Scientific Data: Increasing Transparency and Reducing the Grey

Bonnie C. Carroll, June Crowe, and J. R. Candlish
Information International Associates, Inc. (IIa)

Abstract

The foundation for all scientific research begins with data, however most scientific datasets are not publicly available and are an increasingly important part of the body of scientific grey literature.

We provide an overview of the current scientific data landscape, primarily in the United States with regard to both policies and tactical approaches to better scientific data management and access. This includes how to improve bibliographic control (a metric for the definition of grey) as well as approaches to make datasets more usable. Since few scientific datasets are commercially produced, the historical framework for differentiating grey literature must adapt to new ways of making scientific datasets less grey.

To illustrate how the identification, collection, management and access to scientific data is increasingly transparent (less grey), we will examine a couple case studies from communities of practice including medicine and earth sciences. We will examine what has been done to make them less grey and more discoverable, and the policies that are driving the change. We will conclude with a look to the future – the policies and technologies that will facilitate additional progress.

Keywords: Scientific Data, Information Policies, Grey Literature, Data Discovery

Introduction

Science in the 21st century will be conducted in a fully digital world. The results of research and development are born digital and have digital life cycles. From numbers to text to images and audio, all knowledge is reduced to bits. New ways of accessing these bits and new concepts of returning investment on their generation had generated an increased emphasis on scientific datasets which are often the elementary particles of science. In today's world, collaborative science demands the sharing of data. Advances in computing technologies allow for the collection of and analysis of data on a previously unimagined scale. Researchers and scientists are finding new ways to use old data (re-purpose) in which the original creator may not have even considered, and the idea that datasets be made available for the benefit of all science has gained significant momentum in recent years. A dynamic push-pull relationship drives the need for scientific data. The technology that enables data creation is the "push" while demand for data intensive science is the "pull" (Figure 1).

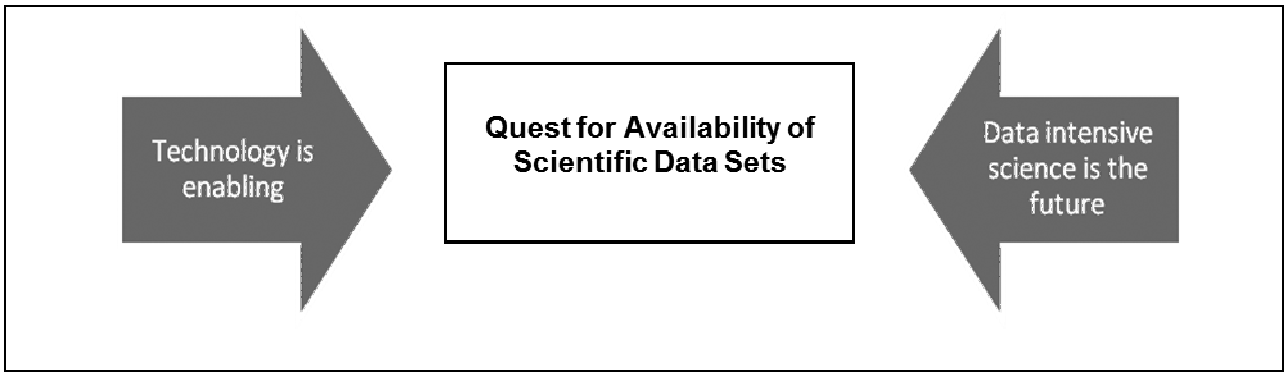

Figure 1. The drivers of scientific data**.**

More and more, researchers and policy makers alike adhere to the notion that "data are not consumed by the ideas and innovation they spark but are an endless fuel for creativity."[1] In many countries around the world, scientific data management (SDM) is being discussed at the strategic and policy level. The focus is to bring together the best practices to provide an overarching framework for SDM that includes the types of data and their expected impact; the relevant standards; and the provisions for protection, access, and continuing preservation.[2]

Why are data grey?

The exponential growth of scientific research in relation to the multitude of distribution channels has helped create a situation that unsurprisingly became a challenge to professionals in both information and scientific communities of practice. In order to set the stage of datasets as grey literature and to look at how transparency is being increased, it is important to review a definition of the term grey literature. The most widely utilized definition today is the following:

> *Grey Literature -- "information produced on all levels of government, academics, business, and industry in electronic and print formats not controlled by commercial publishing i.e. where publishing is not the primary activity of the producing body." (Luxembourg, 1997 - Expanded in New York, 2004)*[3]

Although not formally stated, the implicit nature of grey is also "obscure" or difficult to find and use. Essentially, a plethora of data "publishers," both traditional and new, have been operating without a uniform bibliographic control mechanism and without common methods to identify and access datasets that exist. This is a classic grey situation.

Scientific Data Landscape

The Challenge (Data Complexity: Heterogeneity and Volume)

The combination of two interwoven factors, data volume and heterogeneity, have created a very complex data landscape. Much has been discussed in a variety of disciplines about the magnitude of data and information creation. The supply and demand of digital data have exponentially increased in recent years, and all areas of science (i.e., experimental, observational, theoretical, modeling) have been transformed by the continuous cycle of data generation, access, and storage of an ever increasing volume of digital data[4]. Data volumes in 2005, for example, were growing at a doubling rate, and datasets were forecasted to reach petabyte size in the near future[5]. As a result, scientists have predicted dramatic changes in the way science is and will be conducted, and there are speculations that "few traditional processes will survive in their current form by 2020[6]". The term exaflood has been applied to the challenges presented by massive amounts of data being generated. Within the scientific spectrum, there are many contributing examples to the exaflood. The Large Hadron Collider (LHC), for example, produces 15 petabytes annually. This avalanche of digitally stored data provides a foundational platform for data to be dissected and analyzed, shared, and combined in innovative ways to better quantify unique characteristics of matter properties[7].

The challenge to data management created by volume is compounded by the disparate forms and types of datasets that constitute the heterogeneous scientific data landscape. In the 2005 National Science Board report, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century" [8], datasets were characterized as (Figure 2):

- Research database collections that are specific to a single investigator or research project.
- Resource or community database collections that are intermediate in duration, standardization, and community of users
- Reference collections that are managed for long-term use by many users.

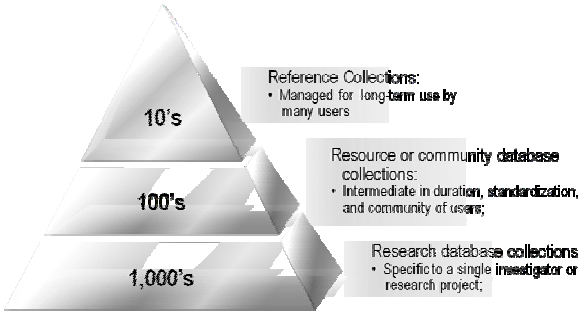

Figure 2. The types of data collections.

Each of these types requires very different management approaches and considerations but is rarely distinguished in discussion of "data." There is also a factor of heterogeneity across data types such as remotely sensed, field data, large and small scale experimental, and model data. Finally, each discipline has uniqueness in its data types and formats.

Publishers of Scientific Data

Part of the data landscape developed in efforts to increase transparency of datasets is a variety of venues in which scientific data are published. Currently, some of the primary sources that publish scientific data include:

1) Commercial publishers *(e.g. ProQuest)*;

2) Professional societies *(e.g., Ecological Society of America (ESA), Optical Society of America (OSA), International Council for Science: Committee on Data for Science and Technology (CODATA))*;

3) Repositories/clearinghouses/data archives *(e.g. Dryad, Sustainable Digital Data Preservation and Access Network (DataNet))*;

4) Information analysis centers *(e.g. Carbon Dioxide Information Analysis Center (CDIAC))*;

5) Research centers and researchers;

6) Metadata clearinghouses *(National Biological Information Infrastructure (NBII) Metadata Clearinghouse, Mercury Information Clearinghouse, Electronic Cultural Atlas Initiative (ECAI) Clearinghouse, Data.gov)*

Each of these has varying missions, goals, and reasons for existence. Traditional proprietary publishers (like ProQuest) have long competed in the marketplace by offering "value added" services designed to simplify the experiences of the end user. ProQuest offers a tool called "Deep Indexing" that allows users to search and retrieve information normally embedded in scholarly publications. This includes tables, figures, graphs, and illustrations within publications that have not previously been utilized separate of published material[9].

Professional societies have also historically published scholarly articles for the benefit of their paying members and affiliations. With the advent of newer technologies, many of the societies have embraced the open access concept, and/or participated in unique partnerships and collaborative agreements to increase transparency of scientific datasets. For example, the Ecological Society of America (ESA) established a data archive for publication material associated with ESA journals or partnering publishers. Three types of published content are archived for search and retrieval: appendices, supplements, and data papers. Additionally, authors are encouraged to officially register their published data to accompany archived content in efforts to strengthen the announcement of the existence of their data to a broader audience[10].

Repositories have long existed to house print format, and are equally essential as part of the digital landscape as well. These institutions may be clearinghouses or data archive centers, but serve essentially the same function. One significant recent initiative for digital data is Dryad, a collaborative partnership that originated in the research triangle of North Carolina, whose mission is to uphold an international repository that enables scientists to explore current and new methodologies of bioscience research, endorse published findings for intramural and extramural research initiatives, and potentially repurpose archived data for new applicable focus areas[11].

Another is National Science Foundation's (NSF) DataNet (Sustainable Digital Data Preservation and Access Network) Program that originated in 2007. To date, DataNet has funded two major projects that focus on integrating a plethora of diverse science areas (i.e., library and archival sciences, cyberinfrastructure, computer and information sciences) to enable and launch the development of long term sustainability of digital access, discovery, and preservation of science and engineering data. One of DataNet's additional initiatives is to accommodate the evolutionary pace of technologies and ensure reliable, long-term services to its audience[12].

The Information Analysis Center (IAC) is a concept proposed by the Weinberg Report[13] in 1963 that recommended the creation of centers of excellence where data were collected and analyzed to create new science from existing knowledge. This was a forerunner of today's data intensive science in a world without the digital technologies of today. There was a proliferation of IACs in the 1960s. On center that was a product of the movement and remains a leader in

scientific data today is the Carbon Dioxide Information Analysis Center (CDIAC) [14] at the Oak Ridge National Laboratory (ORNL). In many cases, because of the highly technical focus of these organizations, access to scientific data from these sources has been fairly limited to interested researchers and employees affiliated with the Center. However, today's computing technologies (i.e., portals, embedded links) are aiding in increasing transparency above and beyond the traditional users of said data. Today, CDIAC has been a significant contributor to global climate change research and its data are easily found and accessed by scientists throughout the world.

Research centers and individual researchers have also historically created data but little was actually "published" or made available for reuse. In the internet world, more of these are in fact "published" or made available if you know how to find them and use them. This syndrome is classic grey literature and reminds us of the state of technical reports in the 1950s and 60s.

Metadata clearinghouses are a final source of published scientific datasets. Although Data.gov[15] compiles data sets from the United States government above and beyond the field of science, it is an example of a metadata clearinghouse for datasets and clearly illustrates the national policy level of commitment to better provide bibliographic and metadata control over as well as access to scientific data sets. Other examples that are more topical in nature include: the National Biological Information Infrastructure (NBII) Clearinghouse[16], and the Electronic Cultural Atlas Initiative (ECAI) Clearinghouse[17]. More specific to the scientific community is Mercury[18], and this enterprise is discussed in more detail later.

Key Thrusts that Reduce the Grey

It is clear that no single entity can solve the digital data dilemmas alone and results are often achieved through leveraging a unique set of perspectives and efforts by multiple stakeholders and Communities of Practice. Senior management is ultimately responsible for the "top down" approach by developing data policy and for data management planning. A recent workshop that addressed scientific data management (SDM) in the United States[19] brought federal agencies together to look at data policy and planning across the federal sector as an initiative that will help develop an effective top down approach. Professionals engaged in the everyday routines of manipulating scientific data represent the "bottoms up" approach and have emphasized data citation and digital object identifiers or other types of persistent identification of datasets as an essential element in their communities of practice. The information management and technology communities of practice have helped by developing improved data discovery tools and expanding capabilities in their quest for a "semantic web." One recent initiative undertaken by CODATA (the ICSU Committee on Data) was to form a Task Group on Data Citation Standards and Practices[20]. The work of this group should help in increasing the transparency of datasets by creating guidelines on how they should be identified and cited.

A look at Figure 3 below will illustrate a brief snapshot of the scientific data landscape with regards to policies, technologies, and efforts by stakeholders to increase transparency of scientific datasets.

Policy, Culture and Management	Technology Trends & Applications
National Policy – government taxpayers funded projects should be accessible	Digital object management technology
Enhanced metadata	Growth of scientific workflow software
Journals supporting links to some published datasets	Adaptation of "netcentric" way of doing business
"People getting the message that data has to be accessible."	Use of embedded links in publications
Increased involvement of libraries and lifecycle management of data	Increased number of portals serving datasets
Younger generation post data as they go – expectation that data should be shared	

Figure 3. A high level view of the current scientific data landscape.

Data Activities in the United States/Case Studies

The following are a few case studies of initiatives that are working toward the better management of scientific data. With better management comes increased transparency.

Interactive Science Publishing (ISP) --- Medical

An innovative approach to scientific images and corresponding datasets has been implemented in a joint effort between the Optical Society of America (OSA) and the National Library of Medicine (NLM) entitled Interactive Science Publishing (ISP)[21]. OSA developed the ISP software in coordination with Kitware, Inc. and with support from the NLM and allows authors to publish large 2D and 3D datasets accessible through scientific articles. End-users can view and interact with original source data by downloading the ISP software. Ultimately, this software allows both readers and editors the ability to view, analyze, and interact with source data published in conjunction with an article. See figure 4 below for illustration.

... through the left main bronchus and into the distal section of the trachea, acquiring a 3D scan of the airway lumen. As shown in the axial view of Fig. 3, the *a*OCT scan enabled quantification of the lumen diameters at the time of the bronchoscopy.

A strong correlation was observed between CT and *a*OCT estimates of airway lumen diameters. A representative site in the proximal left main bronchus was selected for the purposes of illustration, with the same anatomical site visually identified for comparison. Using CT, the airway diameter was estimated to be 17.8mm x 14.1mm (Fig. 2). In the *a*OCT scan, the diameter was measured as 17.3mm x 13.9mm. Note that with the CT scan, we have used the oblique (not axial) view, so as to orient the measurement perpendicular to the central axis of the airway.

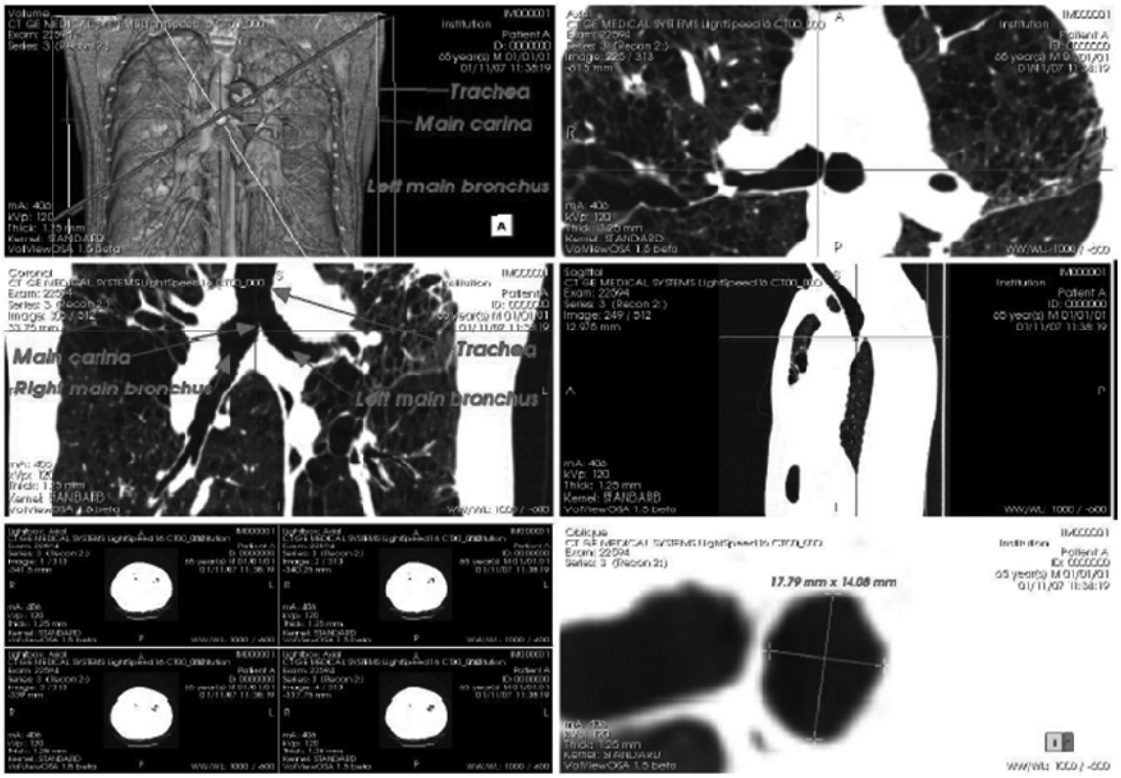

Fig. 2. Patient A. Chest CT depicting the lower airway (View 1). Top row (L-R): 3D view; Axial slice at the level of the main carina. Middle row (L-R): Coronal view; Sagittal view. Bottom row (L-R): Lightbox view; Oblique view measuring airway diameter.

Figure 4. An interactive 2D and 3D dataset derived from ISP software[21]**.**

Dryad's Repository --- Bio-diversity

Dryad is an international data repository created by the National Evolutionary Synthesis Center (collaborative effort by Duke, North Carolina, and North Carolina State Universities, and

the National Science Foundation) and the University of North Carolina Metadata Research Center. The data underlies peer-reviewed articles in basic and applied biosciences and enables scientists to achieve multiple objectives, including validation of published findings, repurposing data for research in new and innovative ways unanticipated by the original authors, and performing synthetic studies, to name a few. Dryad is governed by a consortium of journals that promote data archiving and will ensure the sustainability of the repository. Figure 5 below provides an illustration of the general concept and architecture of Dryad[22]:

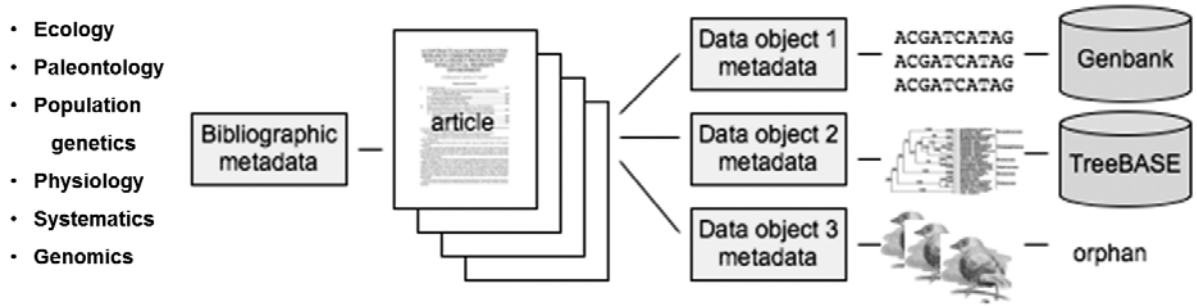

Figure 5: The Dryad repository conceptual flow diagram and general architecture**.**

As of Nov 23, 2010, Dryad contains 354 data packages and 868 data files, published in 50 journals.

Oak Ridge National Laboratory (ORNL) --- Earth Sciences

ORNL is a preeminent center for environmental scientific data management responsible for archiving, managing, and distributing data, and for enabling the distribution, use, and analysis of this data[23]. Three major data repositories (Atmospheric Radiation Measurement Archive, Carbon Dioxide Information and Analysis Center (CDIAC) [24], and the ORNL Distributed

Active Archive Center[25]), comprise the central resources in environmental data. At the core of this enterprise is the Mercury Metadata Clearinghouse[26] architecture that harvests metadata from multiple nodes and compiles information in a Metadata Index. The end user is able to access the data through a single portal with numerous search capabilities, including RSS feeds and other web alert services.

Opportunities/Look to the Future

Collaborative efforts to address the complex issues of making research data more available have been a major focus in recent years, and results are undeniably being achieved as a result. Recent survey results of data management practitioners[27] (see Figure 6 below) highlighted in a Scientific Data Management (SDM) Workshop reveal that 20% of respondents claimed "yes" that they had sufficient knowledge of datasets when planning projects and research programs, while only 14.3% responded no to the same question. However, over 65% responded to the same question with "sometimes" identifying that considerable improvement can still be obtained[28].

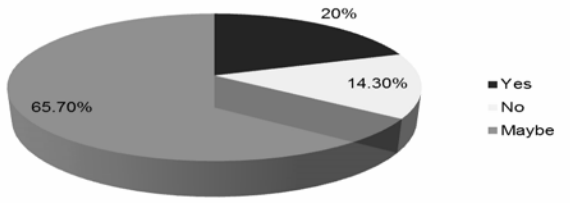

Figure 6: The percentage of survey responses that indicated sufficient knowledge of datasets when planning projects and research programs[29].

The case studies addressed in this paper identified a trend in increasing transparency in their respective disciplines, and comments from STI managers indicated that "people are getting the message that data needs to be accessible"

and that organizations have truly adopted a "netcentric" way of doing things. There will be an increasing sense that data belongs to the publication not just to the object. We will be looking at enriched publications that are supported by data and all tied together as bits in a web of science.

Conclusions

This thought paper briefly addressed the issue of scientific data as grey literature. It suggests that making scientific data more transparent and accessible is a major focal point of a lot of discussion at the national data policy level as well as from the communities of practice who are taking steps to gain better management control and provide better accessibility to scientific data.

In just the past few months since the paper was given at the Twelfth International Conference on Grey Literature (GL12) in Prague, Czech Republic[29], laws and Presidential directives in the US have focused on the need for better accessibility to scientific data[30]. In the UK a new Data Corporation has been proposed which will is a new but powerful "publisher" in the data landscape.

Although grey today under most definitions, the management and accessibility of scientific datasets are well on their way to more transparent access and control.

References

[1] National Science and Technology Council (NSTC), Office of Science and Technology Policy. Networking and Information Technology Research and Development (NITRD) Program. 2009. Harnessing the power of digital data for science and society. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Available: http://www.nitrd.gov/about/harnessing_power_web.pdf

[2] CENDI: Scientific Data Management Workshop, Washington DC, June 29-July 1, 2010. Accessed on October 15, 2010. Available: http://www.cendi.gov/activities/06_29_10_SDM_workshop_agenda.html

[3] Grey Literature definition from GreyNet International. Accessed on November, 15, 2010. Available: http://www.greynet.org

[4] Szalay, A., & Gray, J. 2006. Science in an exponential world. *Nature, (440),* 413-414.

[5] Gray, J. (2005). Scientific Data Management in the Coming Decade. SIGMOD Record, Vol. 34, No. 4, 34-41.

[6] Gray, J. (2005). Scientific Data Management in the Coming Decade. SIGMOD Record, Vol. 34, No. 4, 34-41.

[7] European Organization for Nuclear Research. The Large Hadron Collider. 2008. Accessed on January 19, 2011. Available: http://public.web.cern.ch/public/en/LHC/LHC-en.html

[8] National Science Board. National Science Foundation. 2005. Long-lived digital data collections: enabling research and education in the 21st century. Available: http://www.nsf.gov/nsb

[9] ProQuest. Cambridge Information Group. 2011. Accessed on January 18, 2011. Available: http://www.csa.com/e_products/pqdeepindex.php

[10] Ecological Society of America. Ecological Archives. 2011. Accessed on January 18, 2011. Available: http://esapubs.org/archive/

[11] Dryad Repository. 2010. Accessed on November 17, 2010. Available: http://www.dryad.org

[12] DataNet information. National Foundation. 2010. Accessed on November 17, 2010. Available: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141&org=OCI

[13] The Weinberg Report to the President's Science Advisory Committee (PSAC), 1963. Science, government and Information.

[14.] Carbon Dioxide Information Analysis Center (CDIAC). Oak Ridge National Laboratory (ORNL). 2011. Accessed on November 17, 2010. Available: http://cdiac.ornl.gov/

[15] Data.gov. 2011. Accessed on January 18, 2011. Available: http://data.gov

[16] National Biological Information Infrastructure Clearinghouse. Unites States Geological Society. 2011. Accessed on January 18, 2011. Available: http://metadata.nbii.gov/clearinghouse/

[17] Electronic Cultural Atlas Initiative (ECAI) Clearinghouse. The University of Sydney and Time Map™. 2010. Accessed on January 18, 2011. Available: http://ecaimaps.berkeley.edu/clearinghouse/

[18] Mercury Metadata Clearinghouse. Oak Ridge National Laboratory (ORNL). 2008. Accessed on January 18, 2011. Available: http://mercury.ornl.gov/

[19] CENDI: Scientific Data Management Workshop, Washington DC, June 29-July 1, 2010. Accessed on October 15, 2010. Available: http://www.cendi.gov/activities/06_29_10_SDM_workshop_agenda.html

[20] International Council for Science: Committee on Data for Science and Technology (CODATA). 2011. Accessed on January 18, 2011. Available: http://www.codata.org/index.html

[21] Interactive Science Publishing Optics InfoBase. The Optical Society. 2011. Accessed on January 19, 2011. Available: http://www.opticsinfobase.org/isp.cfm

[21] Childs, J. 2010. Interactive science publishing: a joint OSA-NLM project. Presented on January 12, 2010 at the National Library of Medicine (NLM). Available: http://www.cendi.gov/presentations/01-12-10_Ackerman_Michael_ISP.pdf

[22] Dryad Repository. 2010. Accessed on November 17, 2010. Available: http://www.dryad.org

[23] Oak Ridge National Laboratory (ORNL). 2011. Accessed on January 18, 2011. Available: http://ornl.gov

[24] Carbon Dioxide Information Analysis Center (CDIAC). Oak Ridge National Laboratory (ORNL). 2011. Accessed on

November 17, 2010. Available: http://cdiac.ornl.gov/

[25] Distributed Active Archive Center for Biogeochemical Dynamics. Oak Ridge National Laboratory. 2010. Accessed on November 17, 2010. Available: http://daac.ornl.gov/

[26] Mercury Metadata Clearinghouse. Oak Ridge National Laboratory (ORNL). 2008. Accessed on January 18, 2011. Available: http://mercury.ornl.gov/

[27] CENDI: Scientific Data Management Workshop, Washington DC, June 29-July 1, 2010. Accessed on October 15, 2010. Available: http://www.cendi.gov/activities/06_29_10_SDM_workshop_agenda.html Survey results available: http://www.cendi.gov/presentations/06-29-10_IWGDD_SDM_Survey_Carroll.pdf

[28] CENDI: Scientific Data Management Workshop, Washington DC, June 29-July 1, 2010. Accessed on October 15, 2010. Available: http://www.cendi.gov/activities/06_29_10_SDM_workshop_agenda.html Survey results available: http://www.cendi.gov/presentations/06-29-10_IWGDD_SDM_Survey_Carroll.pdf

[29] Twelfth International Conference on Grey Literature (GL12), Prague, Czech Republic. December 6-7, 2010. Available: http://www.textrelease.com/gl12program.html

[30] PCAST Report, Report, Designing A Digital Future: Generally Funded Research and Development in Networking and Information Technology, Executive Office of the President, December 2010, Section 6.3 Large-Scale Data Management and Analysis. Accessed January 19, 2010. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf; America Competes Reauthorization Act of 2010, H.R.5116, Section 103(b). Accessed January 19, 2010. Available: http://democrats.science.house.gov/Media/file/Commdocs/H.R.%205116_bill_text.pdf; and the Scientific Integrity Memo from the Office of Management and Budget (OMB). Office of Science and Technology Policy (OSTP). December 17, 2010. Accessed January 19, 2010. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/scientific-integrity-memo-12172010.pdf