Data Governance Workshop, Final Report

Arlington, VA

December 14-15, 2011

Supported by NSF #0753138 and #0830944

Abstract

The Internet and related technologies have created new opportunities to advance scientific research, in part by sharing research data sooner and more widely. The ability to discover, access and reuse existing research data has the potential to both improve the reproducibility of research as well as enable new research that builds on prior results in novel ways. Because of this potential there is increased interest from across the research enterprise (researchers, universities, funders, societies, publishers, etc.) in data sharing and related issues. This applies to all types of research, but particularly data-intensive or "big science", and where data is expensive to produce or is not reproducible. However, our understanding of the legal, regulatory and policy environment surrounding research data lags behind that of other research outputs like publications or conference proceedings. This lack of shared understanding is hindering our ability to develop good policy and improve data sharing and reusability, but it is not yet clear who should take the lead in this area and create the framework for data governance that we currently lack. This workshop was a first attempt to define the issues of data governance, identify short-term activities to clarify and improve the situation, and suggest a long-term research agenda that would allow the research enterprise to create the vision of a truly scalable and interoperable "Web of data" that we believe can take scientific progress to new heights.

Table of Contents

Introduction

Data governance is the system of decision rights and responsibilities that describe who can take what actions with what data, when, under what circumstances, and using what methods. It includes laws and policies associated with data, as well as strategies for data quality control and management in the context of an organization. It includes the processes that insure important data are formally managed throughout an organization, including business processes and risk management. Organizations managing data are both traditional and well-defined (e.g. universities) as well as cultural or virtual (e.g. a scientific disciplines or large, international research collaborations). Data governance ensures that data can be trusted and that people are made accountable for actions affecting the data.

Sharing and integrating scientific research data are common requirements for international and interdisciplinary data intensive research collaborations but are often difficult for a variety of technical, cultural, policy and legal reasons. For example, the NSF's INTEROP and DataNet programs are addressing many of the technical and cultural issues through their funded projects, including DataONE, but the legal and policy issues surrounding data are conspicuously missing from that work. The ultimate success of programs like DataNet depends on scalable data sharing that includes data governance.

Reproducing research – a core scientific principle – also depends on effective sharing of research data along with documentation on its production, processing and analysis workflow (i.e. its provenance) and its formatting and structure. Without access to the supporting data and the means to interpret and compare it, scientific research is not entirely credible and trustworthy, and this access again depends on data governance.

The research community recognizes that data governance issues, such as legal licensing and the related technical issue of attribution of Web-based resources would benefit from wider community discussion. The Data Governance Workshop was convened to discuss:

- Legal/policy issues (e.g. copyrights, sui generis database rights, confidentiality restrictions, licensing and contracts for data);
- Attribution and/or citation requirements (e.g. as required by legal license or desired by researchers);
- Repositories and Preservation (e.g. persistence of data and its citability, persistence of identifiers for data and data creators);
- Discovery and provenance metadata, including its governance (e.g. licenses for metadata);
- Schema/ontology discovery and sharing, including governance (e.g. licenses for ontologies)

The primary goal of the workshop was to develop a better shared understanding of the topic, and a set of recommendations to research sponsors and the broader community of scientific stakeholders for useful activities to be undertaken. In particular, the workshop discussed how NSF OCI (e.g. DataNet) projects might address these data governance questions as part of a sound data management plan, as mandated by the current NSF grant proposal guidelines. Additional goals for the workshop were to define useful short-term actions and a long-term strategic and research agenda.

Workshop participants are listed in appendix II and included scientists and researchers from the life, physical and social sciences, and representatives of data archives, research universities and libraries, research funding agencies and foundations, legal and advocacy organizations, scholarly publishing companies, and scholarly societies. This cross-section of the research community brought diverse perspectives to the discussion that informed and enriched the resulting recommendations.

Legal Landscape

The workshop commenced with a review of the current legal landscape surrounding data. Copyright law, while complex and nuanced, is largely harmonized world-wide, unlike other types of intellectual property law (e.g., sui generis database rights or patent rights). The law limits copyright protection for some types of data (e.g. facts and ideas are never protected) and the legal distinction between facts or collections of facts and protected "databases" are murky. Furthermore, different legal jurisdictions distinguish various types of data (like "factual" versus creative products) with different protections. For example, a database of factual sensor readings that is automatically in the public domain in one country may fall under intellectual property control in another, making it difficult to combine data produced by researchers in both countries without complex legal negotiation or development of a customized contract to harmonize the different laws for the purposes of the research project. Another nuance of research data is the distinction in many jurisdictions between a database and its contents – the former is often copyrightable while the latter may or may not be, depending on what it is and where it came from. While some approaches are more straightforward than others (as described below), the mere existence of these legal differences can make it necessary to involve legal counsel in establishing research project data sharing norms.

Privacy and/or confidentiality law is another important part of the legal landscape for data produced by medical research, and in the social, behavioral, and health sciences. These laws and regulations impose restrictions on storage, dissemination, exchange, and use of data, and are even more fragmented and diverse than in the area of intellectual property. In addition, institutions release this data with ad hoc, custom contracts (usage agreements) which are often incompatible with restrictions from other institutions using the same regulatory framework.

The overview covered copyrights, sui generis database rights, and the public domain as they apply to various types of research data, and the current legal tools and remedies to protect and share data: contracts, public licenses, and waivers. The merits of, and problems with, each approach was discussed, along with the merits of an open, commons-based approach to data sharing.

The complexities surrounding research data make it difficult to answer questions like "who has the right to decide which legal approach to take for a given dataset" or "is it allowable to combine datasets that were released under completely different contractual 'terms of use' each requiring that its terms and conditions continue to apply to the data in the resulting derivative dataset". Many researchers rely on scientific norms or conventional wisdom to resolve these questions since they lack resources to help them with any other approach, and this leads to behavior that may or may not be legally defensible and has questionable side effects for research reproducibility and data reuse.

Certainly the laws affecting data are not sufficient to insure that the norms of scientific research are followed. For example, there is an important distinction between releasing data at all (i.e. just making it accessible to other researchers) and making it effectively reusable or re-purposable for new research, with only the latter supporting research strategies that require combining multiple existing datasets. So part of data governance that exceeds the reach of law is specifying *how* data is to be shared so that it supports follow-on research and is not merely findable, if sought. Insuring data reusability requires additional policy to cover data quality and metadata provision, and separate mechanisms for policy enforcement such as contractual agreement (e.g. as a condition of funding) or dependence on scientific social norms of practice.

An issue that raises considerable concern among scientists and others in the community is around the meaning of "open" and some of the subtler points like the meaning of "non-commercial use" in various copyright licenses. While advocacy organizations like the Open Knowledge Foundation have published widely recognized definitions of "open" [1], these do not allow for limitations on the reuse of data for commercial purposes, e.g., by pharmaceutical companies or software start-ups. Many scientists desire protection from that type of commercial reuse, but would agree that imposing any legal limitations on reuse of their data might also create a barrier for colleagues wanting to use the data for non-commercial purposes. Since there is no standard or legal definition of "non-commercial" use[2], how such a condition will be enforced is uncertain. For example, if research results that drew from multiple reused datasets generate a patent or are included in a new textbook, is that a violation of the non-commercial terms of the license?

Because the laws around data do not insure scientific norms nor researchers' expectations, a typical approach for sharing scientific data is to impose a "terms of use" or "data usage agreement" on a data archive as a condition of searching, viewing and downloading data it holds. These are private contracts that apply to the person interacting with the data archive but do not "follow the data" if the person who agreed to the terms subsequently re-publishes the data, so they can be difficult to enforce. A further problem with these usage agreements is that they are often very complex, written in 'legalese' and incompatible with each other, making effectively data reuse and re-purposing impossible if their terms are observed.

Technology Landscape

Following the legal overview, the group reviewed the technological landscape for data sharing as it relates to governance. To make data sharing effective at Web-scale and thereby enable international e-science or network science we need a way to support automated, machine-processable information on what can be done with data, as well as its properties and quality. The technologies involved include media types and formats, metadata (for description and provenance), identifiers, persistence strategies, and software required to use the data. What is needed is nothing less than a new layer of the Web architecture to support research and scholarship at the social and legal level, and with minimal process – "ungovernance". This layer would then support the functional needs of data discoverability, accessibility, interpretability, reproducibility, and reusability.

The metadata required to support data governance includes both discovery and provenance metadata. Descriptive metadata supports the ability to learn about and locate the data on the Web, and basic information about its purpose, type, creator, funding source, etc. Descriptive metadata will include one or more identifiers that link to the data so that it can be accessed and cited. Provenance metadata includes information about the source of the data and the workflow and methodology to generate it and that is necessary for the interpretability, reproducibility and reusability goals of data sharing, as well as determining its quality. There are many challenges associated with metadata for data (sometimes called "paradata") among which is its transitive nature. Datasets not only evolve over time, but can be combined, derived from or built upon. In order to properly manage some types of data, we can imagine using something like the Github software development infrastructure to allow data to transition in multiple directions, while still retaining the original core in a retrievable form. If done properly, having

[1] See http://opendefinition.org/ for example

[2] More information on the noncommercial license terms is available from the Creative Commons website at http://wiki.creativecommons.org/Defining_Noncommercial and in particular, the report *Defining Noncommercial*. A more recent article on the subject is also available from http://dx.doi.org/10.3897/zookeys.150.2189

this metadata would be a great boon for scientists, since it would reduce the need to redo experimental and other data-production work. However producing this metadata is usually difficult and time consuming, since it is not integrated into the research process and we lack tools to make it easy and standards to store and share it. It would be useful to identify case studies of when this was done, how, and with what benefit to researchers.

Another type of metadata necessary for data governance and applicable to a governance layer of Web standards is rights metadata. Even in cases where the legal metadata exists, i.e., a license or rights waiver, such as CC0, CC-BY or ODC[3], has been applied to the data, there is no common way to technically communicate that information with the data today[4]. It is typically published on a web page connected to the site, or referenced from a separate metadata record describing the dataset (again, with the terms and conditions published as a Web page at the referenced URL). So ultimately a human has to interpret the conditions for access and reuse of the data. If a common and well-known license or waiver such as those from CC is used, then the researcher can avoid extra effort in determining the eligibility of the data to their research, but if it is a custom contract (e.g. a 'data usage agreement' or 'terms of use') then the researcher might have to resort to finding a lawyer to interpret the contract for them. This has a chilling effect on research, since few researchers have the time, funding, or access to expertise required to take that last step.

Another technological aspect of data governance relates to the software used to generate, process, analyze, or visualize the data – i.e., any software needed to interpret or reproduce the data. All information in digital formats, including research data, requires software to intermediate it and is otherwise a meaningless collection of bits. So for a researcher to validate research results by recreating or re-processing data the software used for the original research is usually necessary. And software that is integral to re-creating or re-processing data should be openly available (e.g. as open source software, ideally platform-independent) and publically shared to support future uses of the data. Therefore the software must also have metadata (descriptive and provenance, including versioning) so it can be discovered and its quality assessed, and it must be preserved just as carefully as the data associated with it.

Data Management Plans

The overview of the current landscape concluded with a review of data management and archiving. Many disciplines are moving to systematize data archiving, either in large, centralized repositories (e.g. GBIF, Dryad) or in institutionally-supported repositories (e.g. DSpace or EPrints instances). In a few cases, journals mandate data archiving (e.g. the Journal Data Archiving Policy, or JDAP[5], imposed by the majority of evolutionary biology journals, or Nature's policy[6] on depositing genomic data into GenBank prior to publication). Increasingly, research funding agencies are also requiring data archiving and open

[3] http://creativecommons.org/about/cc0, http://creativecommons.org/licenses/, and http://opendefinition.org/licenses/odc-by/

[4] One possibility for providing this metadata is in a separate "license.txt" file packaged with the data files (similar to the convention for open source software code), but it would still be difficult to automate discovery processes that incorporate the information in these separate files.

[5] "The JDAP is a policy of required deposition to be adopted in a coordinated fashion by Dryad partner journals. The Joint Data Archiving Policy (JDAP) is distinct from Dryad. However, it is recognized that Dryad is designed in order to make the JDAP easier, and without JDAP there would likely be limited adoption of Dryad; thus the two efforts are mutually reinforcing." For the text of the policy see http://datadryad.org/jdap

[6] See http://www.nature.com/authors/policies/availability.html for details

sharing as a condition of funding. These range from blanket policies (e.g. the Wellcome Trust[7]) to proposal guidelines from the NIH[8]and NSF[9], among others[10]. While not all agencies mandate sharing in all cases, and could not in some cases (e.g. where privacy laws apply) their intent is to encourage that behavior.

To take a recent example, data management plans for NSF proposals require description of the types of data to be created or used, the standards in which the data will be stored and preserved, and policies for insuring access to the data and under what terms and conditions. The requirement was created to protect the agency's investment in the research's outputs and optimize their value. In addition to their benefits of achieving research goals and supporting research reproducibility, the NSF believes that data management plans and their evaluation by review panels will evolve over time to become a more influential part of both the broader impact and merit review criteria, so adding it as a proposal requirement was just the first step. Review panels currently take direction from program officers and review plans inconsistently, but already there are knowledgeable PIs using their plans as a competitive advantage and including references to it in the proposal narrative. But the NSF is aware that the ultimate success of their plans requires community-driven guidelines, norms, and expectations.

To improve the quality of data management plans, tools like the DMPTool[11] are emerging to help researchers understand the issues and options, and create credible plans. Such tools are popular and present an opportunity to guide outcomes and behavior in good directions. But it will be important for tools and templates to be developed by the research community and not just librarians and research administrators.

A gap in both the data management plan guidelines and the emerging tools to create them is in the area of policies, and particularly for copyrights, database rights, and other applicable intellectual property policies, related to data sharing. There is a lack of expertise to guide researchers and research administrators, and uncertainty about who controls policy effecting the distribution and archiving of data (which have serious legal aspects and long-term costs). There are also uncertainties about how to create policies that are discipline-agnostic, and how to centralize policy in a time of rapid change. And this is exacerbated by a general unwillingness to tackle these issues by all parties concerned.

Section I: Current Conventions for Data Sharing and Reuse

A recent survey[12] of the research community undertaken by the DataONE project showed that 80% of scientists are willing to share their data with others in the research and education community. But the question was never raised to them as to how, legally, they might include a statement in their data about that willingness to share, and under what terms, so that their expectations are clear to others who are interested in using the data. The consensus view was that most researchers have simply never thought about these issues beyond their obligations in relation to human subjects and IRB regulations.

[7] See http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm
[8] http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
[9] http://www.nsf.gov/bfa/dias/policy/dmp.jsp
[10] Among other agencies requiring some degree or form of research data archiving and sharing are the CDC, DOD, DOE (Energy and Education), NASA, NEH, and the USDA.
[11] The Data Management Plan tool, or DMP tool, is a project of the California Digital Library and other organizations. For more information see https://dmp.cdlib.org/.
[12] http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101

A research discipline that has already worked through many issues of data sharing and governance is the social sciences, and particularly for census and survey data. They have the legal and technical tools to archive and share data, and well-established behavioral norms. But there are still weaknesses, especially with regard to personal privacy and subject consent agreements. And the legal contracts ("terms and conditions") for access are often very complex and incompatible across data repositories.

The main incentives for any research activity are recognition and credit for the work, with a secondary incentive being improved efficiency, quality, and impact of the research (e.g. avoiding replication of effort, or ability to better verify results). Compliance with funder, institutional and publisher mandates are also a consideration but are insufficient by themselves to insure good behavior unless enforced. Data publishing and citation standards and practices are needed to support better credit allocation and reward mechanisms for good data sharing behavior. And in the short-term there are simple measures that would increase awareness and begin to build expectations – proposal questions such as "how has your research data been reused by others in the past" or asking researchers who download datasets to publish their own data and link it back to the source data. Even with proper credit, researchers express uncertainty about what data they should share, when, with whom, in what form, for what purposes, etc., and lack the resources or expertise (or awareness that they lack expertise) to do what is necessary or even get help to find out what options are available to them.

From the publishers' perspective, there is renewed interest in the relationship between data and the publications that capture the research results from the data. Tighter integration of the data and publications is desirable for a variety of reasons, from making it easier to give credit to data providers to enabling "enhanced publications"[13] that simplify the mechanism of locating available data on particular topics. Publishers are increasingly interested in making sure that supporting data is available, often in advance of the article, but are uncertain of their own role in making that possible. Some are developing archiving solutions, some are partnering with institutions to link data to publications, and some are setting policy but remaining silent on implementation. At a minimum, publications should cite data in a similar manner to related publications, and could include statements about the data's availability and metadata for where to access it, and terms and conditions.

On the technology side, software used by researchers often makes subsequent data sharing and reuse difficult, since the "one tool per lab[14]" phenomenon is still common and there are few standards for structuring or encoding data to make it useful beyond its creators and the software they used. In other words, even if the data is successfully shared, without the software that produced, processed, analyzed or visualized it, the data is often not understandable by itself.

Whatever the current policies and intentions of funding agencies and publishers, unless researchers have access to appropriate infrastructure – repositories with long-term preservation capability, means of creating identifiers and metadata (or "paradata") for datasets, etc., the policies and goals for improved sharing of research data will not succeed. Where the infrastructure and support services exist,

[13] For more on this see http://www.articleofthefuture.com/

[14] The phenomenon of labs writing their own custom software tools rather than adopting those developed by another lab. While this is sometimes due to perceived or actual lack of documentation and support for the externally-produced tool, it is often unnecessary and leads to a proliferation of tools that do the same (or nearly the same) thing and are unsustainable over time.

recruiting data from research to comply with policies and scientific norms is easier, although the infrastructure itself is not sufficient to insure good compliance.

In order to achieve a good level of appropriate and effective data sharing, several things are needed:

- Clear and consistent statements of policy and enforcement practices by funders, publishers, institutions, societies and other research stakeholders;
- Easy-to-use and trustworthy infrastructure to accommodate the data and associated metadata
- Credit mechanisms to reward researchers for the effort of sharing their data;
- Better clarity around the researchers' (and other stakeholders') rights in and responsibilities for the data, including privacy/confidentiality regulations and copyright status;
- Harmonization of Data Usage Agreements, including privacy restrictions.

A further issue is that, even with all of these pieces in place, researchers shared concerns about the possibility of misuse of their data – of it being reused without regard to the expressed conditions (e.g. citation, commercial use restrictions), of reuse becoming a support burden, and of other less tangible fears. The counterweight for these concerns is the understanding that data underpins scientific research reproducibility and can help advance scientific progress (both core values of science).

Barriers to sharing include the inverse of the main incentives – i.e., lack of recognition or credit for the work required to share data effectively. Additional barriers are uncertainty about what to share, when, with whom, in what form, etc., and lack of resources or expertise to do what is necessary. Finally, some researchers fear unwanted exposure from providing full access to their research data and tools, leaving them open to criticism that would be difficult without such access.

Reusing data has its own challenges, since researchers are often uncertain of the provenance of a given dataset and whether it can be trusted, and are also often faced with significant effort to reformat the data for integration with other data or use by a different tool than the original research used. Legal issues play into this, since researchers rarely understand what rights adhere to their data and who holds those rights, i.e., themselves, their institution (the grantee), their funder, or no one (i.e. public domain). And even if the determination is made, what contract, license or waiver to apply to the data is another source of confusion. International, interdisciplinary and cross-sector collaborations raise further questions, for some of which there may not be clear legal answers.

Each stakeholder community wants a degree of control over policy affecting research data – researchers, funders, institutions/grantees, data archives, publishers, etc. There is recognition that policy needs to be cognizant of the research discipline it affects, but at the same time work across disciplines to support interdisciplinary research and achieve economies of scale. A general framework with room for discipline-specific detailed policies may be necessary to achieve everyone's goals.

Data governance affects all stakeholders in the research enterprise: funders, institutions, government, legislatures, disciplines, publishers, data centers, standards bodies, researchers, libraries, consumers, and the public. First we should consider the activities of data governance, including creating, enabling, enforcing, promoting, educating, and managing policies over time. Thinking about roles in relation to activities, we can begin to delineate a few:

- Institutions ensure compliance with funder policies, for which they need incentives;
- Libraries can provide services for education and advising, in addition to data curation activities;

- Governments and research disciplines (e.g. societies) create policies according to the discipline's and the public's interests, on both national and international levels;
- Publishers create policy for data related to publications to insure research validity and reproducibility, and manage them over time.

Section II: Short-Term Opportunities and Challenges in Data Governance

Participants discussed which opportunities to improve data governance might be achieved in the near-term (e.g., in the next 1-2 years) and what corresponding challenges need to be addressed. The examples provided were of common, normative "data usage agreements" that could be proposed to replace the ad hoc and heterogeneous contracts in place today; improved methods for embedding rights and licensing terms into metadata; developing a best practice for the terms and conditions of sharing metadata (as opposed to the primary data); creating educational modules on data governance for general distribution. Corresponding challenge examples included lack of clarity about the law with regard to research data, or conflicting beliefs about who "owns" and controls data sharing decisions. Our discussion went beyond issues of access to research data to include consideration of possible policies of funding agencies to avoid unnecessary replication of data if it has already been collected (or at least justification of overlap with existing data). Breakout groups were asked to develop "top five" lists of opportunities and challenges for the community, and what follows is an integrated and summarized description of the results.

Opportunities: Legal

1. All participants were clear about the need for clarification of current policy across all stakeholders involved in the research enterprise, and particularly of research funders and research institutions. While data-related policies should evolve over time with input from the research community, the lack of clear policies today makes discussion of the issues very difficult since it requires speculation about what agencies "meant" by statements they have made or "would say" in different scenarios;
2. The group was similarly clear about the need to clarify the legal situation with regard to data ownership in the context of research data of all types, from factual scientific data to media collections used by humanists or statistical data used by social scientists;
3. Given the common use of custom contracts or "Data Usage Agreements" by data archives and websites that govern access to and use of data, we agreed that a template with limited options would be beneficial to the community. Such a template should include options for data deposit, publishing and reuse (e.g. re-licensing terms), and encourage open sharing through exclusion of terms that would prohibit reuse or re-purposing. Such a template would require wide community review, with existing archives and range of data-producing projects and research funders and institutions;
4. Develop a legal framework for data that is linked to the data life cycle, mapping current records management practices to scientific research data. Adopt existing copyright management and transfer agreements whenever possible, adopting existing copyright management and transfer agreements whenever possible;

Opportunities: Social

5. Once the community of stakeholders involved in governing research data is identified, it will be possible to create a governance process that is community-led and empowered. Such an organization could be a quasi-formal organization, similar to the Smart Grid Interoperability

Panel[15]. It could consider all aspects of data governance: necessary standards and best practices, funder mandate recommendations, cost models, etc.

6. Given the current lack of information available about data governance, creating a Web-based resource intended as the default destination for the community to learn about the issues, get recommendations, communicate with each other, etc., would be timely and useful;
7. Another clear opportunity is in the area of education about data governance. Modules could be developed for either in-class or online dissemination to all levels of students and research practitioners, as well as librarians, data archivists, university administrators, publishers, etc. These can be initially general, then tailored to different disciplines and types of stakeholders overtime;
8. In parallel with creating general educational resources and a destination website for them, the community needs examples of how data sharing is beneficial and which governance regimes best serve the purposes of advancing research. Referring to these as "science design patterns", we mean them to be common, clearly defined scenarios with actions and consequences (both good and bad) provided. These could initially be case studies for the use of particular waivers, licenses or contracts, for particular types of data, with particular consequences;
9. Begin development of a "Digital Science Code of Conduct" (i.e. statement of goals and principles, as opposed to best practices at this early stage). Code would include practices like data citation and appropriate data description.

Opportunities: Technical

10. Infrastructure supporting "good behavior" in data sharing and interoperability is also needed and appropriately part of data governance, i.e., insuring that researchers can act on the stated policy of their funders and institutions. A specific example of that infrastructure is the DMPTool[16] to help researchers create credible and appropriate data management plans for grant proposals, as mandated by the NSF and NIH. Other examples are long-term archival data repositories, persistent and unique identifiers for published data, standards for descriptive and rights metadata, etc.;
11. In general, we need a better definition of data and some kind of taxonomy to help the community communicate more clearly about it. For example, how can we represent the different types of data, the relationship of data standards to each other, or the relationship of data to associated metadata or software. Developing that taxonomy or model for research data would facilitate both discussion of the issues and development of tools to handle interdisciplinary data uses;
12. Just as CrossRef created a large-scale registry of articles with persistent, unique identifiers (i.e. DOIs) and associated metadata to support interlinking, we could consider similar centralized infrastructure to associate datasets with identifiers and related metadata and rights/licensing information. Such a dataset registry or catalog could enable activities like discovering relevant research data that is eligible for reuse or re-purposing in a particular manner (e.g. data mining or combining with other datasets). In fact, CrossRef and DataCite already provide this service for datasets and could be further exploited as data discovery services.

[15] NIST convened the Smart Grid Interoperability Panel http://www.nist.gov/smartgrid/priority-actions.cfm to advise it on technical gaps and priorities identified by the Smart Grid community.

[16] Currently in development by the California Digital Library, more information is available at https://dmp.cdlib.org/

Challenges

1. The main challenge observed by workshop participants was the lack of clear leadership and expertise in data governance, and the lack of materials to support either education or collaboration on the topic;
2. There is a tension in data governance between generic laws governing data, which broadly apply to all instances of a type of data, and the variations across research disciplines in how they wish their data to be published, accessed and reused. Research funding agencies have been clear that they want solutions to emerge from the disciplines, but the law makes no such distinction for differences across scientific communities. We should not expect legal reform to solve this tension, so alternative methods must be sought such as well-crafted and standardized "terms of use" agreements that promote good scientific norms within disciplines;
3. Alongside intellectual property rights, another set of challenges for data governance stem from issues of national security, privacy rights, and other kinds of non-intellectual property legal protection. There are few mechanisms to balance the tension between open sharing of data and protecting the privacy rights of individuals involved in the research, or the interests of governments in the effects of that research;
4. Another major challenge to data governance is that there is not yet agreement about which parts of the community should be involved, and which should bear the costs of data sharing, based on defined benefits. We need a better understanding of the risks and rewards of data sharing, reuse and re-purposing, such as the "science design patterns" and case studies of impact described in the opportunities section above.
5. A related challenge is that there is no clear authority in the research enterprise for data governance, since most stakeholders feel they have some role to play in setting policy, and in compliance considerations, but there is little precedent or consensus for which stakeholders are empowered to make decisions if there are conflicts of interest and no consensus on a particular data sharing opportunity. In other words, we do not yet agree on who "owns" research data and has ultimate authority to set policies affecting it.
6. A huge challenge we observed for all of the opportunities identified is where to find the resources (both human and financial) to implement them, or sustain them over time. Data governance is a part of the larger data curation undertaking, which is similarly challenged, and has a few unique aspects like the need to understand international legal differences for data that has a global audience.
7. The opportunity identified above to create infrastructure supporting data governance, as well as the data-related activities that governance demands, has a parallel challenge: without that infrastructure governance has little teeth since researchers simply cannot comply with policy to effectively share their data.
8. The final challenge we recognized was simply the need for cultural change, which is never easy, to get everyone involved in the research enterprise to recognize the value of data sharing and the need for data governance as a means of achieving that goal.

Recommendations for Initial Activities

1. Related to opportunity #5 outlined above, we should begin exploring the creation of a "Data Governance Interoperability Panel" as an open, participatory and community-driven process to address the opportunities and challenges described above. Given the importance and timeliness of good strategies for data governance in parallel with the lack of clear leadership or locus of responsibility to advance solutions, such a Panel would be a way to begin to form that locus of activity in a positive, community-led way. The stakeholder groups represented at the workshop could form the nucleus of such a Panel to identify initial issues (possibly based on the workshop

findings), but the process should allow all stakeholders to identify themselves and participate as desired. It will be necessary to reach out to other countries and identify how to coordinate with them (e.g., in the UK, the Royal Society has established a group to do this type of activity, and there is a similar group in Australia). Another important stakeholder to consult is the National Science Board's committees focused on data, such as the National Science and Technology Council;

2. While a community forms around data governance, we can begin to develop some model practices for policies and legal practices related to research data, and do research to clarify the legalities of ownership of different types of data. To continue the momentum from this workshop, we could organize more focused workshops on topics like available data sharing licenses, current funding agency policies, or technological challenges of data interoperability under particular governance regimes (e.g. "attribution stacking" caused by certain licenses like CC-BY or ODC-BY). We can also collect existing information on the issues produced in the U.S. and other countries, such as at the Oak Ridge National Labs, or the Digital Curation Centre in the UK;

3. Institutions should be encouraged to provide resources to their researchers to create and implement good data management plans, e.g., infrastructure for data preservation, data identifiers, or appropriate waivers or licenses. Channels to promote these ideas include e-research forums (e.g. sponsored by EDUCAUSE or ARL), community forums (e.g. CNI's biannual forums or SPARC's Open Access events) or communities of practice around particular tools (e.g. DMPTool). Important initial audiences for this outreach are chief research officers at institutions and research funders, so identifying ways to engage those offices is a useful first step. Identifying ways to define the benefits of good data governance and associated data sharing and reuse will be a prerequisite to this outreach;

4. Defining metadata standards to describe data types and properties (including terms of use) is another priority for advancing data sharing. Developing the data taxonomy described in the "opportunities" above will facilitate discussion and clarify where there are problems and where there is general consensus. This could begin with an inventory of data types and related metadata being used today in various data repositories and archives. Important to this effort is distinguishing practices that are discipline-specific from those that cut across disciplines. Finding current research on this topic, such as existing data taxonomies or library research on metadata practices across the data lifecycle, would also accelerate the work. The "Data Governance Interoperability Panel" described above would be a natural group to vet and advance these standards in the future, but standards groups like the W3C and NISO will also have an interest;

5. Related to metadata standards are the development and promulgation of domain-specific data citation practices, as part of a "Code of Conduct" or as standalone recommendations from a governance body such as the above;

6. In the area of education, we could begin with simple steps like authoring a Wikipedia article on the subject, and collecting existing teaching materials from institutions that are actively promoting data curation to local researchers. Creating or collecting webinars on the topic that could be adopted by institutions would also be relatively easy. Authoring short blog posts on specific topics, e.g., how can you license a data mashup based on the terms of use for the original data, is another way to raise awareness and build a community;

7. Existing tools like the DMPTool should be refined to include some policy-awareness, such as suggesting specific waivers or licenses available to researchers to share their data. The tools can evolve to include all funding agencies and understand their particular policies. This suggests that recommended tools should be Open Source Software that is community-extensible, having Open APIs and Open data standards to allow, for example, extensions to DMPTool to attach default CC0 waivers to new Data Management Plans, with opt-out available.

Section III: Long term R&D agenda

Social Agenda

1. A recurrent theme of discussion at the workshop was the lack of a locus to work on data governance and policy development, so a long-term goal is to identify or build an organization to provide that locus. Such an organization would be necessarily interdisciplinary and cross-sector, e.g., university-based but collaborating with scholarly societies, funding agencies, publishers and other stakeholders. It could be a single organizations or a virtual one consisting of programs or activities at multiple organizations. Going beyond the scope of the "Data Governance Interoperability Panel" described above, the activities of this organization would include education, outreach and advocacy. An example activity might be to create an inter-agency institute for data governance that provides specific guidance to agencies in setting appropriate policy within different research domains. Another area of possible activity is advocating for data integrity, including quality, provenance, avoiding misuse, etc. and monitoring these on behalf of the relevant stakeholders. The organization would provide input to the "Data Governance Interoperability Panel" across the range of activities identified above, as well as providing a way to take forward recommendations of the Panel and the international data curation community;

2. Organizations working to advance open sharing of data, such as Creative Commons and the Open Knowledge Foundation, should work together with organizations developing technical infrastructure for data sharing and interoperability, such as DataCite and CrossRef, to develop basic principles for implementing the legal constructs surrounding both primary data and metadata. The "Interoperability Panel" could provide the venue for discussing proposed principles to insure their acceptability to the broader research community. An example of this is the development of a standard 'terms of use' contract template for data repositories that encourages open sharing of data with appropriate metadata;

3. Much more needs to be learned about the attitudes of researchers towards data sharing and how they see policy and the law as helping or hindering their goals. A survey, possibly conducted by an NSF project working with scalable data sharing such at DataONE, and involving scholarly societies from a range of disciplines, would be an initial activity, and the long-term outcome would be a set of recommendations for research funders and institutions for how best to support the research enterprise;

4. Exclusive reliance on researchers to supply needed metadata for their primary research data is unrealistic, even with improved tools for metadata production and extraction from the data. A better understanding of which agents could produce the necessary metadata, in what relation to the researchers and at what point in the data lifecycle, will be important for developing

practical workflows and policies for what can be required. Staff from data archives, libraries, publishers and societies may all play roles in this work and so forming some models for how this could work will clarify future roles and responsibilities;

5. Finally, beyond simple articles and educational modules, a broad-based communication and outreach plan that reaches both practitioners and decision makers will insure that the importance of data sharing and supporting data policies and infrastructure are well understood by the entire research community. Changing researchers' attitudes to recognize the scientific value and importance of sharing research data will be a long-term effort, but it should begin soon.

Technical Agenda

6. Understanding what policies are implicated in each stage of the data life cycle will facilitate discussion of data curation and governance. The analysis should include the metadata and software tools needed at each stage, and the motivations to comply with policies and produce needed metadata. Having this framework for data governance linked to the life cycle will allow prototypes for new, more efficient data production tools to be designed, built, tested and refined in operational settings.

7. Metadata for primary research data takes many forms and there are few standards for the various type of relevant metadata (i.e., discovery, provenance, preservation, policy) that interoperate across disciplines. There are efforts to define more Web-friendly metadata standards that build on the Semantic Web architecture, but these are also mainly within disciplines. Long-term work to harmonize metadata between disciplines and using common technical infrastructure (e.g., Web standards) will be necessary to achieving the goal of large-scale data reuse, integration and re-purposing for new research problems. The data governance community is well-positioned to work on this problem since it is interdisciplinary and international in scope, and includes representatives of all the necessary stakeholders. Especially important to include will be representatives of standards organizations (e.g. NISO, W3C) to insure broad input and help with outreach.

8. In parallel with creating better, more interdisciplinary standards for metadata to associate with primary research data, new tools and workflows are needed to automate as much metadata production as possible, to enable better capture of that metadata in useful contexts, and to make it easier to integrate data sourced from different communities (e.g., visualization tools to explore new datasets). Motivating researchers to produce metadata is a long-term social goal, but if instruments and tools could do the bulk of their work then curation costs would drop and the goals of data sharing could be achieved much sooner. And improving the ability to integrate data would increase data reuse and re-purposing, leading to more examples of the benefits of making data available.

Conclusion

The data governance workshop was felt by participants to have covered important new ground in the field of data curation, focusing as it did on the aspects of curation that are affected by policy and law, and characterized as data governance. Unlike some other types of research outputs such as publications, primary research data is not yet well-understood as a research asset or intellectual

property with defined roles and responsibilities for producing, sharing, reusing, and preserving it. How data fits into the scholarly record and relates to other research outputs, who has responsibility for it and can set policy for its management, how the research community should think about the relationship between policy and actions by researchers: these are all important issues that should be addressed if we want to have a realistic, credible, and ultimately successful system of data sharing and reuse.

This report identifies a set of activities for the short-term (next steps) and the longer-term research agenda, and we hope that the scientific research community will find these of sufficient importance and interest to continue the discussion and collaborate on future work on them.

Acknowledgements and Credits

The Data **Governance Workshop** had the following major contributors:

Conveners

MacKenzie Smith, Science Fellow, Creative Commons
Trisha Cruse, Director, UC Curation Center, California Digital Library
William K. Michener, Director, DataONE, University of New Mexico

Co-organizers

MacKenzie Smith, Science Fellow, Creative Commons
Trisha Cruse, Director, UC Curation Center, California Digital Library
Micah Altman, Senior Research Scientist, Institute for Quantitative Social Science, Harvard University
Geneva Henry, Executive Director, Center for Digital Scholarship, Rice University
John Wilbanks, VP for Science, Creative Commons

Rapporteurs

Geneva Henry, Executive Director, Center for Digital Scholarship, Rice University
Joy Kirchner, Scholarly Communications Coordinator, Digital Initiatives, University of British Columbia Library
Rebecca Koskela, Executive Director, DataONE, University of New Mexico
Ann Riley, Associate Director of the Access, Collections, and Technical Services Division, University of Missouri Libraries

Presenters

Sarah Pearson, Legal Counsel, Creative Commons
Jonathan Rees, Principal Scientist, Creative Commons
Carly Strasser, Project Manager, California Digital Library

Funders

National Science Foundation under award numbers #0753138 and #0830944

Appendix I: Data Governance Workshop Agenda

December 14-15, 2011
Hotel Palomar, Arlington, VA

Day One: Defining Data Governance
Wednesday, December 14

9:00-9:40	Welcome and overview of motivating problems and workshop goals
9:40-10:10	Participant introductions
10:10-11:00	Review of the legal status quo
11:00-11:20	Break
11:20–12:00	Review of the technical status quo
12:00–12:45	Review of the Data Management Plan status quo
12:45-1:45	Lunch
1:45-2:30	Common practices and implementations
2:30-4:00	Breakout: data sharing and reuse conventions and expectations
4:00-4:15	Break
4:15-5:00	Reports from breakout groups and discussion of norms
5:00	Adjourn
6:00	Workshop Dinner

Day Two: The Data Governance Roadmap
Thursday, December 15

8:30-8:45	Welcome, logistics and goals for the day
8:45-10:15	Breakout: top 5 opportunities and challenges, short-term practical objectives
10:15-10:30	Break
10:30-11:00	Reconvene to consolidate top 5 lists
11:00–12:00	Breakout: defining near-term next steps and commitments
12:00-1:00	Lunch
1:00-2:00	Breakout: defining a long-term R&D road map
2:00-3:00	Reconvene to define R&D agenda and review next steps
3:00	Adjourn

Appendix II: Attendees

Conveners

MacKenzie Smith, Science Fellow, Creative Commons
Trisha Cruse, Director, UC Curation Center, California Digital Library
William K. Michener, Director, DataONE, University of New Mexico

Participants

Micah Altman, Senior Research Scientist, Institute for Quantitative Social Science, Harvard University
Kevin Ashley, Director, UK Digital Curation Centre
Chris Biemesderfer, Publisher, American Astrophyiscal Association
Robert Chadduck, Principal Technologist for Advanced Research, National Archives and Records Administration
Bob Cook, Dist. Research Scientist at Environmental Sciences, Oak Ridge National Laboratory
Lee Dirks, Director, Education & Scholarly Communication, Microsoft Corporation
Josh Greenberg, Program Director, Digital Information Technology and the Dissemination of Knowledge, Alfred P. Sloan Foundation
Chris Greer, NIST
Gerry Grenier, Director of Publishing Technologies, IEEE
Geneva Henry, Executive Director, Center for Digital Scholarship, Rice University
Richard Huffine, Program Director, USGS Libraries
Joy Kirchner, Scholarly Communications Coordinator, Digital Initiatives, University of British Columbia Library
Rebecca Koskela, Executive Director, DataONE, University of New Mexico
Cliff Lynch, Director, Coalition for Networked Information
Daniel Mietchen, Open Knowledge Foundation
Sarah Pearson, Legal Counsel, Creative Commons
Jonathan Rees, Principal Scientist, Creative Commons
Ann Riley, Associate Director of the Access, Collections, and Technical Services Division, University of Missouri Libraries
Angela Rizk-Jackson, Research Scientist, UCSF
Carly Strasser, Project Manager, California Digital Library
Anita de Waard, Advanced Technology Group, Elsevier
Ruth Wilson, Publisher, Nature Publishing Group