



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Data Initiatives in Federally Funded Research

BERAC Meeting
Gaithersburg, MD
June 6, 2012

Jonathan Petters
AAAS Science and Technology Policy Fellow
DOE Office of Science

Dealing with Data: Challenges and Opportunities

- More data is being collected than we can store
- Many data sets are
 - too large to download
 - too poorly organized to be usable
 - Many data sets are heterogeneous in type, structure, semantics, organization, granularity, accessibility ...
- Utility of data is limited by our ability to interpret and use it



<http://www.sciencemag.org/site/special/data/>



Response to Dr. Brinkman's Charge -Feb 2011

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

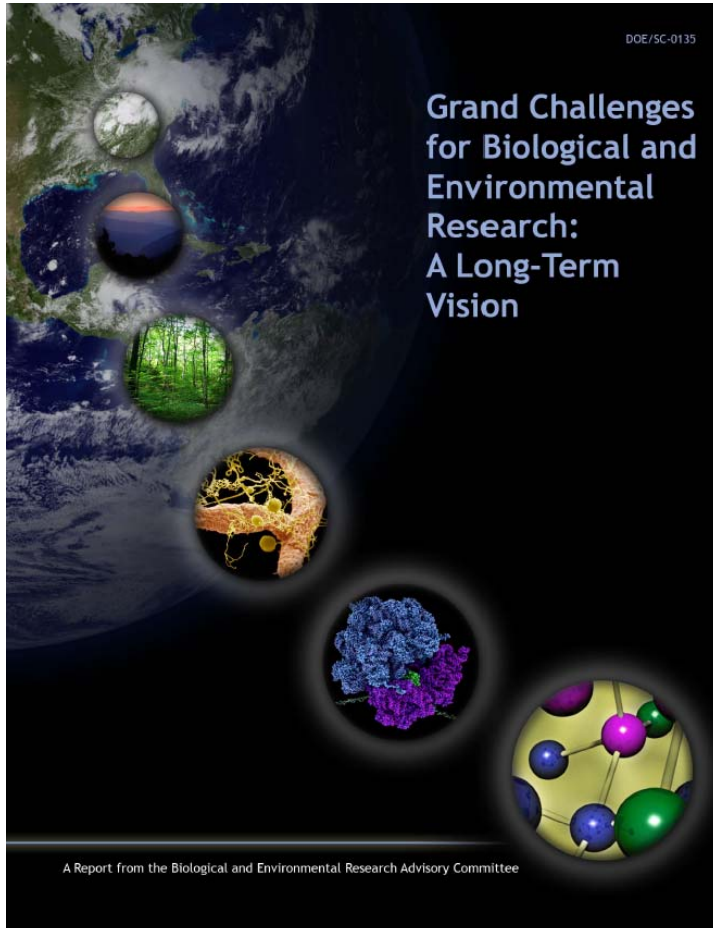
A Report of the Biological and Environmental Research Advisory Committee

Approved June 17, 2011

- Charge in response to COMPETES Reauthorization Act, section 103
- Interagency Working Group on Digital Data formed to “coordinate federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research supported by funding from federal science agencies”



Grand Challenges in Computing for BER



- Establish a new data management paradigm...
- Create a new publishing paradigm...
- Develop new computing paradigms...
- Standardize experimental and computational protocols and methods...
- Improve data usability and model accuracy...
- Design and build software solutions...
- Develop virtual laboratories and tools...

March 2010



U.S. DEPARTMENT OF
ENERGY

Office of
Science

National Big Data Research and Development Initiative

Big Data Senior Steering Group – chartered in spring 2011 under the Networking and Information Technology R&D (NITRD) Program

- Members from DARPA, DOD OSD, DHS, DOE-Science, HHS, NARA, NASA, NIST, NOAA, NSA, and USGS, Treasury
- Co-chaired by NIH and NSF
- Long-term, National initiative with four major components
- Initiative announced March 29th, 2012



U.S. DEPARTMENT OF
ENERGY

Office of
Science

National Big Data Research and Development Initiative

- **White House Big Data R&D Initiative announced March 29, 2012**
 - Six Federal departments and agencies announced more than \$200 million in new commitments ...to...improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.
 - Office of Science announced the creation of the SciDAC Institute of Scalable Data Management, Analysis and Visualization (SDAV Institute)
www.sdav-scidac.org



National Big Data Research and Development Initiative

Core Technologies - Foundational research to develop new techniques and technologies to derive knowledge from data

BIGDATA solicitation (joint between NIH and NSF)

To advance the core techniques and technologies for managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets



U.S. DEPARTMENT OF
ENERGY

Office of
Science

National Big Data Research and Development Initiative

Domain Research - New cyberinfrastructure to manage, curate, and serve data to research communities

Group identifying potential data- and computational-intensive research domains where further interagency coordination could be beneficial

Climate-health nexus and NSF's EarthCube project among other possibilities



U.S. DEPARTMENT OF
ENERGY

Office of
Science

National Big Data Research and Development Initiative

Workforce Development - New approaches for education and workforce development

McKinsey report (May 2011)

Looking at building communities of fellows across agencies (ex. DOE Computer Science Graduate Fellowship conference this summer will be made available to similar fellows outside DOE)



U.S. DEPARTMENT OF
ENERGY

Office of
Science

National Big Data Research and Development Initiative

Challenges - Challenges and competitions to create new data analytic ideas, approaches, and tools from a more diverse stakeholder population

Ideation Challenge forthcoming, BER is involved (ARM)

- **responders are asked to generate ideas for applications/tools that would**
 - address participating agencies' missions
 - allow heterogeneous collections of data to become more homogeneous and searchable
 - involve more than one data set (at least one of which is “large”)



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Data Citation and Attribution

ARM has coordinated with OSTI to assign digital object identifiers to datastreams available through ARM website

The screenshot shows the ARM website interface. At the top, there is a navigation bar with links for 'CART', 'Home', 'People', and 'Site Index', along with a search bar for 'arm.gov'. The ARM logo and 'CLIMATE RESEARCH FACILITY' are on the left, and the 'U.S. DEPARTMENT OF ENERGY Office of Science' logo is on the right. Below the navigation bar, there is a menu with tabs for 'About', 'Science', 'Campaigns', 'Sites', 'Instruments', 'Measurements', 'Data', 'News', 'Publications', and 'Education'. The 'Measurements' tab is selected, and a dropdown menu shows 'Location Table' and 'Contacts'. The main content area displays the 'Datastream : RAIN' page, which includes a breadcrumb trail 'ARM.gov >> Data >> Datastreams >> rain', a title 'Datastream : RAIN', and sections for 'Rain gauge', 'Active Dates' (2006.03.01 - 2012.05.27), 'Measurement Categories' (Atmospheric State), 'Originating Instrument' (Rain Gauge (RAIN)), and 'Primary Measurements'. A sidebar on the right contains sections for 'Documentation' (Data Object Design, Data Quality Plots), 'Citation' (DOI: 10.5439/1025264, [What is this?]), 'Order Data' (BUILD AN ORDER button), and 'Comments?' (We would love to hear from you! Send us a note below or call us at 1-888-ARM-DATA. Email Address input field).



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Data Citation and Attribution

ARM has coordinated with OSTI to assign digital object identifiers to datastreams available through ARM website

The screenshot shows the ARM website interface. At the top, there is a navigation bar with links for 'CART', 'Home', 'People', and 'Site Index', along with a search bar for 'arm.gov'. Below this is a banner for the 'U.S. DEPARTMENT OF ENERGY Office of Science'. A main navigation menu includes 'About', 'Science', 'Campaigns', 'Sites', 'Instruments', 'Measurements', 'Data', 'News', 'Publications', and 'Education'. The 'Data' menu is expanded to show 'Location Table' and 'Contacts'. The main content area is titled 'Datastream : RAIN' and provides details about a rain gauge, including active dates (2006.03.01 - 2012.05.27), measurement categories (Atmospheric State), and the originating instrument (Rain Gauge (RAIN)). A 'Primary Measurements' section explains that the following measurements are scientifically relevant and refers to the netCDF File Header Description. A sidebar on the right contains sections for 'Documentation' (Data Object Design, Data Quality Plots), 'Citation' (highlighted in red, showing DOI: 10.5439/1025264 and a link '[What is this?]'), 'Order Data' (with a 'BUILD AN ORDER' button), and 'Comments?' (with an email address input field).



Data Citation and Attribution

ARM has coordinated with OSTI to assign digital object identifiers to datastreams available through ARM website

The screenshot shows the ARM website interface. At the top, there is a navigation bar with links for 'CART', 'Home', 'People', and 'Site Index', along with a search bar for 'arm.gov'. The ARM logo and 'CLIMATE RESEARCH FACILITY' are on the left, and the U.S. Department of Energy 'Office of Science' logo is on the right. A main navigation menu includes 'About', 'Science', 'Campaigns', 'Sites', 'Instruments', 'Measurements', 'Data', 'News', 'Publications', and 'Education'. The 'Data' menu is expanded to show 'Location Table' and 'Contacts'. The main content area displays 'Datastream : RAIN' with details such as 'Rain gauge', 'Active Dates: 2006.03.01 - 2012.05.27', 'Measurement Categories: Atmospheric State', and 'Originating Instrument: Rain Gauge (RAIN)'. A sidebar on the right contains 'Documentation' (with links for 'Data Object Design' and 'Data Quality Plots') and a 'Citation' section (highlighted with a red box) showing 'DOI: 10.5439/1025264' and a link '[What is this?]'. At the bottom, there are contact options for 'Tropical Western Pacific (TWP)' and 'ARM Mobile Facility (GAN)', a phone number '1-888-ARM-DATA', and an email address input field.

Goal is to improve discoverability of and access to ARM datastreams

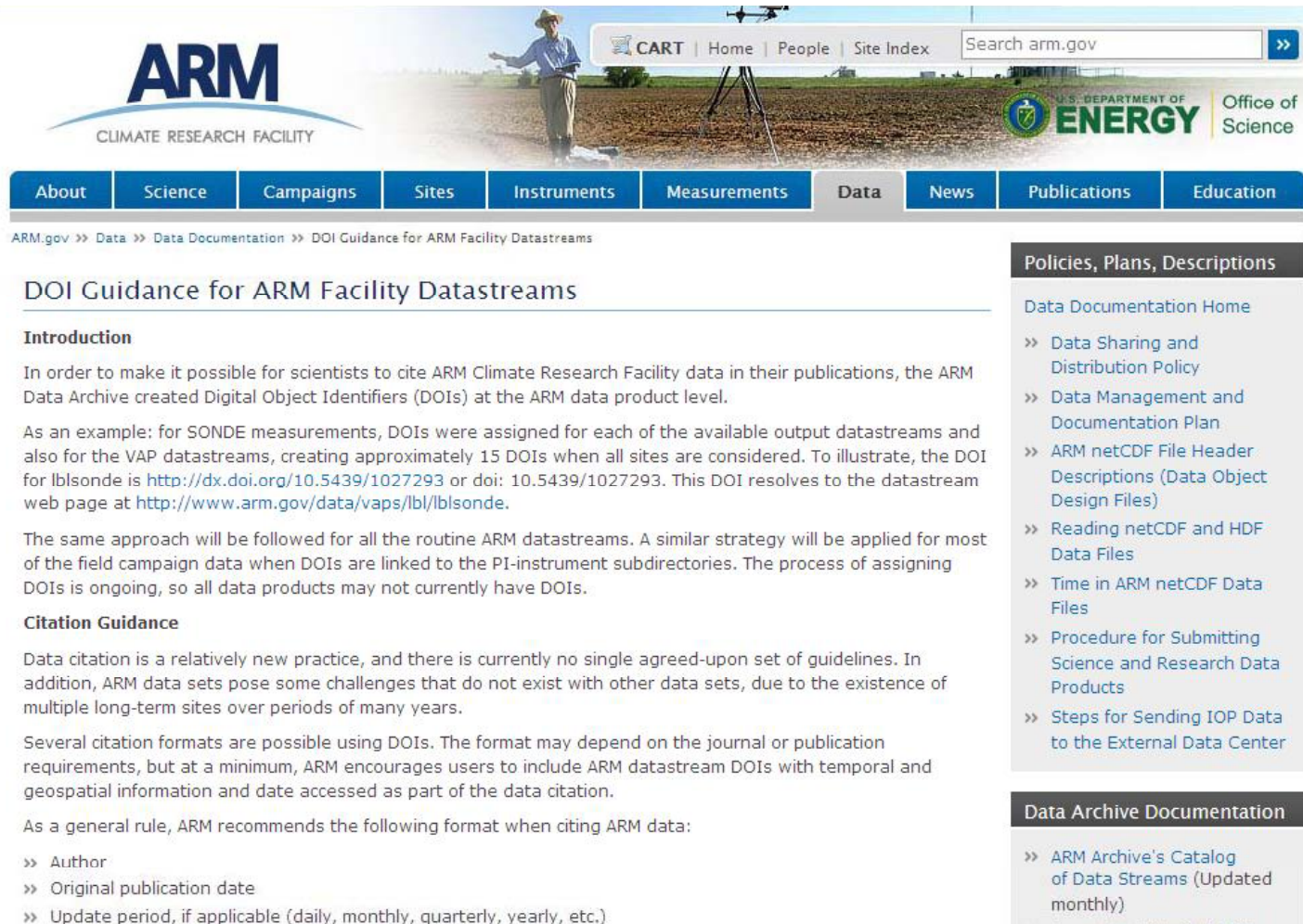


U.S. DEPARTMENT OF
ENERGY

Office of
Science

Data Citation and Attribution

ARM has coordinated with OSTI on data citation for ARM datastreams



The screenshot shows the ARM Climate Research Facility website. The header includes the ARM logo, navigation links (CART, Home, People, Site Index), a search bar, and the U.S. Department of Energy Office of Science logo. A navigation menu below the header lists: About, Science, Campaigns, Sites, Instruments, Measurements, Data, News, Publications, and Education. The main content area is titled 'DOI Guidance for ARM Facility Datastreams' and includes an introduction, citation guidance, and a list of links under 'Policies, Plans, Descriptions' and 'Data Archive Documentation'.

ARM
CLIMATE RESEARCH FACILITY

CART | Home | People | Site Index | Search arm.gov

U.S. DEPARTMENT OF ENERGY | Office of Science

About | Science | Campaigns | Sites | Instruments | Measurements | Data | News | Publications | Education

ARM.gov » Data » Data Documentation » DOI Guidance for ARM Facility Datastreams

DOI Guidance for ARM Facility Datastreams

Introduction

In order to make it possible for scientists to cite ARM Climate Research Facility data in their publications, the ARM Data Archive created Digital Object Identifiers (DOIs) at the ARM data product level.

As an example: for SONDE measurements, DOIs were assigned for each of the available output datastreams and also for the VAP datastreams, creating approximately 15 DOIs when all sites are considered. To illustrate, the DOI for lblsonde is <http://dx.doi.org/10.5439/1027293> or doi: 10.5439/1027293. This DOI resolves to the datastream web page at <http://www.arm.gov/data/vaps/lbl/lblsonde>.

The same approach will be followed for all the routine ARM datastreams. A similar strategy will be applied for most of the field campaign data when DOIs are linked to the PI-instrument subdirectories. The process of assigning DOIs is ongoing, so all data products may not currently have DOIs.

Citation Guidance

Data citation is a relatively new practice, and there is currently no single agreed-upon set of guidelines. In addition, ARM data sets pose some challenges that do not exist with other data sets, due to the existence of multiple long-term sites over periods of many years.

Several citation formats are possible using DOIs. The format may depend on the journal or publication requirements, but at a minimum, ARM encourages users to include ARM datastream DOIs with temporal and geospatial information and date accessed as part of the data citation.

As a general rule, ARM recommends the following format when citing ARM data:

- » Author
- » Original publication date
- » Update period, if applicable (daily, monthly, quarterly, yearly, etc.)

Policies, Plans, Descriptions

- Data Documentation Home
- » Data Sharing and Distribution Policy
- » Data Management and Documentation Plan
- » ARM netCDF File Header Descriptions (Data Object Design Files)
- » Reading netCDF and HDF Data Files
- » Time in ARM netCDF Data Files
- » Procedure for Submitting Science and Research Data Products
- » Steps for Sending IOP Data to the External Data Center

Data Archive Documentation

- » ARM Archive's Catalog of Data Streams (Updated monthly)



Data Citation and Attribution

ARM has coordinated with OSTI on data citation for ARM datastreams



- Author
- Original publication date
- Update period, if applicable (daily, monthly, quarterly, yearly, etc.)
- Dataset name
- **Dates used***

>> Author
>> Original publication date
>> Update period, if applicable (daily, monthly, quarterly, yearly, etc.)

- **Locations*** (latitude/longitude, site name, and facility identifier)
- Editor(s) or compiler(s)
- Place of publication
- Publisher
- **Date accessed***
- **DOI***

>> ARM Archive's Catalog of Data Streams (Updated monthly)



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Giri Palanisamy, ORNL and Mark Martin, OSTI

SC - internal Working Group on DD

- Share information about data issues, policies, and research across SC programs
- Consider pathways to improve SC digital data management (e.g. data standards, access, preservation)
- Formed partially in response to COMPETES 103 and Brinkman charge (Feb 2011) to respond to guidance from **Interagency Working Group on Digital Data**; currently that guidance is....



SC - internal Working Group on DD



U.S. DEPARTMENT OF
ENERGY

Office of
Science

SC - internal Working Group on DD

- In lieu of guidance from above, moving at own (slower) pace
- Learning from other agencies/communities' experiences in data management (including from BER)
- Currently considering options



Thank You!
Questions/Comments?



U.S. DEPARTMENT OF
ENERGY

Office of
Science