

Article

Data Always Getting Bigger—A Scalable DOI Architecture for Big and Expanding Scientific Data

Giri Prakash ^{1,*}, Biva Shrestha ¹, Katarina Younkin ², Rolanda Jundt ², Mark Martin ³
and Jannean Elliott ³

¹ Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831, USA; ivashrestha@gmail.com

² Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99354, USA; soap@pnnl.gov (K.Y.); Rolanda.Jundt@pnnl.gov (R.J.)

³ DOE Office of Scientific and Technical Information, 1 Science.gov Way, Oak Ridge, TN 37830, USA; martinm@osti.gov (M.M.); ElliottJ@osti.gov (J.E.)

* Correspondence: palanisamy@ornl.gov; Tel.: +865-241-5926

Academic Editor: Jamal Jokar Arsanjani

Received: 24 March 2016; Accepted: 9 August 2016; Published: 31 August 2016

Abstract: The Atmospheric Radiation Measurement (ARM) Data Archive established a data citation strategy based on Digital Object Identifiers (DOIs) for the ARM datasets in order to facilitate citing continuous and diverse ARM datasets in articles and other papers. This strategy eases the tracking of data provided as supplements to articles and papers. Additionally, it allows future data users and the ARM Climate Research Facility to easily locate the exact data used in various articles. Traditionally, DOIs are assigned to individual digital objects (a report or a data table), but for ARM datasets, these DOIs are assigned to an ARM data product. This eliminates the need for creating DOIs for numerous components of the ARM data product, in turn making it easier for users to manage and cite the ARM data with fewer DOIs. In addition, the ARM data infrastructure team, with input from scientific users, developed a citation format and an online data citation generation tool for continuous data streams. This citation format includes DOIs along with additional details such as spatial and temporal information.

Keywords: Atmospheric Radiation Measurement; DOIs; Digital Object Identifier; data citation; arm data citation; continuous data citation

1. Introduction

Any public digital resource, including scientific datasets, can be referenced by appropriate information and a Uniform Resource Locator (URL) for the resource. However, URLs are not always persistent, and thus not trusted to be the permanent resource scientists expect for a formal citation [1]. Linking to digital scientific data using persistent identifiers such as Digital Object Identifiers (DOIs) provides stability that can benefit data providers, data users, and publishers. By assigning DOIs to data, scientists link their articles to the exact data used for the research, which is critical if (1) other researchers wish to reproduce the same results; (2) people responsible for generating the data wish to get credit for their contribution; (3) funding agencies wish to assess the usefulness of the data generated by their projects; and (4) publishers wish to link the journal article to the cited data and help readers access the data over the years.

Data citation is not a new concept; many data providers and data centers have been following this practice for a long time. Scientific focus groups and publishers have conducted various working groups and discussions to define best practices and policies for data citation [2,3]. These best practices often deal with datasets that are typically static in nature with DOIs assigned to one or more digital objects. However, for big and continuously growing datasets such as those generated by streaming

data from various sensors, a different granularity is needed. There are very few community-defined policies available for citing large and continuous datasets. These policies are either limited or contain very high-level information without many specifics. Many big data archives and projects will greatly benefit by providing more detailed and proven strategies for citing large, constantly growing datasets.

The objective of this paper is to explain the structure of big datasets that are available in the atmospheric research observational networks and our attempts to come up with a practical solution for linking these datasets with publications using DOIs.

2. Background

Climate change research projects involving observational networks, satellites, and large-scale simulation outputs are data intensive and typically produce large and complex datasets. In many cases, the output data could be multiple terabytes of data in millions of data files. Users who analyze this data have issues with citing these datasets in their journal articles appropriately. Users of the Atmospheric Radiation Measurement (ARM) Climate Research Facility, which provides observational data to study global climate change, would face these same issues were it not for ARM's strategy of handling DOI granularity.

The U.S. Department of Energy (DOE) created the ARM Program in 1989 to develop several highly instrumented ground stations to study cloud formation processes and their influence on radiative transfer [4,5]. As the program evolved, the original ground sites were supplemented with three mobile facilities and an aerial facility.

The ARM program collects and archives (Figure 1) several different types of data: regular instrument data streams, processed data, special collections of data from Principal Investigators (PIs), data from field campaigns, and data from external sources [2]. A data stream is a collection of different variables sampled over the same time interval and packaged together. Information about ARM instruments, measurements, and data products is compiled and presented via the ARM website [6]. ARM data can be discovered and downloaded using the ARM Archive Data Discovery tool, which is available from the ARM Data Archive page [2]. As part of discovering ARM data, users can refine the search to data of interest by using hierarchical keywords grouped in instrument and measurement categories. Additional information such as data plots, time grids, and Data Quality Reports (DQR) also aids the data selection process. Figure 2 shows the workflow of a typical search and access of ARM data using the ARM Data Discovery tool.

The Atmospheric System Research (ASR) community has scientific working groups [7,8] and focus groups [8] that determine the measurements needed to perform research. Based on community input from various working group meetings, these groups collaborate with the ARM data infrastructure group to define the data product and the correct granularity.

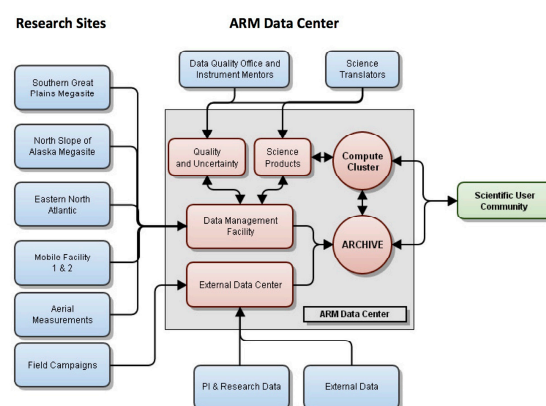


Figure 1. High level—ARM data flow: The blue circles are primary data sources, but data are also generated in various parts of the red circles. All of these data need to be traceable with DOIs.

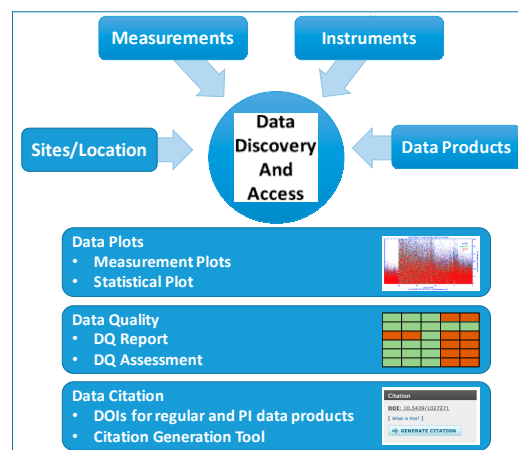


Figure 2. Accessing ARM data using the ARM portal.

3. Using ARM Data

ARM collected data have been used for the study of cloud lifecycles, aerosol lifecycles, radiative processes, and their effect on precipitation for 20 years. Scientists are currently still using ARM data in a variety of ways. The usage ranges from PIs performing in-depth analysis of a few data streams which look at specific atmospheric processes such as examples in Reference [9], to more integrated analysis that uses a large number of data streams [10]. Scientists also use ARM data to improve global climate change models [11].

Traditionally, a unique DOI is assigned to static datasets, then cited using a specific citation structure. The DOI will resolve to a landing page, which may lead to one or more data files, along with ancillary information such as data dictionary and data quality information. This approach to defining the boundary of a DOI works well for many static data products, including PI-contributed products that are available in the ARM Data Archive. ARM has been following existing best practices defined by the Committee on Data for Science and Technology (CODATA) [12], the Joint Declaration of Data Citation Principles by FORCE11, DOE's Carbon Dioxide Analysis [13], and NASA's Distributed Active Archive Center for Biogeochemical Dynamics [14]. However, with ARM's continuous data streams, assigning DOIs for individual files could potentially lead to the ARM Data Archive creating and managing large volumes of DOIs. Citing these DOIs in a journal would also be a challenging or an impossible task for scientists. In addition, this could be unacceptable to most publishers; therefore, the ARM facility needed a new practice and method for assigning DOIs and using citations for continuous data streams.

4. Methodology

Considering the complexity of the ARM datasets defined in the "Background" section, the ARM Data Archive followed a combination of DOI and citation structure to avoid assigning DOIs for each data file and build a scalable architecture for citing data. ARM Data Archive assigns DOIs at the data product level. As an example: for the SONDE (Balloon-Borne Sounding System) measurements [15], the ARM Data Archive assigned DOIs for each of the available output data streams and also for the Value-Added Products (VAPs) data streams, which includes about 15 DOIs when all sites are considered. One of the derived outputs for SONDE measurement is LSSONDE (Microwave Radiometer-Scaled Sonde Profiles). The DOI for the LSSONDE is 10.5439/1027294 or it can be expressed as a hyperlink when it is written as <http://dx.doi.org/10.5439/1027294>. This link will not change, but the underlying URL that the DOI redirects to may. Currently, this link corresponds to the ARM dataset page of lssonde product [16]. This makes the DOI a persistent link to the digital resource.

The authors discussed this particular strategy for assigning DOIs for larger datasets in various working groups such as the CODATA Task Group on Data Citation Standards and Practices (2012, Taipei) [17] and the CENDI-NFAIS Workshop on Big Data [18]. Feedback from these discussions was incorporated in the implementation phase.

Assigning DOIs at the data product level allows ARM to use the same DOIs for new deployments of the same instruments for future ARM sites. As an example: the SONDEWNP data collected at Southern Great Plains (SGP) and the recently deployed ARM Mobile Facility at the McMurdo Station, Antarctica (AMF2), uses the same DOI of 10.5439/1021460.

For ARM field campaign data products and special data provided by PIs, ARM Data Archive applies DOIs and formats data citations using specific metadata provided by PIs. Figure 3 indicates the DOI assignment workflow for these products.

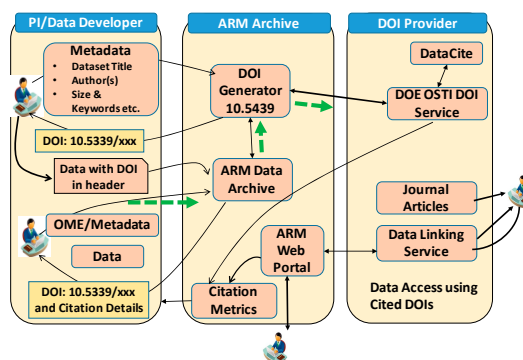


Figure 3. ARM DOI assignment workflow.

PIs work with the ARM Data Archive in two different ways. During the data creation phase, PIs request DOIs from the ARM Data Archive with the minimum metadata required to reserve a DOI from the DOE's Office of Scientific and Technical Information (OSTI). The ARM Data Archive submits this metadata to OSTI's DOI Web Service, obtains the DOIs, and provides them to the PIs. The PIs add the DOIs to headers of data files in different formats such as ASCII and NetCDF [19] as global attributes. Then, they submit the data to the ARM Data Archive for storage and distribution. A second approach occurs as part of the data registration process. PIs use the ARM Online Metadata Editor (OME) ([20], as shown in Figure 3), to create the scientific metadata and upload the data for the review process. After the ARM data reviewers receive the metadata and data, they determine if the submitted data needs DOIs and work with the PIs to assign DOIs at the appropriate level. In this case, ARM reuses the metadata already submitted by the PIs. The ARM data tracking process and ARM management typically reach out to the PIs to make sure they submit their data. If data is not coming to the ARM Data Center, the reserved DOI gets reused for other data products with new metadata.

4.1. Digital Object Identifier Generation

The ARM Data Archive in collaboration with OSTI established a DOI service for the ARM datasets. OSTI is a member of DataCite [21] and acts as an allocating agent to assign DOIs to dataset records submitted by DOE organizations. After a DOI is assigned to the dataset described in the submitted metadata, OSTI then provides the metadata and the DOI to DataCite, where the DOI is minted and becomes resolvable on the web. After that, the DOI will always resolve to the dataset's landing page on the ARM website, and the user can freely order from the ARM Data Archive any of the data file components covered by that DOI.

OSTI requests a unique prefix from DataCite that appears on the front end of every DOI generated for that particular client and provides it to each data client. The dedicated prefix makes it easy for the client to collect citation metrics and other usage information from various sources. The end part or

suffix of the DOI is the unique number assigned to a newly submitted record by OSTI's intake and processing system, Energy Link (E-Link) [22].

For high-volume data centers, such as the ARM Data Archive, OSTI provides a web service for automated submission and ARM implemented a web service client on its end using a Java framework. The ARM script extracts metadata from information provided by PIs and stored in the ARM database. It constructs XML records, which are sent to OSTI, authenticated by OSTI's web service, parsed, link-validated, and loaded into E-Link. If the metadata is successfully loaded, a DOI is immediately assigned and notification is provided back to the ARM server. If there is a problem with the metadata, a DOI is not assigned; instead, error messages are sent back to the ARM server to facilitate correction. The ARM Data Archive resends the corrected metadata when it is ready.

OSTI processes each record that was successfully submitted and registers the DOI and associated metadata with DataCite and will be active within 24 h.

If the data product gets major revisions, this will be released as a higher-level data product. As an example: if the data quality checks apply to original netCDF files (a1 level), the processed data will be published as b1 level data product. The data management and documentation process is explained at ARM documentation page [23]. If minor changes or reprocessing are done to the data, this is captured in the global attribute (header) of the data file. The ARM-recommended citation field "data accessed date" will also help retrieve the data version.

4.2. Citing ARM Data Using Proposed Citation Structure

In addition to DOIs for ARM data products, ARM also provides a recommended citation structure to help users understand how to cite the exact ARM data that they are referencing in their articles. ARM encourages the users to include the ARM data stream DOI, temporal and geospatial information, and date accessed as part of the data citation. ARM continuously reprocesses the current and historical data to address various data quality issues and these revisions are captured in an internal system. Typically, users get the latest processed data from the ARM Data Center. The previous versions of these data files are deep-archived and could be retrieved only for specific requests. The data-accessed date allows ARM to identify and retrieve the specified processed data that the journal article cites. For example, let us assume that a user downloaded the ARM data product Balloon-Borne Sounding System (SONDEWNP) for the temporal range of 1 October 2010 to 30 March 2011 from the Southern Great Plains (SGP) Central Facility on 13 April 2014. The DOI of this data product is 10.5439/1021460. Using the above information, the user will follow one of the citation structures below to cite the ARM data in an article:

Using publisher of data: *"Atmospheric Radiation Measurement (ARM) Climate Research Facility. 1994, updated daily. Balloon-borne sounding system (SONDEWNP). Oct. 2010–March 2011, 36° 36' 18.0" N, 97° 29' 6.0" W: Southern Great Plains Central Facility (C1). Compiled by R Coulter, J Prell, M Ritsche, and D Holdridge. ARM Data Archive: Oak Ridge, Tennessee, USA. Data set accessed 2011-04-13 at <http://dx.doi.org/10.5439/1021460>".*

This structure arranges the information in the following order: [Publisher]. [Data Publication Year]. [Title and Processing Level], [Location], [temporal and geospatial subset used]. [Date accessed] and [DOI].

Using author of data: *"Coulter, Richard., Jenni Prell, Michael Ritsche, and Donna Holdridge. 2010. Balloon-borne sounding system (SONDE), sondewnpn b1 data stream, Oct 2010–Mar 2011, 36° 36' 18.0" N, 97° 29' 6.0" W. Atmospheric Radiation Measurement (ARM) Climate Research Facility Data Archive, Oak Ridge, Tennessee, U.S.A. Data set accessed 2011-04-13 at <http://dx.doi.org/10.5439/1021460>".*

Here, the arrangement of information is in the following order: [Author(s)]. [Data Publication Year]. [Title with Processing Level], [temporal and geospatial subset used]. [Publisher]. [Location]. [date accessed] and [DOI].

Using specific measurements extracted from ARM data files: *"Coulter, Richard., Jenni Prell, Michael Ritsche, and Donna Holdridge. 2010. Balloon-borne sounding system (SONDE), sondewnpn b1 datastream,*

Oct 2010–Mar 2011, 36° 36' 18.0'' N, 97° 29' 6.0'' W, relative humidity. Atmospheric Radiation Measurement (ARM) Climate Research Facility Data Archive, Oak Ridge, Tennessee, U.S.A. Data set accessed 2011-04-13 at <http://dx.doi.org/10.5439/1021460>".

For this structure, the arrangement of the information is in the following order: [Author(s)]. [Data Pub Year]. [Title with version], [temporal and geospatial and measurement subset used]. [Publisher]. [Location]. [date accessed] and [DOI].

These three options allow users to cite ARM data based on author/publisher requirements. For example: if the author/publisher wants to cite it using the data publisher (project) as the data author, they will pick the first option. If the author/publisher prefers to highlight the person involved in generating the data, then they will use option 2. The third option helps users cite specific measurements extracted from the data streams. This flexibility will still ensure the data reproducibility using these citations.

The above examples demonstrate the suggested citation structures, but many other structures are also possible. The data-accessed date is critical for the routine ARM data streams, because the version of data used is determined based on this information. The DOI guidance for the ARM Facility data streams is available at [24].

4.3. ARM Citation Generator

In addition to providing the above citation guidance, ARM also developed an ARM Citation Generation tool to help users create a citation text. Figure 4 shows that the Citation Generator tool is linked from all the ARM data product pages. In this figure, the button "GENERATE CITATION" activates the Citation Generation tool (Figure 5). The tool helps users generate a citation by asking them to answer a few simple questions. This tool also provides the citation currently distributed to users in data notification emails. The Citation Generator tool was designed and developed within the ARM data discovery and delivery workflow.

The screenshot displays the ARM Climate Research Facility website. The header includes the ARM logo, navigation links (Home, People, Site Index), a search bar, and the U.S. Department of Energy Office of Science logo. A secondary navigation bar lists categories: About, Science, Campaigns, Sites, Instruments, Measurements, Data, News, Publications, and Education. The main content area is titled 'Datastream : AOSCP' and provides details about the Aerosol Observing System (AOS), including active dates (2011.03.09 - 2016.03.21), measurement categories (Aerosols), and the originating instrument (Aerosol Observing System (AOS)). A table lists measurements: Aerosol concentration (units: parts/cc, variable: cpc) and Particle concentration, preliminary calibration. Below this, locations are listed for North Slope Alaska (NSA), Southern Great Plains (SGP), and ARM Mobile Facility (MAO, PGH, PVC). Two callout boxes highlight key features: 'DOI and a link to text explaining how DOIs are assigned to ARM data' pointing to the DOI 10.5439/1025152 and the 'GENERATE CITATION' button, and 'Browse data to order specific data files included in this datastream' pointing to the 'Browse Data' button. A 'Comments?' section at the bottom right allows users to send feedback. The footer includes logos for partner organizations like Argonne, Brookhaven, Los Alamos, NREL, Oak Ridge, and Sandia.

Figure 4. ARM landing page.

The ARM data landing page, shown in Figure 4, allows users to browse and order data using the ARM Data Discovery tool and create a citation for the data they ordered.

The Archive's landing pages are an important key to the success of the high-level granularity of ARM DOIs. DataCite recommends that all registered DOIs link users to a landing page rather than directly to the dataset itself. This allows users to refer to further details such as the data dictionary, data plots, data quality information, etc., before they download the data.

The landing pages designed for ARM data do all this and more.

Figure 5. ARM Citation Generation web tool.

5. Summary

The ARM Data Archive is pioneering the DOI concept for large-scale continuous data streams. ARM data citation not only aims to promote access to data, but is also trying to provide proper attribution to the data creators for their efforts. Using this policy, ARM can further improve the discovery and access of ARM data through publications and citations. These citations are a means by which authors can order and expedite ARM data by providing a simple template of information. Since these citations are automatically generated and uniform, they follow certain templates. For example: "Coulter, Richard, Jenni Prell, Michael Ritsche, and Donna Holdridge. 2010. Balloon-borne sounding system (SONDE), b1 data stream, 36° 36' 18.0'' N, 97° 29' 6.0'' W (SGP), Oct 2010–Mar 2011. Atmospheric Radiation Measurement (ARM) Program Archive, Oak Ridge, Tennessee, U.S.A. Data set accessed 2011-04-13 at <http://dx.doi.org/10.5439/1021460>" always provides users the version of SONDE data streams from ARM's Southern Great Plains (SGP) location from October 2010 to March 2011.

We are actively presenting this data citation approach in various forums, including during the ARM science team and DOE's Atmospheric System Research (ASR) working group meetings. This method allows ARM to educate data users on proper data citations and also allows the collection of feedback from scientists to improve the citation structure.

Acknowledgments: The ARM Climate Research Facility is funded through the DOE Office of Science and is managed through the Biological and Environmental Research (BER) Division. ARM is supported by a large number of dedicated individuals—far too many to mention here. We would like to specifically thank Sally McFarlane, Rick Petty, and Ashley Williamson at DOE headquarters as well as the members of the joint ARM/ASR Science and Infrastructure staff (www.arm.gov/about/contacts) for their guidance and support. We would also like to thank DOE Office of Scientific and Technical Information (OSTI) staff for their support in providing DOI service.

Author Contributions: Giri Prakash: Principal Investigator, project design, supervisor, and primary drafter; Biva Shrestha and Katarina Younkin: Software development; Rolanda Jundt: Website customization and communications; Mark Martin and Jannean Elliott: DOI service providers; Laura Aslinger and Alka Singh: Help with citation, formatting, and proof reading.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Callaghan, S.; Donegan, S.; Pepler, S.; Thorley, M.; Cunningham, N.; Kirsch, P.; Ault, L.; Bell, P.; Bowie, R.; Leadbetter, A.; et al. Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *Int. J. Digit. Curation* **2012**, *7*, 107–113.
- ARM Archive User Services. ARM Archive, Atmospheric Radiation Measurements. 1 January 1994. Available online: <http://www.archive.arm.gov> (accessed on 1 March 2013).
- FORCE11. Joint Declaration of Data Citation Principles. FORCE11. 1 January 2014. Available online: <https://www.force11.org/group/joint-declaration-data-citation-principles-final> (accessed on 1 March 2013).
- Ackerman, T.P.; Stokes, G.M. The Atmospheric Radiation Measurement Program. *Phys. Today* **2003**, *56*. [CrossRef]
- Turner, D.D.; Ellingson, R.G. The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years. *Am. Meteorol. Soc.* **2016**, in press.
- ARM Climate Research Facility. Atmospheric Radiation Measurement, ARM Archive and User Services. 1 January 1994. Available online: <https://www.arm.gov> (accessed on 1 March 2013).
- ASR. ASR-Focus-Groups. 1 January 2000. Available online: <http://asr.science.energy.gov/science/working-groups/focus-groups> (accessed on 1 March 2013).
- ASR. ASR-Working-Groups. 1 January 2000. Available online: <http://asr.science.energy.gov/science/working-groups> (accessed on 1 March 2013).
- Hudson, J.G.; Noble, S.; Tabor, S. Cloud supersaturations from CCN spectra Hoppel minima. *J. Geophys. Res. Atmos.* **2015**, *120*, 3436–3452.
- Mather, J.H.; McFarlane, S.A. Cloud classes and radiative heating profiles at the Manus and Nauru Atmospheric Radiation Measurement (ARM) sites. *J. Geophys. Res. Atmos.* **2009**, *114*, D19204.
- Gustafson, W.I., Jr.; Berg, L.K.; Easter, R.C.; Ghan, S.J. The Explicit-Cloud Parameterized-Pollutant hybrid approach for aerosol-cloud interactions in multiscale modeling framework models: Tracer transport results. *Environ. Res. Lett.* **2008**, *3*, 025005.
- CODATA. Available online: <http://www.codata.org> (accessed on 4 March 2013).
- CDIAC. Carbon Dioxide Information Analysis Center. U.S. Department of Energy (DOE). Available online: <http://cdiac.ornl.gov> (accessed on 1 March 2013).
- ORNL DAAC. NASA's Earth Observing System Data and Information System. Available online: <https://daac.ornl.gov> (accessed on 1 March 2013).
- Holdridge, D.; Coulter, R.; Kyrouac, J. Instrument: Balloon-Borne Sounding System (SONDE). ORNL-ARM: Oak Ridge, TN, USA, 1993. Available online: <http://www.arm.gov/instruments/sonde> (accessed on 1 March 2013).
- ARM Archive User Services. ARM LSSONDE. Available online: <http://www.arm.gov/data/vaps/lbl/lssonde> (accessed on 01 January 2013).
- CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Sci. J.* **2013**, *12*, CIDCR1–CIDCR75.
- CENDI. CENDI (the Commerce, Energy, NASA, Defense Information Managers Group). Defense Technical Information Center (DTIC): Fort Belvoir, VA, USA, 2000. Available online: http://cendi.gov/activities/12_11_2012_CENDI_NFAIS_FEDLINK_post.html (accessed on 11 December 2012).
- Rew, R.; Davis, G. NetCDF: An interface for scientific data access. *IEEE Comput. Graph. Appl.* **1990**, *10*, 76–82.

20. ARM. ARM Data Product Registration and Submission, ARM Archive User Services. 1 January 2013. Available online: <http://archive2.ornl.gov/armome/> (accessed on 1 January 2016).
21. DataCite. Available online: <https://www.datacite.org> (accessed on 1 March 2013).
22. DOE-STI Management System. United States Department of Energy. Energy Link System (E-Link). United States Department of Energy: Washington, DC, USA, 2000. Available online: <https://www.osti.gov/elink/index.jsp> (accessed on 1 March 2013).
23. File Naming Conventions. Available online: <http://www.arm.gov/data/docs/plan#naming> (accessed on 1 January 2013).
24. DOI Guidance for ARM Facility Datastreams. Available online: <http://www.arm.gov/data/docs/doi-guidance> (accessed on 1 January 2013).



© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).