Integrating Domain Repositories into the National Data Infrastructure: Meeting Report

Inter-university Consortium for Political and Social Research (ICPSR)
Ann Arbor, Michigan

November 20-21, 2014

Contents

Background 3
Emerging National Infrastructure Projects 3
Shared Access Research Ecosystem (SHARE) 3
Clearinghouse for the Open Research of the United States (CHORUS) 4
National Data Service (NDS) 4
Biomedical and Health Care Data Discovery Index Ecosystem (BioCADDIE) 5
Comparing the Projects 5
Science drivers 5
Sustainability 5
Collaboration across projects 6
Related Themes 6
Access versus integration 6
Data index 6
Metadata and other challenges 7
Designated communities 7
Federating data discovery across domains 8
Engaging researchers in data sharing, data management, and documentation 9
Incentives 9
Other ways to engage researchers 10
Linking data to publications/data citation 12
Sustaining repositories 13
Concluding Remarks 13
Data sharing 13
Data discovery 13
Guiding principles for new infrastructure 14
Next Steps 14
Appendix: List of Participants 17

Background

With support from the Alfred P. Sloan Foundation and the Research Data Alliance (RDA), ICPSR convened a meeting of domain repository representatives on November 20-21, 2014, in Ann Arbor, Michigan, to discuss the ways in which domain repositories can work together with emerging national infrastructures to provide access to data across all fields of science.

This meeting built on a Sloan-funded meeting held 18 months earlier on the topic of sustainable funding for domain repositories, which led to a white paper, "Sustaining Domain Repositories for Digital Data." The paper highlighted the important role that disciplinary repositories play in the scientific enterprise: they add value to data through curation, preserve data for future generations, and connect with researchers in the various domains. Leveraging these contributions is critical to the success of the new national infrastructure projects. This meeting provided a venue to start the conversation about the role of domain repositories in this new data landscape.

Emerging National Infrastructure Projects

Stimulated in part by recent mandates to provide open access to federally funded research outputs, four national infrastructure projects have been launched to enable discovery and sharing of research data and publications. The projects all have a strong open access focus but their objectives and funding models differ. Below are summaries of the projects.

Shared Access Research Ecosystem (SHARE)

Funded in part by a joint $1 million grant from the Institute of Museum and Library Services (IMLS) and the Alfred P. Sloan Foundation, SHARE is a higher education and research community initiative to ensure the preservation of, access to, and reuse of research outputs, including research data. SHARE will develop solutions that capitalize on the compelling interest shared by researchers, libraries, universities, funding agencies, and other key stakeholders to maximize research impact, today and in the future. SHARE aims to make the inventory of research assets more discoverable and more accessible, and to enable the research community to build upon these assets in creative and productive ways.

The Association of Research Libraries (ARL), the Association of American Universities (AAU), and the Association of Public and Land-grant Universities (APLU) have partnered to develop SHARE with significant input from the three associations' member institutions and their broader stakeholder communities. ARL, AAU, and APLU strongly believe that ensuring broad and continuing access to research is central to the mission of higher education. Research publications,

research data, other research outputs, along with their associated metadata, should be publicly accessible and available for reuse, text mining, data mining, and machine reading. This accelerates further research and discovery.

The higher education community also has the responsibility to collect and preserve their researchers' scholarly output for reasons beyond ensuring access, such as for institutional planning and compliance with the growing number of funder requirements. ARL, AAU, and APLU envision SHARE as the means for the academy to tackle these critical issues.

Clearinghouse for the Open Research of the United States (CHORUS)

CHORUS is a broad collaboration of over 100 scholarly publishers formed to develop, implement, and steward a partnership with the federal research funders for providing public access to the peer-reviewed publications that report on federally funded research. CHORUS serves as an information bridge, supporting agency search portals and leveraging publishers' existing infrastructure to facilitate a simple compliance process, optimized search and dashboard services, and multi-party archiving and preservation capabilities.
To initiate CHORUS' services, authors identify their funding sources when submitting a paper for publication with a participating publisher. The data that results from the tagging and subsequent public access is collected by CrossRef and provided by CHORUS to all at no cost through an open Application Programming Interface (API); this can be used by anyone to create new and customize available search and analytic tools. Applications to optimize search and enable funders to track and ensure compliance and analyze funding impact have already been developed by CHORUS and are available to participating agencies. CHORUS also partners with CLOCKSS and Portico to ensure the archiving and preservation of research papers.

National Data Service (NDS)

The National Data Service (NDS) is an emerging vision of how scientists and researchers across all disciplines can find, reuse, and publish data. It is an international federation of data providers, data aggregators, community-specific federations, publishers, and cyberinfrastructure providers. It builds on the data archiving and sharing efforts under way within specific communities and links them together with a common set of tools.

The National Data Service should make it easy to find, use, and publish data. It should provide a common set of services that can work across communities and disciplines, but it should also build on top of existing infrastructure already put into place by those communities. NDS services should have a strong connection to publications and publishing process to ensure that robust links between published literature and the data they discuss. Through these links, it should be possible to *make data as citable as literature*.

Core capabilities of the NDS will focus on helping researchers to find, use, share, and publish data, and linking data and literature.

Biomedical and Health Care Data Discovery Index Ecosystem (BioCADDIE)

Funded by NIH, BioCADDIE aims to engage a broad community of stakeholders in the development of a data discovery index (DDI) to do for data what PubMed (and PubMed Central) did for the research literature. BioCADDIE will engage stakeholders in the discussion of a sustainable ecosystem for making biomedical research and healthcare data discoverable and accessible.

BioCADDIE will develop and test tools to enhance data discoverability and usability in a way that respects those whose business depends on information/data production (e.g., researchers, authors) and dissemination (e.g., publishers and funding agencies), as well as data consumers who seek information to enhance their own health (e.g., patients) or the health of others (e.g., biomedical researchers, healthcare providers, and data scientists).

Comparing the Projects

Science drivers

In terms of the research drivers and research goals behind the projects and how they were attempting to advance research and help researchers, differences across the projects emerged. While SHARE and CHORUS are not as strongly driven by research needs, NDS and BioCADDIE will actively support verification and reproducibility of research results as well as the capability to produce new science with old data.

Sustainability

Each of the projects is in a different position with respect to sustainable business models. CHORUS will have a membership model. SHARE has not yet settled on a model but in general is seeking broad sustainability of the overall system to support the need to rationally manage research data over time. The NDS has no funded project yet, but is looking at all options. This will likely be a combination of things: large communities that will rely on a national infrastructure will need to build these costs into their budgets; national agencies will need to do more than what they are doing now; and there may also be interested individuals with the financial means to fund this project as a public good.

The BioCADDIE perspective is that data policies related to sustainability should be data-driven themselves, and a period of exploration is needed to ensure sufficient understanding of the big picture. As we move forward with large-scale sharing, we will learn that some studies need more investment than others and we will be able

to undertake more intelligent triage. If we determine that community repositories are critical to data sharing and preservation, then sustainable funding models for them will be needed.

More generally, there was a view that we may be getting ahead of ourselves in creating these national infrastructures without sustainable models for the vision. The analogy that we are building parts for cars with no design for the vehicle seems apt.

Collaboration across projects

SHARE and CHORUS have interdependencies, and these two efforts should work together in a cooperative way. SHARE and CHORUS are about keeping track of things while the NDS and BioCADDIE have the goal of changing scholarly practice.

Collaborations are essential in this very complex problem space, and we must bring all the stakeholders to the table. The cooperative model needs to be hammered out through a shared vision that we are all working toward or we will continue to reinvent the wheel. We need to learn how to do cooperative science now and define best practices with people agreeing to work on specific aspects of the problem and coming together at various times.

Related Themes

Access versus integration

Discovery and technical access get a lot of attention, but understanding the data and their respective ontologies and the actual science behind integrating data are also challenging issues. Researchers need to understand the relationships across datasets to integrate them in a way that is scientifically sound. In a sense, the intense focus on discovery undercuts integration needs.

At the same time, in some disciplines good data discovery aids for search and retrieval are lacking; researchers don't know what data are out there, and this constrains research.

Data index

The infrastructure projects all need to enable data discovery through some type of index. How many indexes do we need and how should search work most effectively since bringing disparate data together can result in minimal metadata? An approach that shows promise is multi-level searching that will direct the user from more general search results to a discipline-specific search. This kind of tiered search model is potentially the most powerful. A data index cannot substitute for deep data integration but the opposite is also true.

We are in a period of transition in which our notion of the traditional data catalog is changing. We should not constrain our vision of what the future of data discovery might look like, but at the same time we need to get started to develop tools that people need now.

Metadata and other challenges

DataONE surveys have shown that for data repositories and researchers, the biggest challenge in terms of effective data use and deep integration is metadata. We need more user-friendly tools to capture robust metadata. So far our approach has been to try to capture some minimal metadata, but we really need comprehensive metadata to do innovative types of analyses. We might look at alternative solutions to focusing on the metadata hurdle.

Some data are document-like and others are more dynamic – we need to recognize this diversity as well as differences in granularity. Not everything is immediately accessible in the discovery process because there are different levels of information. The first step is to learn that the data exist and then it should be possible to drill down to do more. Large databases only become useful with fine-grained specific queries. A hierarchical approach to discovery works well in the Virtual Observatory and may be useful in other fields as well.

Another huge barrier for discovery is getting data in electronic form that search engines of the future can find and others can mine. We need to show agencies that there will be a loss of return on investment unless these data are digitized, curated, and shared. In today's environment analog resources are harder to discover and integrate than digital and more likely to be ignored. Investing in digitization extends the life-cycle of resources, prevents their loss from contemporary consciousness, improves usability, and connects current research to the foundations on which it was built.

And aside from metadata for discovery, there is still a need for metadata that enable the understanding and interpretation of the data 100 years from now. The argument that a lot of science isn't getting done because people can't find the data is perhaps overstated. Use cases differ across disciplines.

Designated communities

To understand what to do and do it well, we need to bring in stakeholders who know the data and want to use the data. As domain repositories, we often think we know what researchers want, but this is not always the case. We need another kind of forum to bring together producers and analysts.

One of the most useful concepts that came out of the OAIS reference system is the Designated Community. OAIS ensures that the information to be preserved is

independently understandable to its Designated Community. The Designated Community needs to be able to understand the data without the assistance of the experts who produced the information. Most of us know our own Designated Communities fairly well, but how do we go beyond our Designated Communities to work across disciplines? Also, some Designated Communities have broadened over time and have new needs.

There is a United Nations independent advisory group working on the problem of data for integrated sustainable development. The effort is pulling in a set of science disciplines that need to be integrated, and this is an opportunity to support new user scenarios. How can we build an infrastructure that is responsive to those needs?

Unless we are aligned with Designated Communities that will explain why funders should invest in data that are reusable, we can't make a compelling case.

Federating data discovery across domains

Federation of data discovery is complex and means different things to different people. For purposes of this discussion, it was defined as "the ability to perform a set of operations (specifically search) across heterogeneous content regardless of its location." While there are different models of federation, the best approach appears to be a tiered one: first look broadly and then go deep for progressive revelation of data. It is also useful to think about types of queries – one might start out knowing what one is looking for but another user might start with the attributes of the data instead.

Use cases for federated discovery need to be developed. In medical informatics, for example, it is often the case that geographic data are needed (location of parks, factories, etc.).

Branding and ease of use are critical when federating across repositories. Using common conventions like displaying the data license on the website and consistent placement of a search box can enhance the user experience.

There are some concerns that arise with respect to federating discovery. For example, one might encounter an ontological collision when a word from one domain means something different in another (for example, "wave"). In these cases, the system should try to determine what the user is really interested in.

Engaging researchers in data sharing, data management, and documentation

Incentives

Researchers are motivated to share for several reasons: they believe their data are useful to others, they want to be recognized for their work, they are required to by granting agencies or their institutions, they are taught to do so while still students.

There are disincentives as well. Often, researchers cite the investment they have made in their own data and the desire to exploit the data before sharing as reasons not to share data. They sometimes fear that others will discover errors in the data or that the data will be misused. Sharing also places a burden on the researcher in terms of time and effort.

We can learn from each other by exchanging information in meetings and publications about what incentivizes researchers within a particular domain. Incentives can differ greatly across domains.

Helping people see how organizing/managing data helps *them* is also a good approach – for example, showing evidence that sharing data leads to a higher citation rate, increased impact of the data, and more visibility for a researcher. We should work on changing criteria for promotion to recognize non-standard research outputs and enable highly cited/high impact data to be counted in tenure and promotion decisions.

We also need to identify larger-scale incentives, those key scientific and societal objectives where more open data sharing would significantly accelerate research, e.g., saving lives during an Ebola crisis, sharing data on rare or difficult diseases, or finding sustainable development solutions before it's too late.

Another idea is to set up selection for long-term preservation as a prize/privilege/ sign of quality to get people to want to submit their most valuable data with good metadata.

Developing training courses and materials across disciplines for data sharing is needed. The data repositories could "open the wall" and allow the research community to get training from them; at the same time, repositories could learn from the scientists about what they are trying to do and what they need.

Another incentive is that publishing data is a protection against fraud and allows for other scientists to reproduce findings. The key here is the need for full transparency and for the deposit of all information necessary to replicate the research, including the workflow that was used, program code, software/versions used, etc. Also, we need to encourage researchers to share all results, not just those that will garner the most attention.

In addition, we can point out that, although researchers may believe their data are not reusable, it is quite difficult to predict when, how, or why another researcher may choose to reuse that data. For example, in human language technologies there are several cases of datasets being used for an array of unanticipated purposes.

Providing help with data maintenance, active curation, metadata, search, etc., as the Sustainable Environment Actionable Data (SEAD) project is doing, can also stimulate more data sharing. The SEAD project deploys the active curation concept with a model based on incentives. One of the biggest barriers to getting data into repositories is the fact that this step occurs after the project is complete and it involves yet more work. A solution is to make curation part of the ongoing work process by planning for data curation at the pre-proposal stage and tagging with appropriate metadata from the very beginning. One of the things SEAD has discovered is that it is possible to engage and align communities better if they are not thinking about curation as a second step. Rather, the focus is on progressive refinement – it's a loop.

Researchers should also begin interacting with repositories from the beginning of their projects. Some funding agencies require this and enable the cost of archiving to go into grants.

The greatest impact will be felt if and when funders make it a requirement to deposit the data in a repository where they can be accessed, reused, and preserved. Domain repositories and national infrastructure projects need to go in concert to the funding agencies and emphasize this.

Working with journals and professional associations is also important. A recent editorial in *Science* on data reproducibility advocated for access to full datasets for further analysis and scrutiny; editors of 30 scientific journals committed to this. This is a powerful approach to take. It should be noted, however, that data alone without the whole workflow may be useless depending on the domain. While depositing the entire workflow is an even higher bar, it could lead to a world where a reviewer could reanalyze the data and validate the results before publication.

Other ways to engage researchers

- Establish and promote the value of trusted repositories and digital preservation.

- Showcase groups doing things well. There are projects that are demonstrating best practices in terms of data management, and we should trumpet these examples. For example, TAIR (plant genome domain repository) has partnered with journals to insert language into author instructions asking for data deposit at the time of publication. Compliance is not perfect but is significant and increasing over time - culture change is gradual.

- Provide training on digital data curation and workflow. Higher-level courses on conducting research should deal with data management. There could also be required Sponsored Projects training (like IRBs require human subjects training, often on-line). In addition, the research libraries are helping people with data management plans, and more generally, libraries are an untapped resource for training and assistance.

- Enable repositories to provide a research workspace to encourage people to store their data early on while it is in process, permitting an easier transition to data sharing.

- Link people, publications and data. We should encourage the assigning of ORCIDS (Open Researcher and Contributor Identification) to students and researchers and link ORCIDs with existing user registration at repositories. We all suffer from registration fatigue and need to ease this burden.

- Build tools that work for everyone to do impact evaluation (citation studies, user feedback surveys, customer satisfaction surveys, etc.).

- Develop a Kickstarter for research with a list of projects that are not currently available or curated. This would be a way to identify data projects needing funding for curation, and people could contribute their curation expertise. Such a mechanism could provide a way to evaluate "drawer data" and then go to a funder with a number of projects to obtain money to digitize and curate them. A catalogue of such at-risk datasets could be created and prioritized, and this alone would raise their visibility.

  On a related note, we need to find a way to rescue a lot of data collected by senior scholars before it is too late. These might be data about to go out of existence, data embedded in media about to go out of existence, or data that are linked to researchers who are retiring (retirement may be a triggering opportunity). The community could decide which have high value and focus on those with the greatest potential for reuse, interoperability, etc. To facilitate this data rescue, we might create "data rescue groups" or join the CODATA Data Rescue Task Group (www.codata.org). Although current data rescue approaches do not scale, a few successes could generate publicity, and the publicity would attract not only support for data desperately needing rescue but also new contributions of already-digital data.

Linking data to publications/data citation

Several questions arise in thinking about the value and utility of data citation:

Why cite data? Will it incentivize data production, make knowledge claims potentially reproducible, and be worth the investment increase the value of data which get used over and over?

Are there minimal standards for datasets? Should there be triage, so that only data that meet those standards will be cited? Should there be best practices for disciplines – for example, only those that have adequate documentation, or are anticipated to have a large enough user base will be shared?

With respect to data repositories, should there be a requirement that the data venue has the characteristics of trusted digital repositories? Should the journal be empowered to say a personal website is not good enough? How do data repositories differ?

Why are data not being cited and who has the agency in encouraging data citation? How will solutions differentially empower and incentivize different types of stakeholders and gatekeepers like journal editors, domain repositories, authors, disciplines, departments?

What is the subset of citation problems that are addressed by national infrastructure? We would want the data to be ready to be used by multiple disciplines. Best practices instantiated at a higher level might lead to a better flow. National infrastructures could be the place these best practices are located.

FORCE11 is tackling many of these issues and recently published a Joint Declaration of Data Citation Principles. However, data citation is being handled in a fragmented way currently -- each repository has its own standards, which many or may not be used by journals. Repositories should strive to make citation recommendations uniform.

Replication is important but it is not the only reason for encouraging data citation. By having the data available and citable, we can make science better, faster, easier. It is incumbent on us to make it clear that sharing data for reuse is crucial.

What is the value proposition for a journal to encourage data citation and for a scientific society to encourage this practice? We need to make that clear. If all the claims in an article are backed up with data, it improves the quality of the article, leading to greater transparency. We want our content to appear in systems like SHARE and CHORUS.

Sustaining repositories

Sustained funding for domain repositories is a critical issue, and the repositories need to communicate and cooperate as there is strength in numbers. We need to get to a place where funding agencies understand that the type of data access provided by the domain repositories is an integral part of the data ecosystem. Our group needs to keep the pressure on. Related to this, we might also consider whether the integration of domain repositories could be more cost-effective. As a group we are carrying a lot of overhead and are duplicating services. There is a movement to look at ways for resources to switch locations so others can take them over. This type of distributed world could result in new efficiencies.

Concluding Remarks

Data sharing

With respect to data sharing, we often see both ends of the spectrum, which is the case, for example, in the chemistry community. Many people are advocating for data sharing and open science, but in large part they are speaking among themselves and don't gain traction with others who are not interested in data sharing.

There is the issue of hype vs. reality as well: does open data sharing lead to the expected scientific breakthroughs? A recent special issue of Nature on the human genome at 10 years posed the question of whether human health has really benefited from the sequencing of the human genome.

Similarly, as we move into the world of data publication and data sharing and data citation, there is an element of restraint that needs to be expressed: it may be many years before benefits of these desired practices are realized.

When data and literature are linked, is the data or the article the primary object? This depends on one's point of view at the time. If one is looking at the meaning of the results, it is the article, but the data and the article are interlinked and feed off each other. Also, the experimental procedures underlying the analysis may be complex and involve significant intellectual effort. Are these things given equal weight in terms of data citation and value to the community?

There is an energy barrier to data sharing. Some Electronic Lab Notebooks can serve as a type of data repository. Turnkey solutions like this remove the energy barrier. Standard formats help as well.

Data discovery

One should be able to discover and dive into the data in many different ways. Coming from the literature is one pathway, finding data based on grant descriptions

is another path, using the network of collaborations is another way of getting at the data based on the people who are working on it. And of course there are the traditional things like Google.

Perspectives about metadata have been changing over recent years. Rather than trying to build a system around common metadata, the approach may be to capture metadata and do a data mining approach to see which concepts bubble out of that and which concepts one can leverage. There is a practical way forward to federating diverse data.

We need to identify the things we want to manipulate and know about and manage. Some of these entities are the creators, some are what they create, some are pieces of them, and some are instruments, but we need to identify all of these things in a definitive way in the national infrastructure for robust discovery.

Guiding principles for new infrastructure

When thinking about national infrastructures what should the guiding principles be? The primary goal is for data and information to be as broadly available as possible. Also, the user experience is key. A network that attempts to match people with what they need seems the best approach. We should be building systems around the interests of those who are putting data in and taking them out, making those experiences as easy as possible for users.

We don't yet know all the answers to the big questions about optimal infrastructure, but the argument that trumps everything is transparency. We need to become more articulate about this to convince people about the power of data sharing. In terms of genomic data, the first questions asked of the sequenced genome were small in scope, but our efforts in making the genome available are starting to bear fruit now.

We also need to pay attention to the social systems that manage the technology, focusing on culture and economics and law so we are better able to manage what we are talking about. Data centers can manage data for their communities, but we need to look at the overall landscape. The solutions will bubble up from the bottom but need to be implemented from the top, so we need to identify the pressure points and connections.

Next Steps

After a broad-ranging discussion that surfaced a variety of challenging issues related to the evolving data landscape, the group of domain repositories formulated a plan for action with these steps:

Create a forum for wider discussion and engagement. U.S. domain repositories need to continue to get together on a regular basis to discuss their common interests. Funding will be necessary to accomplish this, and creating a Research Coordination Network (RCN) was suggested as a possible funding mechanism. The RCN could survey the landscape in which the domain repositories operate, compare existing standards and recommendations, and establish collaborations across a wide array of repositories.

The repositories need to hear from a broader group of stakeholders before starting a specific project. Other stakeholders who need to be at the table include but are not limited to the Cambridge Crystallographic Data Centre (CCDC), the National Institutes of Health, the National Oceanic and Atmospheric Administration (NOAA), the National Institute of Standards and Technology (NIST), and the National Center for Biotechnology Information (NCBI), as well as many other domain repositories listed in the re3data.org database.

We can also benefit from working within the Research Data Alliance (RDA). RDA is working in this space to encourage data sharing internationally and to remove some of the barriers by bringing people together around small projects with relatively short timelines. There is an existing RDA Domain Repositories Interest Group that we can leverage. Under the umbrella of this group, we can hold a splinter meeting before or after the next RDA Plenary in San Diego in March 2015.

The Open Repositories meeting June of 2015 in Indianapolis is another good venue for the discussion of these themes. In addition, there will be an NDS meeting in Austin in late March 2015.

Create a pilot project as demonstrator. Once we have established a more representative community, we can create a pilot project to show the vision for what we can accomplish with the collaboration. Our work should be fully transparent to engage the wider audience across diverse domains and to ensure openness in the decision-making process.

We will want to articulate one or more science cases to make the argument for federating across repositories. The approach should be to define the problem to solve first. We could make a good case around resiliency of cities, which need satellite data about the environment plus Census data and land use data. There is work going on at NCAR about this.

Continue to address key concerns, with an emphasis on sustainability. The sustainability gap continues to be an issue of utmost concern to the domain repositories, and the group needs to talk more about business models and investigate alternatives. We need to think about how to approach funders, the message we want to convey, and the types of funding mechanisms we might pursue. Data repositories currently use the same funding mechanism as researchers

producing data, putting the two groups in constant conflict. We should think about ways to ensure that data repositories are not competing for the same funds.

What is the value proposition for domain repositories? We need a cost analysis of the current infrastructure, and this should be done by experts as economic costs are many-faceted. RDA is working on this topic as well and has an Interest Group on Cost Recovery for Data Centres.

One of the opportunities at this juncture is for domain repositories that have deep grounding in particular sets of practices to think about how they can best use national initiatives to change the discussion about funding. The Digital Public Library of America is a good example to review.

Other topics the domain repositories should continue to explore include understanding science impact and what researchers want, sharing exemplar policies, and opportunities for mentoring across repositories.

Appendix: List of Participants

Karen Adolph	Databrary, New York University
George Alter	Inter-university Consortium for Political and Social Research, University of Michigan
Florio Arguillas	Cornell Institute for Social and Economic Research, Cornell University
Brian Athey	Department of Computational Medicine & Bioinformatics, University of Michigan
Chris Bader	Association of Religion Data Archives, Chapman University
Helen Berman	RCSB Protein Data Bank, Rutgers, the State University of New Jersey
Robert Chen	Center for International Earth Science Information Network, Columbia University
Christopher Cieri	Linguistic Data Consortium, University of Pennsylvania
Jonathan Crabtree	Odum Institute, University of North Carolina at Chapel Hill
Stephen Diggs	CLIVAR and Carbon Hydrographic Data Office, University of California, San Diego
Ruth Duerr	National Snow and Ice Data Center
Colin Elman	Qualitative Data Repository, Syracuse University
Carol Ember	Human Relations Area Files, Yale University
Catherine Fitch	Minnesota Population Center, University of Minnesota
Rick Gilmore	Databrary, Pennsylvania State University
Joshua Greenberg	Alfred P. Sloan Foundation
Robert Hanisch	National Institute of Standards and Technology
Margaret Hedstrom	Sustainable Environment Actionable Data, University of Michigan
Eva Huala	Arabidopsis Information Resource, Phoenix Bioinformatics Corporation
Diana Kapiszeski	Qualitative Data Repository, Georgetown University
Rebecca Koskela	DataONE, University of New Mexico
Elaine Kuttner	Cambridge Concord Associates
Carl Lagoze	School of Information, University of Michigan
Jennie Larkin	National Institutes of Health
Jared Lyle	Inter-university Consortium for Political and Social Research, University of Michigan
Clifford Lynch	Coalition for Networked Information
Ron Margolis	National Institutes of Health
David Martinsen	Clearinghouse for the Open Research of the United States
Maryann Martone	National Center for Microscopy and Imaging Research, University of California, San Diego
Marc Maynard	Roper Center
Francis McManamon	Center for Digital Antiquity, Arizona State University

William Michener	DataONE, University of New Mexico
Tom Murphy	Inter-university Consortium for Political and Social Research, University of Michigan
Gilbert Omenn	Department of Computational Medicine & Bioinformatics, University of Michigan
Beth Plale	Research Data Alliance, Data to Insight Center and Pervasive Technology Institute, Indiana University, Bloomington
Ray Plante	National Data Service
Ryan Scherle	DRYAD
Libbie Stephenson	Social Science Data Archive, University of California, Los Angeles
Paul Uhlir	National Academies of Science
Mary Vardigan	Inter-university Consortium for Political and Social Research, University of Michigan