Current Polices and Practices for Disseminating Research Results in the Fields Relevant to the Biological and Environmental Research Program

A Report of the Biological and Environmental Research Advisory Committee

Approved June 17, 2011

On February 25, 2011, Dr. William F. Brinkman, Director, Office of Science, charged the Biological and Environmental Research Committee (BERAC) with preparing a report describing the current policies and practices for disseminating research results in the fields relevant to the Biological and Environmental Research (BER) program. The BER program funds research across an exceptionally diverse range of scientific fields. This report includes research dissemination polices and practices for the following fields as they relate to BER-funded research:

- Climate science
- Atmospheric science
- Climate modeling
- Genomics
- Microbiology
- Plant biology
- Plant-microbe interactions
- Ecology
- Environmental science
- Biogeochemistry
- Subsurface science
- Nuclear imaging instrumentation
- Radiotracer chemistry
- Biological imaging

The following definitions are used as they relate to written research findings and digital data.

- Dissemination – refers to the circulation of research results outside of the origination institutions or scientific collaborations
- Research results – refers to both written research findings (scholarly papers, presentations, reports, etc) and digital data
- Practices – refers to accepted practices within a scientific discipline
- Policies – refers to policies from DOE and other federal and non-federal agencies that have notable impacts on the dissemination of research results in each field.

Criteria of dissemination and who makes this determination

Many common practices and decision-making processes for data dissemination are observed across the range of scientific fields represented by BER-funded research.

Written results are most typically published in peer-reviewed scientific journals at the discretion of the scientists conducting the research. The criteria for whether or where to disseminate the findings are mainly based on the value or scientific impact of the research findings, as determined by the scientists conducting the research and the peer reviewers used by the scientific journals to evaluate the submitted manuscripts. These criteria can include:

- the results and interpretation represent a standalone body of work that contributes to the scientific field
- the results and interpretation clarify and resolve a misunderstanding in the scientific field
- the results and interpretation provide newsworthy information useful in stimulating a field or discipline
- the results and interpretation help expand a body of knowledge into the K-12 or public sector in a way that advances the understanding of science by the public
- the results and interpretation synthesize new and/or existing information in such a way as to advance our scientific understanding

Written results, in addition to publication in peer-reviewed journals, often take the form of progress reports on grants submitted to DOE (and other agencies) and Institutional repositories (albeit a poorly organized and infrequently used one.) Dissemination of the former is achieved via the agency websites whereas publishers (and their respective policies) govern dissemination of published reports. Generally, journal articles are available online, either through the publishers' own websites or via other archives (e.g., PubMed Central, JSTOR, etc.). Increasingly, publishers are developing mechanisms to support delivery to mobile devices, whether through mobile-optimized content (to smart phones) or via Aps (to tablets).

Practices and processes for the dissemination of digital data are much more diverse and vary by scientific field and individual investigator. At the most general level, dissemination of digital data is determined mainly by the developers of the data, with consultation or approval by the principle investigators or managers of the projects that funded the development of the digital data. Digital data may include raw data from laboratory, field, and remote sensing instruments, data produced by numerical models, or the above data that have been processed to enhance their utility (e.g., combining or merging different data bases). The criteria for whether or how to disseminate the findings are based on the potential values or impacts of the data and quality control procedures. Today, many grants and publishers require that the data underlying published articles be made available in public repositories.

Climate science, Atmospheric science, Climate modeling

At the most general level, dissemination of digital data is determined mainly by the developers of the data, with consultation or approval by the principle investigators or managers of the projects that funded the development of the digital data. Digital data may include raw data from laboratory, field, and remote sensing instruments, data produced by numerical models, or the above data that have been processed to enhance their utility (e.g., combining or merging different data bases). The criteria for whether or how to disseminate the findings are based on the potential values or impacts of the data and quality control procedures. Although digital data are often made available by the scientist(s) conducting the research though it was noted that scientists do not always have the resources needed to meet data distribution demands. Example of specific digital data dissemination practices include:

- Policies for data dissemination are defined by the Community Earth System Model (CESM) Data Management and Data Distribution Plan, available at http://www.cesm.ucar.edu/management/docs/data.mgt.plan.2011.pdf. Briefly, the CESM policy requires that the owner (defined in the policy) must make CESM model output freely available on the internet no later than one year from the end of the simulation's

completion. All simulations, except those defined as "evaluation" or "testing", must make the model data publicly available.

- Climate model data is often submitted to a data archive such as the one maintained by the Program for Climate Model Diagnosis and Intercomparison (PCMDI), http://www-pcmdi.llnl.gov/.
- BER CLIMATE RESEARCH DATA SHARING POLICY: Research data obtained through public funding are a public trust. As such, these data must be publicly accessible. To be in compliance with the data policy of the U.S. Global Change Research Program of full and open access to global change research data, applications submitted in response to Funding Opportunity Announcements (FOA) must include a description of the applicant's data sharing plans if the proposed research involves the acquisition of data in the course of the research that would be of use to the climate change research and assessment communities. This includes data from extensive, long-term observations and experiments and from long-term model simulations of climate that would be costly to duplicate. The description must include plans for sharing the data that are to be acquired in the course of the proposed research, particularly how the acquired data will be preserved, documented, and quality assured, and where they will be archived for access by others. Data of potentially broad use in climate change research and assessments should be archived, when possible, in data repositories for subsequent dissemination. Examples of DOE funded data repositories may be found at http://cdiac.ornl.gov/, http://wwwpcmdi.llnl.gov/ipcc/about_ipcc.php. The repository where the applicant intends to archive the data should be notified in advance of the intention, contingent on a successful outcome of the application review. If data are to be archived at the applicant's home institution or in some other location, the application must describe how, where, and for how long the data will be documented and archived for access by others. Applicants are allowed an initial period of exclusive use of the acquired data to quality assure it and to publish papers based on the data, but they are strongly encouraged to make the data openly available as soon as possible after this period. DOE's Office of Biological and Environmental Research defines the exclusive use period to be one year after the end of the data acquisition period for the proposed performance period of the grant application but exceptions to extend this period may be justified for unique or extenuating circumstances.
- BER CLIMATE MODEL DATA SHARING POLICIES
    - BER supports development of the community model "CESM". It is described, and policies on model code and data sharing are described on the CESM website (see above).
    - The Weather Research and Forecasting (WRF) Model is a next-generation mesoscale numerical weather prediction system designed to serve both operational forecasting and atmospheric research needs. WRF was developed at the National Center for Atmospheric Research (NCAR) which is operated by the University Corporation for Atmospheric Research (UCAR). NCAR and UCAR make no proprietary claims, either statutory or otherwise, to this version and release of WRF and consider WRF to be in the public domain for use by any person or entity for any purpose without any fee or charge.

- ATMOSPHERIC RADIATION MEASUREMENT (ARM) RESEARCH DATA SHARING POLICY
    - The ARM data policy is designed to achieve several specific objectives:
        - Establish a data policy in full compliance with the spirit of the U.S. Global Change Research Program.
        - Accelerate the research process by providing timely access to data to facilitate early publication of results.
        - Validate high accuracy measurements through timely intercomparison and/or the establishment of Quality Measurement Experiments.
    - The basic tenets of the ARM data sharing policy are:
        - "Free and open" sharing of data.
        - Timely (e.g., "near real time" where desired) delivery of processed data from the ARM Data Archive to Science Team members.
        - Timely access (e.g., typically "days" for routine processing, housekeeping and archival of data from electronically accessible instrument sites) to data by the general scientific community through the ARM Data Archive.
        - Timely sharing of all data among various participants in ARM-sponsored programs.
        - Recognition of data sources either through co-authorship or acknowledgments as appropriate.
        - Sharing of data of common interest from external sources when possible, Some sources restrict secondary distribution of data. In these cases, ARM will seek specific allowances to distribute such data to members of the ASR Science Team, but will observe restrictions on further distribution from the ARM Data Archive if required.
    - The data policy covers data acquired from ARM instruments as well as ARM-funded guest instruments that participate in Intensive Operational Periods (IOP). Collaborating, non-funded participants are encouraged to provide their data as well. Provisions for data sharing are included in ARM funding mechanisms that require the submission of data no later than 90 days from the completion of the IOP or campaign.

Genomics, Microbiology, Plant Biology

Digital data is made available through a number of public and enhanced databases, including:

- GenBank, http://www.ncbi.nlm.nih.gov/genbank/
- European Molecular Biology Laboratory Nucleotide Sequence Database, http://www.ebi.ac.uk/embl/
- DNA Databank of Japan, http://www.ebi.ac.uk/embl/
- DOE Joint Genome Institute Integrated Microbial Genomes Data Management and Analysis Systems, http://img.jgi.doe.gov/

- Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis, http://camera.calit2.net/
- Minimum Information about a (Meta)Genome Sequence, http://gensc.org/gc_wiki/index.php/MIGS/MIMS
- Ribosomal Database Project, http://rdp.cme.msu.edu/
- Protein Data Bank, http://www.pdb.org/pdb/home/home.do

- JOINT GENOME INSTITUTE DATA RELEASE POLICY
    - The JGI Standard Data Release Policy for all projects is immediate release, which includes the following:
        - All Sanger trace files will be deposited to the public NCBI Trace Archive as the data becomes available. Short reads (454 and Illumina) will be deposited to the NCBI Short Read Archive as the data becomes available. (Note that NCBI recently announced that due to budget constraints, it would be discontinuing its Sequence Read Archive (SRA) and Trace Archive repositories for high-throughput sequence data. However, plans are in place to continue to archive some forms of sequence data, see details at http://www.ncbi.nlm.nih.gov/sra).
        - Draft assemblies and preliminary automated annotation will be posted on the public JGI website and submitted publically to GenBank as the data becomes available. The release of draft assemblies and preliminary automated annotation may not be simultaneous.
        - Improved assemblies and annotation or other value-added data will be publically released as it becomes available.
        - All data will meet a set of quality evaluation criteria prior to release.
        - All data are subject to the data usage policies described at http://genome.jgi-psf.org/pages/data-usage-policy.jsf.
        - After the initial release to the PI, all assemblies and annotations (including finished or other value-added data) will be publically released immediately.
        - This JGI Standard Data Release policy is effective for all new Community Sequencing Projects (CSP) proposals/projects approved May 15, 2009 and retroactively to CSP proposals/projects approved previous to May 15, 2009 for which starting material has not been received by October 1, 2009. For those proposals requiring a User Agreement renewal from previous years, the same data release policy on the previous Agreement will be carried over to the new User Agreement.

- Joint Genome Institute – Bioenergy Research Center Sequencing Data Release Policy
    - While it is the policy of the Office of Biological and Environmental Research (OBER) to support immediate and open data and information sharing within our program, we recognize the need to have a limited and time-sensitive protection of certain types of data for rapid development of Intellectual Property under the special circumstances associated with the DOE Bioenergy Research Centers (BRCs) . Therefore, it is OBER policy that all sequencing work performed by the Joint Genome Institute (JGI) on behalf of the DOE BRCs will have a default 6 month data and information embargo. All data and information related to BRC sequencing projects will be withheld from public display and distribution until 6 months after the BRC Principal Investigator

(BRC PI) has received the initial release of the automatic annotated sequence(s). It is the responsibility of the BRC PI to report all inventions to the BRC Intellectual Property Management Group, which will assign an invention disclosure number (and DOE S-number). Under the BRC contract, the BRC may elect title of these inventions and file provisional or non-provisional applications for patent(s) based on the finding of the work. At the end of the 6 month data embargo, all JGI generated data and information unrelated to an invention will be made publicly available. Extension beyond the 6 month embargo will be considered individually by OBER. This policy is only applicable to the DOE BRC specific sequencing projects and does not include community sequencing projects, regardless of BRC involvement.

- GENOMIC SCIENCES RESEARCH DATA RELEASE POLICY

  Information and Data Sharing Policy

  Final Date: April 4, 2008.

  Introduction:

  Experimental biology has evolved in the past 20 years to include a rapid-access, global scientific community hyper connected through the Internet. The changing scope of scientific inquiry and the astonishing rate of data production drive the development of a new type of cyber infrastructure, which, in turn has promoted the formation of e-science (1). Journals, funding agencies and governments correspondingly have developed information standards and sharing policies, all of which in one way or another address research conducted in an open-access environment. A key hallmark of these policies is the requirement that scientific inquiry and publication must include the submission of publication relevant information and materials to public repositories. For the most part, the policies follow the uniform principle for sharing integral data and materials expeditiously (called UPSIDE)(2). Conversely, when research information is not made publicly available to a global scientific community, a corresponding price is paid in lost opportunities, barriers to innovation and collaboration, and the obvious problem of unknowing repetition of similar work (3).

  This statement summarizes the information and data-sharing policy within the Genomics: GTL (GTL) program at the Department of Energy's Office of Biological and Environmental Research (OBER). OBER recognizes that successful implementation of this policy will require the development of new technologies such as software tools and database architectures, and will be funded, as necessary, from the GTL program subject to funding availability. We affirm our support for the concept of information and data standards and sharing and we believe that a comprehensive policy can be constructed that will encourage GTL researchers to exchange new ideas, data and technologies across the GTL program and the wider scientific community.

  Research information obtained through public funding is a public trust. As such, this information must be publicly accessible. The GTL information-sharing policy requires that all publication related information and materials be made available in a timely manner. All Principal Investigators (PIs) within the GTL program will be required to

construct and implement an Information and Data-Sharing Plan that ensures this accessibility as a component of their funded projects.

Policy Statement:

The Office of Biological and Environmental Research (OBER) will require that all publishable information resulting from GTL funded research must conform to community recognized standard formats when they exist, be clearly attributable, and be deposited within a community recognized public database(s) appropriate for the research conducted. Furthermore, all experimental data obtained as a result of GTL funded research must be kept in an archive maintained by the Principal Investigator (PI) for the duration of the funded project. Any publications resulting from the use of shared experimental data must accurately acknowledge the original source or provider of the attributable data. The publication of information resulting from GTL funded research must be consistent with the Intellectual Property provisions of the contract under which the publishable information was produced

I. Applicability

This policy shall apply to all projects receiving funding in the Genomics: GTL program as of October 1, 2008. For cases where information sharing standards or databases do not yet exist, the information sharing and data archiving plan provided by a project's PI must state these limitations. Data and information that are necessary elements of protected intellectual property and related to a pending or future patent application are explicitly exempt from public access until completion of the patenting process. Adherence to this policy will be monitored through the established procedure of yearly progress reports submitted to GTL program managers. All information regarding data shared by GTL-funded research projects will be made publicly available at genomicsgtl.energy.gov/datasharing.

II. Submission of Information and Data

All investigators are expected to submit their publication related information to a national or international public repository, when one exists, according to the repository's established standards for content and timeliness but no later than 3 months after publication. This includes:

Experimental protocols,

Raw and/or processed data, as required by the repository,

Other relevant supporting materials.

OBER will maintain a website listing all published peer reviewed papers and published patents resulting from GTL funding and PIs are expected to inform OBER on a regular basis when a publication appears in print. OBER is encouraged by the development of the National Institutes of Health open-access policy and, when possible, OBER will link to open-access GTL funded publications. PI's, however, are encouraged to publish in journals appropriate to their fields of research. OBER recognizes that sub-disciplines and experimental technologies have varying degrees of cyber-infrastructure and standard ontology to accommodate this policy. Specific guidelines and suggestions for GTL investigators are provided below.

II. A. Nationally and Internationally-Accepted Databases and Ontologies

II.A.1. Sequence Data

The field of genomic sequencing has a very well developed mechanism for public archiving of experimental data. Nucleotide sequence data will be deposited into GenBank, and protein sequence data will be deposited into the UniProtkb/Swiss-Prot Protein Knowledge database. In fact, beginning in the early 1990s, it has been an editorial policy of most scientific journals that submission of data to the public repositories, especially sequence data, is a requirement for publication of articles that report scientific findings based on the data. Investigators should report to OBER the sequence identifier including the accession number and version. In addition, investigators are encouraged to use the gene ontology annotation database (4) when possible and OBER applauds the work of the Genomic Standards Consortium (GSC) in the development of minimum information about a genome sequence standards (MIGS).

Specifically for large-scale GTL sequencing projects, OBER will adopt the policy that whole genome sequencing data, where genome completion is the stated goal, must be made publicly available 3 months after first assembly of the sequencing reads for that genome. In the case of metagenomic sequencing, data must be deposited to the National Center for Biotechnology Information (NCBI) 3 months after completion of the last sequencing run, which must be specified in the JGI User Agreement. For other types of sequencing experiments, such as expressed sequence tags (ESTs), the data will fall under the guidelines for publication of relevant information and shall be deposited to NCBI 3 months after publication.

II.A.2 Three-Dimensional Structural Data

All coordinates of solved structures, along with structure factors, and related information for structures of biological macromolecules and complexes are to be deposited in the Protein Data Bank (PDB) or Nucleic Acid Databank (NDB), as appropriate. Accession codes are to be reported back to OBER.

II.A.3 Microarray and Gene Expression Data

The Microarray and Gene Expression Data (MGED) Society recommends the use of a MGED ontology for the description of key experimental conditions as, for example, using a MIAME-compliant format (MIAME, Minimum Information About a Micro-array Experiment). OBER's policy will follow the MGED recommended ontology. We further strongly encourage GTL researchers to deposit raw and transformed data sets and experimental protocols to a public microarray database and report back to OBER the accession number and URL. Possible microarray databases for data deposition include the Gene Expression Omnibus (5), ArrayExpress (6) and the Stanford Microarray Database (7).

II.B. Information Sharing Systems and Databases Under Development

II.B.1 Proteomics

The Proteomics Standards Initiative (PSI), a working group of the Human Proteome Organization (HUPO), recently outlined two standard proteomics ontologies: minimum information about a proteomics experiment (MIAPE) (8) and minimum information

required for reporting a molecular interaction experiment (MIMIx) (9). Because this is an evolving initiative and the field is still immature, we cautiously encourage GTL proteomics researchers to adopt the use of MIAPE and MIMIx in their research. We are further encouraged by the development of public proteomics repository databases such as the Open Proteomic Database (10) and PEDRo (Proteome Experimental Data Repository)(11) and encourage GTL researchers to engage with these databases. However, we recognize that standards and ontologies will evolve within the proteomics community and GTL's policy will follow guidelines set forth by HUPO as they develop.

II.B.2 Other Technologies

GTL research makes use of a large variety of technologies for which there are, as yet, no national or international information standards and archival formats. Scientists in the GTL program are encouraged to participate in the efforts of research communities to develop such standards for enabling information sharing.

GTL's long term objective is to encourage the development of infrastructure for technologies that do not as yet have nationally or internationally accepted information sharing standards. In cases where there are no public repositories or community driven standard ontologies, OBER recommends that these types of data and information be made publicly available by the PI.

III. Protection of Human Subjects

Research using human subjects provides important scientific benefits but these benefits never outweigh the need to protect individual rights and interests. OBER will require that grantees and contractors follow the DOE principles and regulations for the protection of human subjects involved in DOE research. Minimally this will require an IRB review. These principles are stated clearly in the Policy and Order documents: DOE P 443.1A and DOE O 443.1A, which are available online at www.directives.doe.gov.

IV. Systems Biology and the GTL Knowledgebase

A long-term vision for the Genomics: GTL program, as outlined in the 2005 roadmap, is an integrated computational environment for GTL systems biology (12). OBER affirms our support for the development of an integrated framework to provide for data sharing, modeling, integration, and collaborations across the program. OBER also recognizes that continued support for development of community driven standard ontologies and data-sharing policies is inherent to the successful implementation of a systems biology network.

V. Computational Software

The International Society for Computational Biology (ISCB) recommends that funding agencies follow ISCB guidelines for open-source software at a "Level 0" availability. ISCB states that research software will be made available free of charge, in binary form, on an "as is" basis for non-commercial use and without providing software users the right to redistribute. OBER will follow ISCB recommendations at a Level 0 availability. OBER recommends that research software developed with GTL funding that result in a peer-reviewed software publication is to be made accessible through either an open source license (www.opensource.org) or deposited to an open source software community such as SourceForge.

VI. Laboratory Information Management Systems (LIMS) for Data Management and Archiving

GTL systems biology research projects involve high-throughput, data intensive research that necessitates use of a data management system to automatically handle this pipeline of data. OBER's goal is that researchers within the GTL program utilize a LIMS system for managing their research data and information. Because different research agendas require different information management systems, an overarching and restrictive policy could place an undue burden on PIs. Therefore, we expect that research projects that involve more than one senior investigator will be required to implement a LIMS or a similar type of electronic system for data and information archiving and retrieval. This plan should balance the clear value of data availability and sharing against the cost and effort of archive construction and maintenance.

VII. Summary

This document outlines the Genomics: GTL program policy and will require GTL funded principle investigators to construct an information and data-sharing plan as a component of their projects. The policy requires information to conform to existing community recognized standard formats wherever possible, to be clearly attributable, and to be deposited, in a timely manner, within a community recognized public database(s) appropriate for the research conducted. OBER is committed to encouraging development of public repositories and standard ontologies for the GTL research community. OBER recognizes that this policy necessarily will be revised to include new standards, data types, and other advances that are pertinent to maximizing availability of data and information across the GTL program. This information and data-sharing policy and related materials can be found at genomicsgtl.energy.gov/datasharing.

References

1. E-science refers to large scale science that is distributed through global collaborations and enabled by the Internet. (see www.research-councils.ac.uk/escience).
2. National Research Council. 2003. Sharing publication related data and materials: Responsibilities of authorship in the life sciences. The National Academy Press, Washington DC.
3. Uhlir, P. F. and P. Schröder. 2007. Open data for global science. *Data Science Journal,* 6:OD36-OD53.
4. Camon, E., M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. 2004. The gene ontology annotation (GOA) database: Sharing Knowledge in Uniprot with gene ontology. *Nucleic Acids Res.*, 32:D262-D266.
5. Edgar, R., M. Domrachev, and A. E. Lash. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30:207-210.
6. Brazma, A., H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-

Serra, and S.-A. Sansone. 2003. ArrayExpress-A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, 31:68-71.

7. Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein and J. M. Cherry. 2001. The Stanford microarray database. *Nucleic Acids Res*., 29:152-155.
8. Taylor, C. F., N. W. Paton, K. S. Lilley, P.-A. Binz, R. K. Julian Jr., A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. R. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yattes III, and H. Hermjakob. 2007. The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, 25:887-893.
9. Orchard, S., L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, V. Stumpflen, A. Ceol, A. Chatr-Aryamontri, J. Armstrong, P. Woollard, J. J. Salama, S. Moore, J. Wojcik, G. D. Bader, M. Vidal, M. F. Cusick, M. Gerstein, A.-C. Gavin, G. Superti-Furga, J. Greenblatt, J. Bader, P. Uetz, M. Tyers, P. Legrain, S. Fields, N. Mulder, M. Gilson, M. Niepmann, L. Burgoon, J. De Las Rivas, C. Prieto, V. M. Perreau, C. Hogue, H.-W. Mewes, R. Apweiler, I. Xenarios, D. Eisenberg, G. Cesareni, and H. Hermjakob. 2007. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology*, 25:894-898.
10. Prince, J. T., M. W. Carlson, R. Wang, P. Lu, and E. M. Marcotte. 2004. The need for a public proteomics repository. *Nature Biotechnology*, 22:471-472.
11. Garwood, K., T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C. F. Taylor, K. Carroll, C. Evans, A. D. Whetton, S. Hart, D. Stead, Z. Yin, A. J. P. Brown, A. Hesketh, K. Chater, L. Hannson, M. Mewissen, P. Ghazal, J. Howard, K. S. Lilley, S. J. Gaskell, A. Brass, S. J. Hubbard, S. G. Oliver, and N. W. Paton. 2004. PEDRo: A database for storing, searching and disseminating experimental proteomics data. *BMC Genomics,* 5:68.
12. U.S. Department of Energy Office of Science Office of Biological and Environmental Research. 2005. *Genomics:GTL Roadmap: Systems Biology for Energy and Environment* (see http://genomicsgtl.energy.gov/roadmap/index.shtml).

Ecology and Environmental Science

Digital data associated with publications are provided as supplementary files, deposited on a website site, and or deposited with DOE Carbon Dioxide Information Analysis Center (CDIAC).

- CARBON DIOXIDE INFORMATION ANALYSIS CENTER (CDIAC)

  The basic CDIAC data policy is that data are available without charge to anyone. Users are encouraged to acknowledge the original data providers in their publications, analyses, and presentations using suggested citations provided by CDIAC.

  An example of such a citation provided by CDIAC is

Houghton, R. A., and J. L. Hackler. 2001. Carbon Flux to the Atmosphere from Land-Use Changes: 1850 to 1990. ORNL/CDIAC-131, NDP-050/R1. Carbon Dioxide Information Analysis Center, U.S. Department of Energy, Oak Ridge National Laboratory, Oak Ridge, Tennessee, U.S.A. doi: 10.3334/CDIAC/lue.ndp050

CDIAC link - http://cdiac.ornl.gov/

Recommended data practices link - http://daac.ornl.gov/PI/pi_info.shtml

Subsurface Environmental Science

Subsurface environmental science datasets are very diverse, commonly including hydrological, geochemical, microbiological, geological, and geophysical datasets. Although the community can and does leverage on community tools available for archiving and dissemination of microbiological datasets, a common community digital data archive for hydrological, geochemical, geological or geophysical data associated with environmental remediation research does not exist in the DOE complex (or even in the wider community). Data dissemination occurs in the form of publication in peer reviewed and grey literature.
There are some exceptions and new DOE activities moving in the direction of open access and systematic protocols. For example, DOE Legacy Management (LM) environmental data associated with closed sites are stored on an LM database (Geospatial Environmental Management System (GEMS) database), http://gems.lm.doe.gov/imf/ext/gems/jsp/launch.jsp, which has open access. The Advanced Simulation Capability for Environmental Management (ASCEM), http://ascemdoe.org/, open source data management tool is being developed with funding from Environmental Management to develop systematic approaches for data input and management (ascemdoe.org), although this tool is still in a very early stage of development it is expected to be useful for facilitating management and dissemination of data associated with BER subsurface biogeochemistry research. datasets. With respect to subsurface geochemical modeling, thermodynamic and sorption databases have been developed globally through a variety of institutions (e.g. Nuclear Energy Agency: http://www.oecd-nea.org/dbtdb/; Japan Atomic Energy Agency http://migrationdb.jaea.go.jp/home_e/home_e.html) and community efforts (http://www.geology.illinois.edu/Hydrogeology/thermo/thermo_phreeqc.dat; http://www.geology.illinois.edu/Hydrogeology/thermo/thermo.com.v8.r6+.dat). However, these are the result of data mining activities (focused on peer-reviewed and grey literature) and not the result of direct interaction of the research community with a data repository or database. Commitments to the ongoing support of these types of databases is often uncertain.

Over the next year, the DOE Subsurface Biogeochemistry (SBR) program intends to develop a more unified approach to supporting the development of community data bases and to work with partners within the DOE and other federal agencies to develop tools for better managing these community data bases. To this end, a first workshop was held in April 2011 to discuss the establishment of data archiving and dissemination protocols for subsurface environmental remediation datasets. Attendees included DOE Subsurface Biogeochemistry (SBR) program managers, key SBR PIs from national laboratories who lead large collaborative projects, and other investigators who represent similar activities outside of DOE (such as from the NSF Critical Zone Observatories). Many topics were discussed, including: identification of high priority datasets for archiving; formats and metadata; opportunities for enhancing synthesis using archived datasets; protocols and incentives for data archiving; and development and maintenance

of webtools to enable data upload and dissemination. Significant discussion was devoted to exploring how DOE environmental subsurface science could leverage on tools that are currently being developed or will be developed soon, such as through the ASCEM or knowledgebase (KBase) tools. The workshop report and more information about this effort will be made available through the SBR website (http://science.energy.gov/ber/research/cesd/subsurface-biogeochemical-research/). DOE SBR has recently released a university FOA, which describes an incipient data sharing policy and requires that applications submitted in response to the FOA must include a description of the applicant's data sharing plan and a commitment to contribute data to the community database, once developed.

Nuclear imaging instrumentation, Radiotracer chemistry, Biological imaging

There is no resource or common practice for the sharing and distribution of digital data beyond investigator to investigator collaboration. In some cases, all raw data is password protected and accessible only to the investigators.

How is access provided and controlled?

Many of the data access practices are included in the discussion above. In general, access could be provided through commercial or not-for-profit publishers or databases including archives, websites, and agency repositories.

Access to published journal articles is determined by the policies of the publishers in whose journals the articles appear. Research indicates that, whether there are access controls or not, the vast majority of researchers in the US and elsewhere do have access to the vast majority of online scholarly literature, although that access is often via institutional subscriptions to journals and/or personal memberships in scholarly societies publishing journals. That said, many publishers are adopting or experimenting with dissemination mechanisms that render articles immediately available upon publication, usually when the author (or her institution) pays a supplemental fee to ensure open distribution. Although such mechanisms may ensure the free availability of the articles (and indeed, freely available articles tend to be accessed more frequently) there is no evidence that they are cited any more frequently than articles that remain under access control.

In many instances, publishers that disseminate articles under access control do so for only a limited period of time, which they establish based on the basis of their understanding of their communities, their business needs, and so on.

In the case of many scientific societies, journals have an "author choice" option that allows authors to pay an additional fee to render their article freely available immediately upon publication. For example, one of the American Society for Plant Biology journals is currently operating a program whereby corresponding authors who are members of the society enjoy complimentary immediate free access to their articles as a benefit of their membership (in effect, an author choice fee of $115).

A consensus is developing among scientific society journals to control access for the first 12 months. The feeling is that this is an appropriate balance between the society's wish to broadly disseminate the articles it publishes and its need to recoup costs through selling site licenses.

Such policies apply to journal content in its entirety. Additional mechanisms used by ASPB (and many other publishers) include single article purchases (so-called "Pay Per View"), which is

offered at a reasonable rate ($10), and – more recently – article "rentals", which allow viewing of non-downloadable article versions for a very nominal fee (typically $1 or so).

Publishers may also choose to make articles freely available to certain institution types and/or in certain parts of the world. For example, upon application from the institution, American Society for Plant Biology will provide free access to up to 10 minority serving institutions per year in the US. The society also participates in UN- and World Health Organisation-managed schemes through which journal content is provided for free to institutions in much of the developing world.

Several additional examples of DOE funded repositories include:

- the Earth System Grid (ESG) for climate simulation model and observational data dissemination, http://www.earthsystemgrid.org/home.htm;jsessionid=2501142E5DC66BDB82365BCE6D969F3E. Most of climate modeling groups in the world use this system to made model available to hundreds of researchers if not thousands.
- DOE-sponsored publications are assigned are submitted for dissemination to the DOE Office of Scientific and Technical Information (http://www.osti.gov/). Unfortunately, the format for the submissions is that of a word document prior to publication layout (i.e., often double sided text with figures at end of manuscript). As such, it appears as an informal web-posted document, which does not have the citation on it and is often not representative of the final, published version. To the uninformed, this would appear to be grey literature (in fact, the posted file names often include an acronym for 'report'), which potentially minimizes the use and impact of disseminated DOE research results.

Access to articles published in American Geophysical Union (AGU) journals is determined by AGU policies; AGU makes the metadata of all articles, including the abstracts, published in our journals freely available upon publication. Recent research indicates [1] that, whether there are access controls or not, the vast majority of researchers in the United States and elsewhere do have access to the vast majority of online scholarly literature via institutional subscriptions to journals and/or personal subscriptions as a member of a scientific society, of which AGU is an example. AGU has a long-established program that offers authors (for a fee) the opportunity to make their article freely available without access control system. This program has the attraction of providing unfettered public access to specific articles, but to date very few authors are taking advantage of the program. Despite the toll-free availability of the articles across many publishers there is no evidence that they are cited any more frequently than articles available through subscriptions.

AGU disseminates articles through subscriptions, and we keep those articles under access control on the basis of our understanding of our communities, our business needs, the "shelf life" of articles, and so on. In many of the fields in which AGU publishes, articles have very long periods of usage (tracked on the AGU Web site) and citations by authors, in many cases exceeding 10 years. AGU subscriptions include not only the current year but also all content back to 1996. All of the older content from all of our journals has been converted to digital format because of continued demand for such older materials.

AGU also disseminates content via single article purchases (so-called "Pay Per View"), which allows viewing of nondownloadable article versions for a very nominal fee.

Journal articles are deemed acceptable only after passing a rigorous peer review evaluation. All AGU journal articles are published online as soon as possible after they have been accepted for publication. Subsequent to acceptance, AGU invests heavily in every article to improve the presentation of content by using professionals (either staff or contractors) to compose the pages for print and electronic publication and to perform copyediting or technical editing, reference verification and correction, insertion of tags to create online links, enhancement and standardization of (in some cases complete preparation of) illustrations or special graphics, typesetting, XML coding for Web dissemination and layout, visual enhancement, reference linking, metadata tagging, indexing, and other technical and editorial input. AGU monitors postpublication citations from other articles to AGU articles and adds forward citations for those articles.

Digital data that are included in articles are accessible under the same conditions as other article content. Access to data stored in specialty repositories or in source repositories such as NASA is accomplished by a robust linking process.

Is access limited in any way?

For both written findings and digital data, the distribution could be limited by, for example, subscription fees, technological barriers, by request only, or limited to the members of a particular research group. Furthermore, access may be exclusive for a limited period of time. For example, access to CESM climate model output can be restricted in accordance with the CESM Data Management and Data Distribution Plan, which allows for a period of up to one year for restricted access to "experiment" and "production" simulations. During this one year restricted time period, access may be granted to other collaborators at the discretion of the owner of the data.

Within subsurface science, there is no community digital data archive for hydrological, geophysical, or geochemical data associated with environmental remediation research in the DOE complex (or the wider community). Many contaminated DOE sites have their own internal data repositories whose access is limited to the research community involved in the remediation effort. These databases are used primarily for regulatory purposes. Collaborative research projects may also have their own database and sharepoint sites, but consistent formats and protocols for these databases do not exist, and access is usually limited to participants. Open access thermodynamic databases are available. However, these are produced as a result of data mining of peer reviewed and gray literature, as described in the previous section.

Another interesting example of a potential unintentional data access limitation is the case of DNA sequence information. While there are not specific limitations beyond those noted above, the problem of excessive DNA sequence data, beyond economic storage capacity is upon us. There is a point at which the value of some of these data versus their storage/retrieval cost must be considered. The impact of the sequencing technology is also seen when considering microarray data, RNA-Seq (transcriptome profiling through deep-sequencing technologies), and the multiplier of application of these tools to metagenomics and communities. Soon proteomics, metaproteomics, metabolomics, and interaction models may reach similar storage and accessibility decision points.

In the case of AGU journal articles, access is controlled by AGU policies and by decisions of authors and their funding agencies. There is variety among funders as to whether or not they specifically earmark grant funds to cover the cost of publication, and even when such funds are

available, many authors do not choose to fund the immediate and open availability of their articles. This means that almost all of the content in AGU journals (and over 90% of all STEM journals) currently can be accessed only after payment of a subscription fee at either the institutional level or member level or by purchase of a single article. Since the transition to online access 10 years ago, AGU has seen a significant increase in access to AGU journals. Previously, access was limited to the print version, and print versions could be used by only one researcher at a time; in the digital age, articles are accessed by a variety of information consumers at all hours and from a wide variety of locations covered by both institutional licenses and individual subscriptions.

Digital data are not restricted by AGU, providing that the author submits data to AGU for publication. However, there are some agencies that control source data and do not make that data available without restriction, while others may choose to protect data that might need to be adequately validated prior to public release or that may require completion of patent applications to protect proprietary interests. This places AGU in the unfavorable position of selecting manuscripts for publication on the basis of data availability. AGU expects that materials and data published in an article should be made available upon request to anyone requesting them, but AGU does not have the resources to "police" author behavior in this area, although egregious misbehavior might render an author in violation of ethical precepts. [3]

Some data are provided to AGU as supplemental or supporting information to journal articles. NISO and NFAIS are already working on developing standards (see http://www.niso.org/workrooms/supplemental) to standardize the way in which such data are published and curated. It is an appropriate—and increasingly urgent—role for science funding agencies, such as DOE, to work with standards and publishing communities to develop and promulgate such standards.

Does access come with any additional functionality?

Perhaps the first and most pervasive interoperable device – established through the operations of a non-profit consortium of scholarly publishers, academic libraries, and others called CrossRef – was toll-free linking to full-text articles from the reference sections of participating journals. This kind of functionality has extended considerably, to include links to underlying database objects and many other kinds of data, to the point that journal articles now serve both as a kind of adhesive that brings together dispersed information, as well as a lubricant that makes it easier for users/readers to find related information of different types.

That said, the establishment of more robust metadata standards would render this kind of functionality much more pervasive and useful. For example the GSC recently published its data standards (Yilmaz. et. al. Nature Biotechnology 29, 415-420 (2011) doic 10.1038/nbt. 1823. Published online 06 May 2011) for metadata accompanying DNA sequence data from soil, water, sediment and plant associated environments, to name the DOE relevant habitat types. One goal of the standards is to facilitate searches across habitat types. It should also be feasible to establish metadata standards that would allow journal articles to link to information on Multimedia are used more frequently by some journals than others; most in the biological sciences have the capacity to host and provide access to peer reviewed multimedia files.

Establishing robust interoperability among datasets and between datasets and journal articles is in early phases in biology, although such connections are more common in other areas of science (especially astrophysics). Again, federal agency attention through the development and

establishment of metadata structures, would be tremendously helpful in advancing this work in biology.

The version of the written material or data provided?

Publishers publish and retain responsibility for the Version of Record (VoR) on their own websites. As mentioned above, though, publishers, too, can publisher earlier versions – although if this is done, those versions are supplanted by the VoR upon the latter's publication.

As noted, other versions may be in circulation or available on other websites (e.g., institutional repositories; PubMed Central). These versions are neither final nor stewarded by their publishers, and – even if they've passed peer-review muster – are sub-optimal at best. This is especially the case if post-publication modifications or changes occur – such as the formal correction or retraction of a published paper.

CrossRef (mentioned above) is working on releasing a validated and robust tag, dubbed CrossMark, to indicate the VoR. Again, CrossRef is an independent organization comprising publishers, librarians, and others that is improving the utility of scholarly information by identifying and solving problems collaboratively.

At present, data associated with journal articles represent a "snapshot", whether or not the underlying dataset continues to evolve. They are the particular set of data that led to (and, per peer reviewers' assessments) underlie the conclusions drawn in the paper.

In climate modeling, work is underway to provide digital object identifiers (DOIs) for CESM model output, so any dataset can be properly referenced and cited. The CESM Data Management and Data Distribution Plan also clearly delineates the "raw" model output from the various kinds of postprocessed data.

In environmental science in general, DOIs are not generally available for published data sets but could be a valuable attribute for ensuring that data are readily linked across multiple publications and for providing both visibility in publication databases and authorship credit. DOI attribution, whereby datasets with DOIs have authors and are indexed in publication databases such as ISI and PubMed, allows datasets to be efficiently cited when used in more than one publication and provides investigators an opportunity to receive citation credit for publishing data on-line – an important incentive not now present.

The criteria for dissemination depend on the form of information, which generally falls into two categories: progress reports on grants submitted to DOE (and other agencies) and formal publication in journals.

Dissemination of the former is achieved via the agency Web sites, and dissemination of the latter is achieved through publishers. DOE does an excellent job of hosting an online database of technical reports (from DOE and many other federal agencies) via the science.gov site hosted by DOE's Office of Science and Technology Information (OSTI). DOE has the responsibility for determining which of the reports are disseminated and under what circumstances.

Grantees achieve formal publication in journals by submitting manuscripts to publishers to be considered for publication. American Geophysical Union (AGU) editors, who are experts in their field, select manuscripts on the basis of a rigorous peer review process. This process is administered by AGU but conducted by scientists acting in two roles, as editors of journals and as reviewers of articles. AGU maintains a significant investment in software systems to make

the peer review process efficient and to speed up the time to quickly disseminate manuscript submissions to article publication.

All journal articles published by AGU are available online at www.agu.org. AGU also permits authors to comply with regulations at their institution by placing their final published article in their institutional repositories after a six month embargo period. Institutional repositories are dispersed across many Web sites and are not currently recognized as an adequately organized or easily searched alternative to publishers' Web sites.

There do not appear to be any specific practices for the maintenance of data beyond that which supports peer review publications beyond the broad data archive exceptions noted above.

Is peer review a condition of dissemination?

Internal (institutional) and external peer review is required for publication of research results in the open literature. Dissemination of preliminary data may occur within and outside the institution for collaborative purposes. Selective dissemination of data or written material may occur as deemed necessary for improving and refining interpretation of research results prior to open literature publication.

Generally, digital data, beyond that included as part of a peer-reviewed publication, is not peer reviewed. One institution-specific exception was noted. Digital data may be shared with external collaborators for purposes of scientific research and project execution, but any distribution of data beyond collaborators must be approved through the formal information release process. A Laboratory procedure is required for establishing a public site for dissemination of digital data, and includes internal peer review for general content and data quality. Updates to data sets may be made within the scope of the initial release, but addition of substantially new or different data requires additional review. A graded approach to information release is followed for digital data (as for written information). Certain types of digital data (including simulation outputs from so-called "Safety Software") may be subject to higher levels of review if they are to be used as the basis of management decisions or risk assessments. These quality assurance (QA) requirements are specified during the project planning stage through the Electronic Prep and Risk tool. Software tools used for such analyses are subject to extensive QA requirements including formal configuration management and metadata recording. Digital data in support of research publications may be published as supplementary material and are subject to the standard publication peer review process.

If there are specific policies, the institution, DOE user facility, or other body by which the policy is currently upheld?

Several examples of DOE and user facility policies were noted above. One additional example was noted. There are no barriers to the dissemination of data below the Federal agency level except for those that involve a brief delay for consideration of IP issues. Hence, it is Federal agencies whose policy is the standard. DOE, for example, has well-vetted and publicized policies on data release, that seem to have broad acceptance. The Department of Homeland Security, however, has a publication bar in its contracts that limit many universities from participation.

Whether, in addition to dissemination, long-term stewardship is accounted for by existing policy or practice?

Several specific examples were provided:

- Written and digital records for all projects at our institution must be managed according to an approved written plan. Records disposition at the end of a project is also governed by this plan, and the TRIM system is used for records with required retention greater than 10 years for projects that utilize electronic recordkeeping.
- Similarly, digital data records follow the same requirements as for written findings. Any digital data generated and used to support research findings must be stored and maintained in the electronic filing system of record. The TRIM system or Electronic Laboratory Notebook system may be used, or data may be stored on a local disk as approved under the formal records plan. Formal data backup procedures must be in place to ensure timely recovery of data in the case of disk failure. For long-term stewardship, any non-standard software needed to access digital data must be maintained along with the data for the required time period.
- Larger publishers that operate their own publishing platforms for the journals they publish are constantly undergoing updating and modifications, as are third-party platforms (e.g., HighWire Press) used by hundreds of smaller publishers to host their journals. Contracts between publishers and such service providers typically stipulate the "back compatibility" of current and previously published materials, regardless of the evolution of the underlying software. In some instances, this back compatibility stretches back hundreds of years, although in ASPB's case, it operates only as far as the first issue of *Plant Physiology*, which was published in 1926.
- The CESM climate data management and data distribution plan requires that disseminated data be made available for a period of no less than four years for all the model output, gradually reducing the total volume by 50% over the next three years. Exceptions can be made to this policy as needed.

Provide a brief description of the kinds of digital data that are generated, the size of the data sets, and how they are stored.

Climate science, Atmospheric science, Climate modeling

- CESM output follows the requirements of the CESM Data Management and Data Distribution Plan for data formats and metadata standards. All CESM output is in netCDF format, and follows the Climate and Forecast (CF) metadata standards. Additional information is in the CESM plan. Dataset sizes can range from megabytes (MB) to many terabytes (TB), depending on the specific model configuration. All datasets are archived on tape.
- Model results are floating point numbers with meta-data that describe the results. Data set sizes range from hundreds of megabytes up to terabytes, and will reach petabytes within the coming decade. They are stored on robotic tape systems at data centers.
- Digital data may include data generated by instruments in laboratory, field, or remote sensing platform, or data generated by numerical models such as climate simulations. Data are typically stored in electronic files with individual file size between a few Mbytes to a few Gbytes. Climate simulation data could amount to Terabytes of data consisting of many data files with different files for different variables and/or time periods. Data files are stored in mass storage backup by tapes.
- For aerosol studies, monthly average 3-D data sets for all aerosol types in different size ranges are often provided. The total data from a single model run might be on the order of 1-2Gbyte. If daily data are needed for analysis, the number would increase by 365÷ 12.

- AGU's published policy declares that data cited in support of manuscripts should be publicly available in order to allow other investigators to replicate experiments and to create shared data sets of similar data across many experimental nodes. AGU does not act as a data repository except to the extent that data sources are cited in journal articles.

Genomics, Microbiology, Plant Biology

- DNA sequence data are the principal data type. An example is soil metagenomic data. They are terabase in size and are stored at the DOE Joint Genome Institute and the research institution that generated the data.
- Large DNA sequencing data include metatranscriptomic information that are being generated at different institutions and generally are stored by the individual scientists responsible for obtaining the data.

Ecology and Environmental Science

- Digital data are very diverse in their types, sizes, and methods of storage. Digital data can range from an Excel spreadsheet containing the results of an experiment in tabular form, to large images or animations, to huge binary digital files created as the output of large simulation runs. Smaller datasets in standard formats (e.g., Excel or Matlab) tend to be stored on local disks using standard records management and backup procedures. For very large simulations, many gigabytes or even terabytes can be generated by a simulation. Large data archival systems exist on supercomputing facilities. However, the user must decide which simulation output require long-term archival versus which can be analyzed and then discarded. In any event, the process, software and metadata used to generate the data must be retained so that data can be recreated if necessary. As external disk drive storage costs have dropped dramatically, some large datasets are now stored and/or physically transferred to another location by placing them on an external disk specifically purchased for that purpose. Data management frameworks such as the Geological Sequestration Software Suite (GS3) tool are intended specifically to provide a consistent framework for data management and integrated analysis that can be used either collectively by a community of researchers or selectively by authorized users as necessary depending on the sensitivity of the data.
- The types of data files generated are
    - Atmospheric trace gas and climatological data
    - Biological and soil chemical and isotope analyses
    - Biological and surface landscape hydrological data
    - GIS map products (raster files)
    - Data are small in size, being generally less than 250KB each.

Subsurface Environmental Science

- Laboratory experimental results yield datasets that are typically small (<1MB) but very diverse. An exception is microbiological datasets, which are large as is described above.
- Wellbore measurements of hydrogeological or environmental parameters (water levels, concentrations, lithology, injection rates, etc) are on the order 1-100 MB and are typically stored on desktop/laptop computers in excel spreadsheets or as ascii documents.

- Geophysical measurement files (such as downhole, crosshole or surface seismic, radar or electrical records) are typically on the order of 10 MB each and are stored as SEG-Y or ascii files or within geophysical data processing packages on desktop/laptop computers.
- Numerical simulation output (hydrological or geophysical) which are often on the order of 50 MB and are stored on desktop/laptop computers.

Nuclear imaging instrumentation, Radiotracer chemistry, Biological imaging

- Imaging files are generated that can be several mByte each. Ten to twelve of these datasets are generated weekly.

ADDITIONAL COMMENTS (by individual BERAC members)

What is your perspective on which dissemination models, if any, successfully maximize the potential benefit of research results in a way that is sustainable within the research community?

Climate science, Atmospheric science, Climate modeling

- For climate model data the community of researchers and users have evolved a workable system through the ESG. In the past the users would have to contact each group to request the data and have to fight through different definitions and formats. All of details have been worked out through the use of the ESG.
- I am a strong advocate for open access publications. When the taxpayers pay to have the research performed, they should not have to pay again to access the results.
- For written results, journal publication maximizes the impacts of research results and potential benefit to the research community and general public.
- For digital data, public access through web download maximizes the use of the data, which also increases potential feedbacks from users of the data to the developers for further improvements.

Genomics, Microbiology, Plant Biology

- Establishing policies that provide links to journal article and involving publishers in discussions about how such policies would operate – is, in my opinion, the most robust, user-friendly, cost effective, and sustainable approach. It is agnostic regarding business model, although would need to accommodate publisher-established access limitation periods, either by incorporating those established at the whole-journal level by the publisher and/or by making use of more innovative approaches (e.g., article "rentals"; toll-free links from specified websites, and so on). Access would be provided to the definitive article version, obviating the need to duplicate effort (and expenditure) by establishing agency-operated databases. Importantly, because different journals in different fields are published at different rates (e.g., 12 monthly issues per year versus 52 weekly issues) and because the measureable online "half-life" of journal articles varies by discipline (generally shorter in the biomedical sciences and longer in the humanities), agency policies re public access to journal articles should not adopt a "one-size-fits-all" approach. Moreover, public access need not be immediate.
- Given that most researchers in most places have ready access to the journal articles they need, working with publishers and others to develop broad utility metadata standards would have a more lasting and potent impact on the progress of scholarship than imposing access policies on publishers.

- The advent of epublishing has generated an unprecedented expansion of scientific publication venues, some of which are more interested in profitability than credibility. We must be continuously vigilant to maintain scientific standards. This situation underscores the value of peer review in maximizing the potential benefits from research results.
- The DNA data deluge coming with the second and third generation of sequencing calls for a new view of how public access can be handled in a useful and cost effective way. Clearly the computation processing, data reduction and visualization are already and will even-more-so, be barriers to useful "dissemination". So, its not a model in the sense I think of models, but the tools to make the data useful.

Ecology and Environmental Science

- For written findings, the open peer-reviewed literature process is the most beneficial and ensures quality and continual evaluation of research results. More detailed information can often be provided in the form of technical reports supporting a concise journal article, or by publication of supplemental information depending on the policy of the publisher.
- For digital data, the processes for disseminating information are less well defined. Current work is underway to develop scientific workflow management tools that will support the collection and dissemination of data and simulation results from complex simulation processes, including metadata and provenance information.
- The dissemination model must include easy-to-use means of going through the vast data to find the data sets of interest.

Subsurface Environmental science

- Subsurface environmental science has a great opportunity to take advantage of the variety of protocols, approaches and tools that are being developed through related projects for data transfer, storage and synthesis.
- A community consensus is needed on types of data that should be archived, as well as location and protocol for archiving (including QA/QC, provenance, metadata, formats).
- Data management and dissemination schema should be developed through close partnership between scientific community, sponsors, PIs, and information experts.
- A plan for a community repository also needs to be developed, which will require long term staffing and funding.
- Needed is community recognition and acceptance that new archiving protocols will increase cost and legwork associated with research projects but will eventually facilitate less expensive synthesis, reanalysis, and discovery.

Nuclear imaging instrumentation, Radiotracer chemistry, Biological imaging

- eJournals with supplemental material on line work well for our group. It would be a hardship if we did not have institutional access to a digital library.

Include any observations regarding opportunities where public access policies or practices could enhance the discovery potential of Office of Science research results.

Climate science, Atmospheric science, Climate modeling

- I think our community of users will general quite satisfied that we have a workable system. I do not hear of complaints about the lack of access. I believe the Office of Science has seen a need for a formal structure and funded its creation and maintenance.

- Open access to publications and data can increase the number of people working with it, and this can lead to additional results.
- Digital data that come with software to read and perform simple analysis and visualization of the data enhance their potential use by the public.

Genomics, Microbiology, Plant Biology

- A reciprocal approach may work best here. By working with other federal agencies and other stakeholders to establish broader metadata standards, DOE will be in a position to push the development of metadata standards that will allow robust linking between agency-operated databases (e.g., of funded grant abstracts/proposals) and journal articles. Such links would operate in the opposite direction, of course, so that readers of the articles could link directly to information about the grant(s) that they cite as supporting the reported research.
- Office of Science policies have been forefront in providing access and usability of digital data, while not disengaging the originator providing added value to the data.

Ecology and Environmental Science

- Movement toward open-source software will facilitate reproducibility and usefulness of digital data by opening access to other investigators. Public access to digital data (including access by scientific investigators not involved in the project) is currently often limited and not nearly as efficient as the dissemination of written findings.