IJDC | *Peer-Reviewed Paper*

Data Producers Courting Data Reusers: Two Cases from Modeling Communities

Jillian Wallis
University of California

Abstract

Data sharing is a difficult process for both the data producer and the data reuser. Both parties are faced with more disincentives than incentives. Data producers need to sink time and resources into adding metadata for data to be findable and usable, and there is no promise of receiving credit for this effort. Making data available also leaves data producers vulnerable to being scooped or data misuse. Data reusers also need to sink time and resources into evaluating data and trying to understand them, making collecting their own data a more attractive option. In spite of these difficulties, some data producers are looking for new ways to make data sharing and reuse a more viable option. This paper presents two cases from the surface and climate modeling communities, where researchers who produce data are reaching out to other researchers who would be interested in reusing the data. These cases are evaluated as a strategy to identify ways to overcome the challenges typically experienced by both data producers and data reusers. By working together with reusers, data producers are able to mitigate the disincentives and create incentives for sharing data. By working with data producers, data reusers are able to circumvent the hurdles that make data reuse so challenging.

Received 29 January 2014 | *Accepted* 26 February 2014

Correspondence should be addressed to Jillian C. Wallis. GSE&IS Building, Room 215, Los Angeles, CA 90095. Email: jwallisi@ucla.edu

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

A recent White House Office of Science and Technology Policy memorandum (Holdren, 2013) is the latest of a long list of policies – including requirements from funders, publishers, institutions and disciplinary norms – that encourage publicly funded researchers who produce data to share those data with others. Given pressure from the top-down, all data producers should be sharing all their data, but the reality is that very few researchers are sharing anything at all.

Data producers are willing to share data when presented with the opportunity to do so, but there are some obstacles that still need to be overcome. Sharing data is difficult. Data producers are being asked to manage their data for unknown and unknowable reuses by future reusers. Whether reused by members of their own domain or from another domain entirely, the data are not going to be reused for the reason they were initially collected. Standard metadata is only guaranteed to support those who are from a specific domain, or even from a specific community of practice. As such, the metadata needs, data formats and other ways the data producers can support data reusers cannot be predicted. Making data available for every possible reuse would be an infinite task, and arbitrarily choosing an audience for the data is likely to be wasted work. Finally, there are very few requests for data to be shared. Given a lack of demand for data, researchers are understandably dubious about going through the trouble of adding metadata when they see very little benefit to doing so. Data sharing and data reuse are really two sides of the same coin. We need to overcome the hurdles to both data sharing and data reuse in order for these data sharing policies and memoranda to be meaningful.

One way data producers can use to avoid these obstacles is to create opportunities for known reuse. This approach narrows the infinite field to a prospective reuser or reuser population so that the data producer can find out the best way to mark up data and support the interpretation process for them. At the same time, the data producers are able to make sure that the data reusers meet data sharing conditions. The effort on the part of the data producer to make their data reusable is rewarded with more immediate consequences, such as the ability to report tangible reuses of their data to their funder. This paper will present two recent cases of data producers reaching out to prospective reusers, how the reaching out was accomplished, and the success of each endeavor.

Background

Data sharing and data reuse practices vary along with the motivations and challenges faced by parties on both sides of the data sharing exchange. What follows is and overview of data sharing practices on the part of data producers, what it takes to reuse data on the part of data reusers, and some of the challenges to both data sharing and reuse.

Sharing Research Data

The majority of research datasets collected fall along the long tail (Heidorn, 2008) in that they are used for the purpose for which they were collected and then they are not used by anyone else. These include one-off experiments, short time series, long time series that are geographically constrained, and all sorts of datasets that do not easily aggregate with other datasets because of a lack of commonalities. Those datasets that do not fall in the long tail are used multiple times. In some domains there is an expectation that data collected for a single purpose will be made available to the larger community where they can be reused (Nelson, 2009). Data from high energy physics, seismology, astronomy, social science surveys, and genetic biodiversity data are relatively easy to access for comparison, aggregation and other forms of reuse because the data lack specificity of purpose (Borgman, 2012). In these cases, there are well-established data standards and repositories, and the punishments for non-compliance with policies are more tangible.

Regardless of whether data are collected along the long-tail or not, data producers are willing to share their data and have done so when asked. In a survey of over 1,300 people, Tenopir et al. (2011) found that 75% of researchers have shared their own data at some point, and 6% have made all of their data available. In a complimentary study from Wallis et al. (2013), based on 43 interviews and ethnography performed at an NSF-funded research center, everyone interviewed was willing to share data, but only half of them had been asked to do so. There are three main ways data producers share their data: through personal exchange, deposit in a repository, or posting the data online (Tenopir et al., 2011; Wallis et al., 2013). Personal exchange is the most common method of data sharing, and it is likely that this is because of the conditions under which these researchers were willing to share. Some of the most popular conditions the researchers provided were: retaining the first right to publish from their results, receiving proper attribution as the source, wanting the requestor to be known to them, the ability to negotiate sharing in advance of exchange, etc. These conditions are all easier to ensure through a personal interaction with the data reuser than through a system that may not encode their conditions (Wallis et al., 2013).

Reuse of Research Data

Data from along the long tail are rarely reused, and are not well studied. A couple studies of long tail data reuse come from ecology (Zimmerman, 2007; 2008) and earthquake engineering (Faniel and Jacobsen, 2010). These studies are from situations where the data being reused does not cross any disciplinary boundaries.

In her studies of ecologists who reuse ecological data, Zimmerman (2007, 2008) found that the reuse of ecology data is not a trivial task. Ecologists must rely on their formal training and informal knowledge from their own time spent in the field to support “finding, acquiring, and validating data collected by others” (Zimmerman, 2007). Having spent time in the field allows the data reuser to visualize the process of data collection and be critical of the methods being used by the data producer. She also found that standardization of data is one way to make it more ‘transportable’ across boundaries, but this is not enough to overcome a lack of formal and informal ecological training (Zimmerman, 2008). When it is difficult for data users to rely on their knowledge and standard, they will instead use their personal knowledge of the person who collected the data to assess the trustworthiness of the data.

Faniel and Jacobsen (2010) studied how earthquake engineering researchers “assess the reusability of colleagues’ experimental data for model validation.” They found that the reusers needed to assess the relevance of the data, understand the data, and determine whether the data were trustworthy in order to use them. Earthquake engineers primarily use documentation to perform this assessment, although in most cases consulting with colleagues provides complementary information about the data. Unlike ecology, where data can be trusted if the data producer can be trusted, in the case of earthquake engineering trust is based on how well the data are documented.

Challenges to Data Sharing and Reuse

Although data sharing occurs, it occurs less frequently than anyone would care to admit. Faniel and Zimmerman (2011) point out that there is an assumption in the literature that making more data available ensures reuse, but research has demonstrated that data reuse is difficult within the same discipline. Zimmerman (2008) provides a list of challenges to data sharing, that includes "issues of data ownership, a lack of incentives for scientists to share; technical hurdles related to incompatible hardware, software and data structures; and costs to document, transfer and store data." Borgman (2012) lists disincentives faced by a data producer who is willing to share data, such as "a lack of reward or credit for sharing, the significant investment of time required to document data in reusable forms, and concerns for misuse or misinterpretation of data." According to Borgman, researchers are more willing to share data within their own discipline, as their colleagues have the expertise necessary to interpret the data and are less likely to misuse them.

Edwards (2010) uses the term 'data friction' to refer to "the costs in time, energy, and attention required simply to collect, check, store, move, receive, and access data." This refers to all the processes of acquiring data from a data producer, which may require that the reuser cross institutional, technological, or even national boundaries. Once the data is in hand, the reuser is also subject to 'metadata friction', or the problems experienced when trying recover to the context of data creation (ibid.). Mayernik et al. (2011) unpacks metadata friction into various metadata frictions, including gauging the correct audience for the metadata and metadata standard proliferation.

All these frictions add challenges to finding common ground when interdisciplinary researchers share data, or 'science friction' (Edwards et al., 2011). Overcoming these frictions is above and beyond the responsibilities of researchers. Their job is to perform research and publish using their own data, not to "describe them for the benefit of invisible, unknown future users, to whom they are not accountable and from whom they receive little if any benefit" (ibid.). The fact that the effort required to overcome these frictions has no immediate benefits increases the effects of the frictions (Mayernik et al., 2011). Faniel and Zimmerman (2011) remind us that the challenges experienced during interdisciplinary data sharing are magnified when non-scientists, such as policymakers, reuse scientific data.

Methods

The data presented here were collected during ethnographic observation at two events: the 2013 Community Surface Dynamics Modeling System (CSDMS)[1] Annual Meeting[2] and the National Climate Predictions & Projections Platform (NCPP)[3] Qualitative Evaluation of Downscaling (QED) Workshop[4]. Both events are introduced briefly below, and described in more detail in the findings. The author was invited by the

[1] CSDMS: http://csdms.colorado.edu

[2] CSDMS Annual Meeting: http://csdms.colorado.edu/wiki/CSDMS_meeting_2013

[3] NCPP: http://earthsystemcog.org/projects/ncpp/

[4] NCPP Qualitative Evaluation of Downscaling Workshop: http://earthsystemcog.org/projects/downscaling-2013/

organizers to participate in these events to evaluate the efficacy of each, and reports were provided to both groups with findings from their event. In addition to observation, meeting talks and materials were consulted. Throughout the paper, the real names of organizations, individuals, and projects are used so that they can serve as role models for other researchers who wish to court data reusers.

CSDMS is a project coming out of the Earth system modeling domain, consisting of a core group of roughly ten developers who are working on cyberinfrastructure to support a 1,000-member community. The developers are working on a model repository, where users can access models, data, and other resources, in addition to a framework that supports the coupling of models. The models in the repository range from describing very small phenomena, such as how particles form an alluvial fan, to models that describe the entire ocean surface system. The CSDMS Annual Meeting is a three-day meeting with a mixture of keynotes, student presentations, special interest group breakouts and training sessions. For the purposes of this paper, a keynote presentation from someone outside of the community is specifically highlighted.

The NCPP is an organization comprised of climate model researchers who are interested in making their data available to researchers from other domains. The QED workshop brought together people who work with climate models and people from other disciplines who are interested in using climate model outputs for a five-day meeting with a mixture of keynotes, panels, group discussions, breakouts, and informal down time between sessions.

Findings

Each of the two observed cases are described, highlighting the situation, the data producers and the data reusers, their motivations, and outcomes of their interaction.

CSDMS: An Experimentalist Reaching Out to Modelers

Dr. Wonsuck Kim, an experimental geophysicist from the University of Texas, gave a talk entitled ‘Building a Network for Sediment Experimentalists and Modelers’ (Kim, 2013) to a group of roughly 120 Earth system modelers at the CSDMS Annual Meeting. During the talk, Kim presented examples of experimental data being collected and discussed what would need to happen in order for modelers to be able to use the experimental data.

The objective of the experimental, geophysical research is to develop new techniques and materials to better simulate surface processes at scale. This is what gets published and receives academic credit in this community of practice. Data about the phenomena being recreated are a by-product of the research that would be of use to other researchers. The members of the CSDMS community are primarily surface modelers, developing models of the very processes the experimentalists are trying to reproduce. Data from experiments can be fed into the models to train them or to benchmark models against observational data.

The experiments performed by Kim and other experimental geophysicists involve the creation of miniature versions of larger-scale natural phenomena, such as deltas, river meanders, riverbed forms, erosion and tsunamis. Materials such as alfalfa and walnut sawdust are used to mimic larger-scale sediments. Once an experimental setup is built, the researchers vary dimensions of the processes, such as using the same water flow rate in a river delta simulation, but changing the tilt angle for each run. Each experiment is captured through a variety of methods – including image capture, video, x-ray tomography and topographic scanners – which make up the data being collected. In addition to these data, metadata about the experiment is recorded, including the initial conditions of the experiment such as the materials being used and rates of any system perturbations. The experiments are designed to create similar phenomena to those found in nature so that the researchers can understand the initial conditions that lead to each phenomenon. To demonstrate how well the experiments are recreating real world phenomena, Kim presented images and video of the experiments next to images from actual deltas and rivers.

Some of the experiments Kim presented were his own, but the majority of experiments presented were from colleagues at other labs and even other institutions. Kim was giving this talk as a representative of the experimental geophysicist community. He explained that the experimental geophysics community is at a point where they want to converge. They want to be able to perform a systematic comparison of their results and answer some grand-challenge questions, such as 'What is the driving equation behind all river flow processes?' In order to begin this process, they are planning to make a repository that is accessible to the public, called the Sediment Experimentalists Network Knowledge Base (SEN KB).

Currently, the data collected during experiments is out in the world, but it is difficult to find. Kim characterizes the data as falling in the long tail of research because it tends to come from one-off experiments rather than long-term or large data collection efforts. Someone who wanted to use the data would need to know that the data existed, and who created, it in order to gain access.

To make their data more accessible, Kim and other representatives from the geophysical community are reaching out to their prospective data users. Kim acknowledged that the experimental and modeling communities had different metadata needs, and one presentation slide stated: "we need to your inputs [sic] for best practices and metadata to effectively share data." He explained that they would use this input to inform education of the experimental students, as well as develop new standards for SEN KB. Not only did Kim indicate that they wanted to share existing data sets with the modeling community, but that they were interested in collecting new data for the modelers. He stated: "you need to tell us the kind of data you need." Running experiments for other people is of benefit to his own research, as it pushes him to perform new experiments, and he can demonstrate his data sharing and collaboration efforts to his funders. The talk ended with an open invitation to the modelers to form collaborations where the experimentalists perform experiments that the modelers need to understand specific phenomena. Kim's talk created a lot of buzz during the meeting, and he has since had modelers approach him for collaborations.

Kim was an invited keynote at the CSDMS Annual Meeting. Members of the core CSDMS team had met him at prior meetings and see a value to his work. His talk differed from the other keynotes, which were all about new models that had been developed, research using models, and research using multiple models coupled together. Although this talk was different from the other keynotes, it fitted with one of the areas the core team has added to the CSDMS community agenda, that of 'tracking uncertainty.' Tracking uncertainty, in this case, applies mainly to model uncertainty, but can also apply to uncertainty in data. Knowing the data source is a step towards tracking data uncertainty.

QED: Modelers Reaching Out to Model Output Users

The Qualitative Evaluation of Downscaling workshop was the first workshop organized by the NCPP, a group of climate modelers who are concerned with making model output data available to researchers beyond climate science. As with other scientific researchers, climate scientists are required by policy at multiple levels to make their data available for reuse, but they are hesitant to do so because climate science data are incredibly complex and difficult to interpret. For instance, only looking at the lower troposphere data from the tropics where satellite measurements are less reliable due to bleeding from other layers of atmosphere gives the impression that there is an overall cooling trend (Nuccitelli, 2013). Given that climate modelers are under close scrutiny from politicians and the public, they want to have control over the data exchange process to ensure appropriate use and understanding of model output data. Even the climate modeling terminology can be an issue. For instance, climate modelers avoid using the word 'prediction' because it implies too much strength, instead they use 'projection' which includes more ambiguity. The entire workshop was designed to bring together the people who are creating climate model outputs and the policy makers who use them.

Everyone who participated in the workshop had been invited by the organizers or recommended to attend by a colleague. Participants in the workshop were comprised of representatives from both sides of the model output reuse process. On the climate model side there were modelers, downscalers and tool developers. On the model output reuse side there were members of the agriculture, ecosystems, human health and water resource management communities. Model outputs are used to predict growing seasons and the economic impact of climate change, migratory patterns of species for land management, and future health outcomes, such as flu season and mortality rates.

The week-long workshop was split up into keynote presentations and panels, and breakout discussions. Presentations and panels covered a variety of topics from both the climate model and output reuse sides. Climate modeler presentations covered the current state-of-the-art of climate models and downscaling, and limitations of climate models, downscaling, and observational data. Model output reuser presentations covered how climate model outputs are currently being used, and problems reusers run into when using climate data. A significant problem reusers ran into was a lack of metadata that would allow them to evaluate the fitness of the model for their application. In order to reduce complexity in global climate models, modelers reduce the complexity of surface phenomena. In many models, all of the world's rivers are collapsed into five giant rivers with each river located on a different continent, and bodies of water within a continent, such as the Great Lakes, have been reduced to marshland. This information is incredibly useful for decision makers looking to evaluate whether a model can be used in a specific geographic region, but would not be included in model metadata.

Breakout discussions brought together participants by output reuse group for more in-depth discussions, with the modelers floating between groups. The groups were charged with developing a 'nutrition label' for models that would contain the metadata the reusers would need to know in order to use or trust the data. During lunch and coffee breaks there were interesting conversations between the modelers and the data users. Output producers and users had conversations that felt like a speed-dating interaction, where the output users would tell the producers how they intend to use the data and then ask whether the modeler's model would be usable for this application. If the interaction proved successful, the participants exchanged contact information; otherwise they moved on to talk with other participants.

The meeting culminated in a common approach for how to move the NCPP forward in their goal of making model outputs useful to other communities. The approach has three main components: metadata to support accessibility, education of output users, and fostering a new workforce to bridge between modelers and output users.

- **Metadata standards** will pool the requirements established by each user group to support the efficient search of available datasets, identification of applicable datasets, and quality assessment of those datasets. The metadata must stand in for considerable climate domain knowledge that members of the various user groups would not be able to bring to bear.
- The metadata is then supported by **education of the output users** regarding appropriate use of climate model outputs and other climate data. Some basic rules-of-thumb were hashed out during the meeting, such as “use a dozen models if you can” and “don't pick the best models, cull the worst.” More in-depth education about output use would be domain-specific and drafted by the members of the domain with input from the climate modelers. For instance, the ecosystems group discussed what would be the best vehicle for transmitting this information to their larger community. They settled on writing a paper for a practitioner journal, and made writing commitments before leaving the meeting.
- Where the metadata and education failed, **‘climate translators’**, people with their feet in both the data producer and data user communities, would be able to bridge the gap. Climate translators could help users to identify possible data sources and assist with the interpretation. Individuals already serve this function, and roughly a quarter of the participants in this workshop identified themselves as such, but until this workshop there had not been a specific title for what they do.

In an exit survey of the QED workshop, participants were asked what they had hoped to get out of the workshop and whether the workshop was useful to them. The majority of the participants coming from the modeling side had hoped to learn more about the reuses to which their data were being put and any feedback the reusers might have about their models or model outputs. Similarly, the majority of the model output reusers had come to the workshop in order to learn more about models, model outputs, and their respective limitations. Participants from both sides found the workshop to be very useful because their hopes for the workshop were fulfilled.

Discussion

By engaging with a specific reuser population, the data producers from both cases were making the reuse of their data more predictable. Rather than dealing with an infinite set of unknown reusers and ways to reuse the data, the data producers were dealing with a concrete, known user population, in the first case Earth system modelers and in the second a series of policy-making communities. By working together, the data producers and data reusers are able to overcome many of the challenges faced by each party in data sharing and reuse, such as determining appropriate metadata standards, various frictions, and incentivizing the process for both the data producer and data reuser.

In both cases presented above, the data producers were able to discuss metadata standards and evaluation metrics with the reusers themselves. Metadata frictions, including gauging the correct audience for the metadata and the proliferation of metadata standards (Mayernik et al., 2011), are each mitigated in these cases by knowing the metadata audience and working together to determine the correct metadata standard. In both the ecology and earthquake engineering communities metadata standards play a key role in giving reusers access to the data (Zimmerman, 2008; Faniel and Jacobsen, 2010). In the first case, Dr. Kim introduced a repository, SEN KB, currently under development by experimentalists that will incorporate metadata to assist the modelers. And in the second case, the QED workshop was the beginning of a conversation to develop 'nutrition labels' to capture the metadata about models and model outputs necessary for evaluation and interpretation by various reuse communities.

The institutional, technological and disciplinary boundaries between the data producers and data reusers that contribute to data friction (Edwards, 2010) are being overcome by reaching across those boundaries to find common ground, something that is necessary to overcome science friction (Edwards et al., 2011). In both cases the data producers were working with members of other communities to establish data sharing and reuse connections. Although in the CSDMS case, both parties are members of the geophysical community, there is a deep divide in methods between experimentalists and modelers that acts like a disciplinary boundary. During the QED workshop, keynote and panel talks from both data producers and reusers established better understanding of the challenges faced by both parties and provided the groundwork for future collaboration. It should be noted that the NCPP climate modelers are also attempting to reach out to non-science communities, mainly policy-makers – a process that is an order of magnitude more difficult than reaching out to other scientists (Faniel and Zimmerman, 2011). The process will still be difficult, and carry costs to all parties involved, but there is a known outcome of the effort required.

From prior work we see that the lack of incentives and the actual disincentives to sharing make doing so very undesirable. Challenges to data sharing on the part of the data producer include data ownership, technical hurdles, and costs (Zimmerman, 2008), as well as a lack of credit and concerns about misuse (Borgman, 2012; Edwards, 2010). These challenges can be seen in the conditions under which data producers are comfortable sharing data, such as retaining the first right to publish from their results, receiving proper attribution as the source, wanting the requestor to be known to them, and the ability to negotiate sharing in advance of exchange (Wallis et al., 2013). Data reusers are similarly disincentivized to reuse data because of the time and resources necessary to make them usable, as mentioned above. In both cases, the benefits of the data sharing exchange were tipped to outweighed the disincentives.

When the data producer identifies and works with a data reuser, as with these cases, the disincentives are reduced. The data producer will be able to have their conditions for sharing met, as they are able to negotiate credit, ownership, first rights to publish and other intellectual property concerns prior to data reuse. Although this was not explicitly brought up in either case, it is likely to come up in the future of both interactions, and will likely happen before any data are shared. Similarly, the misuse of data can be averted by working closely with data reusers. Climate modelers at the QED workshop were very explicit about how easy it is to misuse climate data and both sides came up with strategies to avoid misuse, including the basic rules of thumb, best practices by reuser group, and the development of climate translators to assist with access and interpretation. Being able to report to funding agencies that data from the data producer's project have been reused, by which community, and how many times, even provides an incentive to share. The disincentives for the reuser are reduced, because they have had the opportunity to influence the metadata standards that will provide them with contextual information they find pertinent and they will have a better understanding of who collected the data for purposes of trust.

Conclusion

Data producers face significant pressure to share data with reusers, but are held back by many impediments. The biggest problem of sharing data is not knowing how the data will be reused and by whom. The open-endedness of this situation also means that data producers are performing this never-ending work of adding metadata for all possible audiences without any hope of credit as reuse may happen at any time. Some data are shared, and on the whole data producers are willing to share data when asked, but they also want certain conditions to be fulfilled to overcome the disincentives.

Courting data reusers is one strategy that data producers can use to overcome many of the challenges faced in data sharing and reuse. Two cases where data producers courted data reusers were described to illustrate the process, and what they were able to accomplish by engaging with their reuse populations. Although both cases came from surface and climate modeling communities, the challenges overcome by this approach are faced by many other disciplines. In the first case, a representative from an experimental community invited a modeling community to work with the experimentalists to make data accessible to the modeling community. And in the second case, members of the climate modeling community were reaching out to members of various policy-making communities in order to make climate model outputs available to them. In addition to mitigating the challenges of data sharing, the data producers are able to reap more immediate rewards by making their data sharing efforts visible to their funders and institutions.

The cases described here are still in the process of making data sharing work across their respective boundaries. How successful they are will only be determined by how each situation plays out. Further work is underway to capture the process of reusing climate model outputs from various perspectives, including the output producer, output reuser, and people who play an intermediary role in the transaction, based on interviews collected from participants at the QED workshop.

Acknowledgements

The work presented here was generously funded by NSF Award #0941386, 'Scaling Up: Introducing Commoditized Governance into Community Earth Science Modeling.' The author would like to thank the members of CSDMS and NCPP, specifically Dr. James P. Syvitski and Dr. Richard B. Rood, for inviting and supporting her participation in their respective events. The author would also like to thank Dr. Paul N. Edwards and Dr. Christine L. Borgman for their research support.

References

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059–1078. doi:10.1002/asi.22634

Edwards, P.N. (2010). *A Vast Machine: Computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Edwards, P.N., Mayernik, M.S., Batchellor, A.L., Bowker, G.C. & Borgman, C.L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science, 41*(5), 667–690. doi:10.1177/0306312711413314

Faniel, I.M., & Jacobsen, T.E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work, 19*(3–4), 355–375. doi:10.1007/s10606-010-9117-8

Faniel, I.M., & Zimmerman, A. (2011). Beyond the data deluge: A research agenda for large-scale data sharing and reuse. *International Journal of Digital Curation*, *6*(1), 58–69. doi:10.2218/ijdc.v6i1.172

Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2), 280–299.

Holdren, J.P. (2013). *Increasing access to the results of federally funded scientific research.* Retrieved from White House, Office of Science and Technology Policy website: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Kim, W. (2013). *Building a network for sediment experimentalists and modelers.* Paper presented at the Community Surface Dynamics Modelling System Annual Meeting 2013, Boulder, CO, USA. Retrieved from http://csdms.colorado.edu/wiki/CSDMS_2013_annual_meeting_Wonsuck_Kim

Mayernik, M.S., Batchellor, A.L., & Borgman, C.L. (2011). How institutional factors influence the creation of scientific metadata. In *Proceedings of the 2011 iConference* (pp. 417–425). New York, NY: Association for Computing Machinery. doi:10.1145/1940761.1940818

Nelson, B. (2009). Data sharing: Empty archives. *Nature, 461,* 160–163. doi:10.1038/461160a

Nuccitelli, D. (2013). *The 5 stages of climate dinial are on display ahead of the IPCC report.* Skeptical Science. Retrieved from http://www.skepticalscience.com/5-stages-climate-denial-on-display.html

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE, 6*(6), e21101. doi:10.1371/journal.pone.0021101

Wallis, J.C., Rolando, E., & Borgman, C.L. (2013). If we share data, will anyone use them. *PLoS ONE, 8*(7), e67332. doi:10.1371/journal.pone.0067332

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries, 7*(1-2), 5–16. doi:10.1007/s00799-007-0015-8

Zimmerman, A. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science Technology Human Values, 33*(5), 631–652. doi:10.1177/0162243907306704