be an ongoing need for direct investment both to support research and development needs of existing repositories and to fuel the development of new resources for datatypes or disciplines lacking existing solutions. When it is necessary for the funding model of data services to be dependent on grants, these should be evaluated based on criteria relevant to infrastructure, rather than solely innovation.

Scalability. Scalability of finances for data services can be achieved by including the costs for data management within research budgets, and allowing individual awardees to direct those costs as needed for their project.

Market for data services. Similarly, if funds are allocated to services on a project-by-project basis, that establishes a competitive market for data services within which those of greatest value receive the most support.

International coordination. Insofar as direct funding from agencies is required for certain datatypes, the greatest challenge will be to develop mechanisms for multinationational investment in shared resources (*e.g.*, such as that used by ELIXIR, http://www.elixir-europe.org/).

While the costs of supporting data infrastructure are tangible, funding agencies should also attempt to understand the hidden economic costs of not having infrastructure to support investments in FFSR data, so that the cost and benefits of investment can be fairly compared.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Evaluating plans and tracking compliance are important. Evidence suggests the NIH requirement for data management plans is generally considered toothless (Tucker 2009) and has made little difference to data availability (Piwowar et al. 2010, Piwowar 2011f).

Disseminating research results is the responsibility of the funded researcher. In the short term, better mechanisms for tying outputs to funding are required. As mentioned in response to Question 1, we recommend that annual reporting require researchers to list publicly available datasets derived from FFSR. Research results that have not been disseminated in accordance with policy should not be acknowledged as output of the grant for the purposes of evaluation.

Federal agencies should enthusiastically collaborate with publishers, libraries, universities, and other stakeholders in promoting technological solutions that will promote trackability of research data products and the reuse of those products, such as DataCite (http://datacite.org), ORCID (http://orcid.org) and VIVO (http://www.vivoweb.org).

In the longer term, there is great potential in moving beyond compliance monitoring to fostering enthusiastic reporting through incentives. The impact of both traditional and non-traditional research products (articles, datasets, code, blogs, preprints, slidedecks, etc.) can be collected for investigators, research groups, institutions, grants, and even whole grant programs using traditional and non-traditional metrics (citations, views, downloads, bookmarks, tweets, etc.). These statistics can then be used to demonstrate the impact of individuals and organizations during evaluations, **providing an incentive for products other than only publications to be reported**. We have been working, with others, on a prototype project to demonstrate this potential (http://total.impact.org); an example showing the "impact report" for one of us (HP), including download metrics for archived datasets, is shown here: http://total-impact.org/report.php?id=SIIysw

Achieving compliance through incentives is currently hampered by our closed scholarly communication infrastructure. Existing citation indexing systems do not index datasets -- even when cited in a paper's reference list (Piwowar 2011d), do not make citation data available for innovative impact mashups, and can not be improved through open source contributions. Barriers to text mining the scientific literature are also significant because the context of a citation contains important information about the nature of the attribution. Future funder initiatives could help address these barriers.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

As we mentioned in Question 7, opening up our scholarly communication infrastructure would make the output of funding more generative - more able to produce innovation (Zittrain 2008).

We recommend licence terms for data or other research outputs that do not exclude commercial and derivative products; this will ensure that outputs from FFSR are available for innovative scientific applications and the creation of new business opportunities. Specifically, **nonrestrictive access to all research outputs** (papers, data, code, etc) would permit machine access, text- and data-mining, data integration, third-party curation, and other value-added services.

We recommend that access and preservation of software from FFSR be given the same policy attention as data. Almost all digital data is collected, and statistics are computed, through the execution of software code. Access to the code associated with a dataset increases the comprehension, re-usability, and replicability of that dataset and its analysis.

The accessibility of the scientific literature is also key to fully leveraging associated datasets. The most valuable piece of metadata about a dataset is the publication that describes its original collection and analysis. When this metadata is not available without restrictions on copying and reuse, it limits the reusability of that dataset.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Citation formatting flavors notwithstanding, the scholarly community has fairly efficient and effective norms for citing published papers. Because community norms for citing datasets have been lacking, investigators have adopted a variety of conventions for providing attribution to the authors of datasets (Weber et al. 2010, Enrique et. al 2010, Piwowar et al 2011e). Few stakeholders provide guidance on data citation (Weber et al. 2010); journals, unsurprisingly, are leading the way whereas funders have provided very little guidance thusfar. This diversity of citation practice makes it difficult to track data reuse.

Nonetheless, **even in the current chaotic environment, investigators receive benefit for archiving data**. Several analyses, in diverse disciplines, have found that studies which make their data publicly available receive more citations than similar studies which keep their data private (e.g. Piwowar et al. 2007). In an survey of 1500 corresponding authors in biology, 45% of authors reported that their datasets have been used and formally cited; only 21% said their datasets had been used without citation (Piwowar 2010b).

Data citations standards, coupled with the impact tools we discussed in Question 7, will make collection and interpretation of data attribution simpler, quicker, and more accurate. Various initiatives are underway to establish and promote standard practices for attributing data. The predominant approach, and our recommendation, is to cite datasets very similarly to how we cite papers. This has the obvious advantages of familiarity and easy integration into scholarly communication tools (we hope!). It also provides a distinct author list to fully recognize the names of the people responsible for the data product. Others are working on “data publications” that combine data archives with a data-centric article wrapper.

Standards for Interoperability, Reuse and Repurposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma *et al.,* 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

The array of existing data standards is already very large, and since it evolves with advances in both experimental technology and information technology, it will never be complete. Community interest in the development of a new standard can be stoked by the existence of a critical mass of previously unavailable data with potential for reuse. Thus, funding agencies can indirectly promote new standards by their support for data repositories in particular fields.

It is the role of communities of practice to determine when such standards are applicable to a particular datatype or repository, while it is the role of funders to ensure these are appropriately applied to each project. Rigorous evaluation of data management plans during proposal evaluation can help to ensure this outcome.

It is also the role of the funding agency to ensure that communities of practice are inclusive, transparent, responsibly governed, non-duplicative, and follow best practices in the development of standards. An excellent model of a community of practice working to promote dialog between funders and stakeholders regarding data standards is the Biosharing project (http://www.biosharing.org/standards). Umbrella coordination projects such as this help to rationalize the bewildering diversity of data standards and ultimately help facilitate adoption of appropriate standards by the research community and its repositories.

Two publication outlets for communication of digital data standards in biology include the Open Data Standards section of BMC Research Notes (http://www.biomedcentral.com/1756-0500/3/235/) and the curator collection at PLoS (http://www.ploscollections.org/article/browseIssue.action?issue=info%3Adoi%2F10.1371%2Fissue.pcol.v03.i05). Both outlets, unfortunately, suffer from a relatively low rate of submission.

To help guide public investment in standards efforts, **we recommend federal agencies encourage research into the economic tradeoffs inherent in standard development.** Standards have benefits in ease of data reuse, but also incur costs in development, maintainenance, and compliance. We need to understand better how to balance these costs and benefits.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

See response to Question 10.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

In individual investigator-driven science, international coordination occurs fairly naturally within disciplines provided that funding resources are available for research networks, and that these networks allow international participation. Coordinated calls for funding relevant to digital data standards among agencies in different countries would facilitate the international distribution of effort, and increase the efficiency of investments within each country. One useful model for international cooperation among funding agencies that share a vision of 'making a layer of scholarly and scientific content openly available on the Internet' is the Knowledge Exchange (http://www.knowledge-exchange.info), which includes partners in Denmark, Germany, the UK, and the Netherlands.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Dryad specializes in the linkage between publications (from bioscience and biomedical journals) and associated data, and currently hosts data from over 100 different journals. The model for how to achieve this linkage was developed based on extensive consultation with Dryad's and board and stakeholder community (including journal editors, society officers, publishers, researchers, librarians, and technologists). Some of the elements of Dryad's approach which we think are generally applicable include:

1. Providing a forum for development of shared data policy among journals, including policies regarding embargoes, data citation, terms of use, metadata standards, *etc.*
2. Participation of journal editors, scientific societies, and publishers in governance of the repository and developing its feature road-map.This enables responsiveness to the particulars of journal policies and procedures.
3. Direct involvement of journals, societies and publishers in repository sustainability through payment of membership and deposit fees.
4. Technical integration between article submission and data submission, including direct communication of metadata between repository and publisher.
5. Assignment of DOIs to data through DataCite. Inclusion of article DOIs in online data records and data DOIs within articles.
6. Professional curation to ensure quality metadata, file validity, and preservation actions such as format migration.
7. Exposing metadata through multiple standards, including pushing to third party indexers to enable discovery through standard bibliographic services.
8. Support for secure and anonymous peer review of data associated with articles prior to acceptance.
9. Clear policies regarding data citation, and promotion of technologies to support data citation tracking.
10. Leveraging the preservation infrastructure used by traditional publishers (e.g. CLOCKSS http://www.clockss.org)

References

- Beagrie N, Eakin-Richards L, Vision TJ (2010a). Business Models and Cost Estimation: Dryad Repository Case Study, Proceedings of the 7th International Conference on Preservation of Digital Objects (iPres) 365-370.
- Beagrie N, Lavoie B, Woollard M (2010b), Keeping Research Data Safe 2, Final Report
- Campbell E (2000) Data withholding in academic medicine: characteristics of faculty denied access to research results and biomaterials. Research Policy 29, 303-312.
- Enriquez V, Judson S, Weber N, Allard S, Cook R, Piwowar H, Sandusky R, Vision TJ, Wilson B (2010 Data citation in the wild. Nature Precedings.
- Heidorn, PB (2008) Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2).
- Kuny T (1998) A Digital Dark Ages? Challenges in the Preservation of Electronic Information. International Preservation News, No. 17.
- Michener W, Vieglais D, Vision TJ, Kunze J, Cruse P, Janeé G (2011) DataONE: Data Observation Network for Earth — Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. D-Lib, doi:10.1045/january2011-michener
- Moore AJ, McPeek MA, Rausher MD, Rieseberg L, Whitlock MC (2010) The need for archiving data in evolutionary biology. J Evol Biol. 23, 659-60.
- National Science Foundation, Grant Proposal Guide, Chapter II
- Neylon C (2012) Response to Request for Information - FR Doc. 2011-2862
- Onsrud, HJ, Campbell J (2007) Big Opportunities in Access to "Small Science" Data, CODATA Data Science Journal 6, OD58-OD66.
- Piwowar H, Chapman W (2010) Public sharing of research datasets: A pilot study of associations. Journal of Informetrics 4, 148-156.
- Piwowar H (2010b). Study on Impact Of Journal Data Policies: Towards understanding the impact of journal data archiving policies on attitudes, experiences, and practices of authors. Recruitement ongoing.
- Piwowar H, Vision TJ, Whitlock M (2011b) Data archiving is a good investment. Nature 473, 285.
- Piwowar H (2011c) A New Task For NSF Reviewers: Recognizing The Value Of Data Reuse. Research Remix blog.
- Piwowar H (2011d) Tracking Dataset Citations Using Common Citation Tracking Tools Doesn't Work. Research Remix blog.
- Piwowar H, Carlson J, Vision TJ (2011e). Beginning to Track 1000 Datasets from Public Repositories into the Published Literature. ASIS&T 2011.
- Piwowar H (2011f) Who shares? Who doesn't? Bibliometric factors associated with open data archiving. PLoS ONE 6(7): e18657.
- STM, (2007) Brussels Declaration on STM Publishing http://www.stm-assoc.org/public_affairs_brussels_declaration.php (STM 2007) http://ftp3.dns-systems.net/~stm/2007_11_01_Brussels_Declaration.pdf?PHPSESSID=5cd58816ea0a9087be865c6cf046626f
- STM (2007) Brussels Declaration on STM Publishing
- Tucker J. (2009) Motivating Subjects: Data Sharing in Cancer Research. PhD Dissertation, Science and Technology Studies, Virginia Tech.
- IWGDD (2009) Harnessing the Power of Digital Data for Science and Society.

- Vision TJ (2010) Open Data and the Social Contract of Scientific Publishing. BioScience, 60(5):330-330.
- Weber N, Piwowar H, Vision TJ (2010) Evaluating Data Citation and Sharing Policies in the Earth Sciences. ASIS&T 2010.
- Whitlock M (2011) Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution, 26 (2): 61-65.
- Wicherts JM, Bakker M, Molenaar D (2011) Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PLoS ONE 6(11): e26828.
- Zittrain JL (2008). The Future of the Internet And How to Stop It. Yale University Press.