NCAR/TN-492+STR
NCAR TECHNICAL NOTE

2012-11

Data Citations within NCAR/UCP

Matthew S. Mayernik
Michael D. Daniels
Bob Dattore
Ethan Davis
Kathryn M. Ginger
Karon M. Kelly
Mary R. Marlino
Don Middleton
Jennifer Phillips
Gary Strand
Steven F. Williams
Steven Worley
Michael J. Wright

NCAR Library

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000
ISSN Print Edition 2153-2397
ISSN Electronic Edition 2153-2400

NCAR TECHNICAL NOTES
http://opensky.library.ucar.edu/search/?ky=technotes

The Technical Notes series provides an outlet for a variety of NCAR Manuscripts that contribute in specialized ways to the body of scientific knowledge but that are not suitable for journal, monograph, or book publication. Reports in this series are issued by the NCAR scientific divisions. Designation symbols for the series include:

EDD – Engineering, Design, or Development Reports
Equipment descriptions, test results, instrumentation, and operating and maintenance manuals.

IA – Instructional Aids
Instruction manuals, bibliographies, film supplements, and other research or instructional aids.

PPR – Program Progress Reports
Field program reports, interim and working reports, survey reports, and plans for experiments.

PROC – Proceedings
Documentation or symposia, colloquia, conferences, workshops, and lectures. (Distribution may be limited to attendees).

STR – Scientific and Technical Reports
Data compilations, theoretical and numerical investigations, and experimental results.

The National Center for Atmospheric Research (NCAR) is operated by the nonprofit University Corporation for Atmospheric Research (UCAR) under the sponsorship of the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

NCAR/TN-492+STR
NCAR TECHNICAL NOTE

Data Citations within NCAR/UCP

Matthew S. Mayernik [1] (mayernik@ucar.edu)
Michael D. Daniels [2] (daniels@ucar.edu)
Bob Dattore [3] (dattore@ucar.edu)
Ethan Davis [4] (edavis@unidata.ucar.edu)
Kathryn M. Ginger [1] (ginger@ucar.edu)
Karon M. Kelly [1] (kkelly@ucar.edu)
Mary R. Marlino [1] (marlino@ucar.edu)
Don Middleton [3] (don@ucar.edu)
Jennifer Phillips [1] (jennp@ucar.edu)
Gary Strand [5] (strandwg@ucar.edu)
Steven F. Williams [2] (sfw@ucar.edu)
Steven Worley [3] (worley@ucar.edu)
Michael J. Wright [1] (mwright@ucar.edu)

1. NCAR Library (NCARLIB)
2. NCAR, Earth Observing Laboratory (EOL)
3. NCAR, Computational and Information Systems Laboratory (CISL)
4. UCAR Community Programs (UCP), Unidata Program Center (UNIDATA)
5. NCAR Earth System Laboratory (NESL)

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000

Table of Contents

List of acronyms

AGU - American Geophysical Union
AMS - American Meteorological Society
ARK - Archival Resource Key
BADC - British Atmospheric Data Centre
CDL - California Digital Library
CNRI - Corporation for National Research Initiatives
CODATA - International Council for Science Committee on Data for Science and Technology
DOI - Digital Object Identifier
ESIP - Federation of Earth Science Information Partners
EZID - EZID is the name of a service provided by the California Digital Library
GCMD - Global Change Master Directory
NARCCAP - North American Regional Climate Change Assessment Program
NCAR - National Center for Atmospheric Research
NCL - NCAR Command Language
NSIDC - National Snow and Ice Data Center
NSF - National Science Foundation
ORNL - Oak Ridge National Laboratory
PURL - Persistent URL
RIS - Research Information Systems
UCAR - University Corporation for Atmospheric Research
UCP - UCAR Community Programs
URL - Uniform Resource Locator
WHOI - Woods Hole Oceanographic Institute

Acknowledgements

Many people contributed to the efforts described below. The authors would like to thank Joan Burkepile, John Allison, Mary Haley, Dave Hart, Helen Moshak, Eric Nienhouse, Jonathan Ostwald, Toni Rosati, and Leonard Sitongia for their contributions. We also thank Seth McGinnis for his participation on behalf of the NARCCAP project, and for his review of this report.

Abstract

Federal agencies, professional societies, and research organizations in the geo-sciences are moving towards requiring researchers to formally cite data that led to a given research result. This trend promotes transparency in research by offering a direct pathway to the data so the research can be validated or easily carried forward from a known starting point. Such "data citations" also raise the profile of data, that is, they promote data as being as valued and rewarded in scientific settings as peer-reviewed publications. This paper is the product of an inter-divisional working group created to study, promote, and implement citations to NCAR/UCP digital resources. The paper describes how citations to NCAR/UCP digital resources could make our research products more accessible, and provides a number of recommendations for creating citations and assigning web-accessible identifiers to digital resources.

1. Introduction

Federal agencies, professional societies, and research organizations in the geo-sciences are moving towards requiring researchers to formally cite data that led to a given research result. This trend promotes transparency in research by offering a direct pathway to the data so the research can be validated or easily carried forward from a known starting point. Such "data citations" also raise the profile of data, that is, they promote data as being as valued and rewarded in scientific settings as peer-reviewed publications. Data citations will benefit the NCAR/UCP community in a number of ways, including: 1) formal citations give credit to scientists for their work in collecting and creating data, 2) formal citations will allow data center managers to track the use of data sets and gain the benefits of documenting their services and creating a foundation to design better services, and 3) formal citations will help accelerate scientific progress by tightly coupling scholary publications and data, so that two-way discovery and access are common.

In order for citations to digital resources to serve these desired roles, there must balanced support for information system development, integration of digital resource citation into scientific work practices, bibliometric measurements of citations, and institutional acceptance of citations as indicators of scientific impact. This paper describes how citations to NCAR/UCP digital resources could make our research products more accessible. This is a living document. The recommendations herein will be adjusted as our tools and practices develop, and as more NCAR/UCP groups contribute. We invite the participation of all NCAR/UCP groups to ensure that citation policies and procedures address appropriate needs and requirements for effective data center operations. Comments on this document and requests for additional information should be sent to Matt Mayernik at mayernik@ucar.edu.

2. Historical context

Digital resources, including data, software, and services, are important products of NCAR/UCAR Community Programs (UCP) research and technology development activities. NCAR/UCP digital resources are used widely within and outside the institution. In 2010, a number of NCAR data managers requested library support in developing a coherent approach to data citation across the organization. This interest in data citations is motivated by a desire to better understand the use and impact of data sets. Data citations directly link scholarship and data, and as such provide a mechanism through which data can be discovered and accessed, scholarly use of data can be tracked, and the impact of data facilities can be identified.

The first meetings of an informal NCAR/UCP data citation group took place in the summer/fall of 2010 and spring of 2011. The initial group consisted of representatives from three NCAR data management units, along with members of the NCAR Library. Initial discussions within the group focused on developing a base of knowledge and outlining the work scope for relevant data citation policies and technical implementations in preparation for broader dissemination to and involvement of other data managers.

Since the summer of 2011, the data citation group's work moved forward on a couple of fronts. In the fall of 2011, the group created a membership with EZID, a California Digital Library

service, in order to enable NCAR/UCP groups to assign Digital Object Identifiers (DOIs) to data sets that are then registered with DataCite (See discussion of the registration agency, DataCite, below). DOIs provide unique identifiers/locators for web-based objects, and are an integral component of data citations. Individual data management groups began testing EZID within their data systems and proved the EZID registration service could be successful at NCAR/UCP. The second main focus of work has been the present document, namely, an outline of recommendations for creating citations and assigning DOIs to digital resources.

3. Motivation for data citation

Data citations are growing in visibility in scientific and public policy circles. This increasing visibility is related to the calls among both the scientific and public communities for greater transparency of scientific research [6, 14, 19, 30], and the availability of new tools for identifying and linking to digital resources in a web environment [9, 10, 28, 32]. The interest in data citations is coming from many research stakeholders, including funders, policy makers, professional societies and their publication entities, research organizations, and individual researchers.

At the federal agency level, data management and citation are cross-cutting agenda points. The National Science Foundation (NSF) is increasing the pressure on grantees to make data management and data sharing a priority, as evidenced by their recent institution of a data management planning requirement with each qualified proposal [23]. In another example, the National Academies recently teamed with the International Council for Science Committee on Data for Science and Technology (CODATA) to bring together an international and interdisciplinary symposium on "Developing Data Attribution and Citation Practices and Standards" [22]. Other federal reports also discuss data citations, such as the NSF report entitled *Changing the Conduct of Science in the Information Age* [24], and the report from the NSF-sponsored workshop titled *Geo-Data Informatics: Exploring the Life Cycle, Citation and Integration of Geo-Data* [25].

Professional societies in the earth sciences are also promoting data citations. A 2009 American Geophysical Union (AGU) position statement included the following, "The scientific community should recognize the professional value of such [data] activities by endorsing the concept of publication of data, to be credited and cited like the products of any other scientific activity, and encouraging peer-review of such publications" [2]. Similarly, the 2009 prospectus for an Ad Hoc Committee on Data Stewardship within the American Meteorological Society (AMS) stated that a focal point for that committee would be to "[d]evelop a plan for citing data referenced in publications and preserving data links for the long term" [4].

Another push for data citations is coming from the Federation of Earth Science Information Partners (ESIP). ESIP is "a broad-based, distributed community of data and information technology practitioners who come together to collaborate on coordinated interoperability efforts across Earth science communities" [17]. ESIP's "Preservation and Stewardship" cluster has released an initial set of data citation guidelines for data archives [18], which were formally approved by the ESIP members in January, 2012. ESIP's guidelines draw on similar work done in the context of the International Polar Year project [26].

The same trend is being recommended for the management and citation of software and other digital resources that lead to research results. Ince, Hatton, & Graham-Cumming promote the notion that all software used to produce a research article should be made accessible to any other interested individual [20]. Similarly, Bechhofer, et al., describe how researchers use many kinds of digital resources to produce research results, including data, software, workflow tools, and computational services [8]. Focusing management and curation efforts on data along with the associated resources can provide outside researchers with the most beneficial starting point to validate results or build on the research as a whole.

4. Implementing citations to digital resources

Citations identify a particular resource and indicate where it can be acquired. Before digital resources can be cited, they must be uniquely identifiable. The most common type of unique identifier used within our current global scholarly communication systems are Digital Object Identifiers (DOIs). DOIs are designed to be immutable and overcome the inherent non-permanence of web-based URLs.

Dead URLs - URLs that either return error messages or incorrect pages - are a common problem when citing digital resources [29]. The DOI system addresses this issue by providing a unique identifier scheme and an URL resolution service that allow resource providers to persistently maintain URL-to-digital resource connections. For example, the DOI "10.3334/ORNLDAAC/810" was assigned by Oak Ridge National Lab (ORNL) to a data set titled "USGS Hydro-climate Data Network: Monthly Climate, 1951-1990." This DOI is linked through metadata, in a registration service, directly to an ORNL URL where the data can be accessed. If the need arises to change the URL (e.g. server changes, institutional re-organization) ORNL can merely change the metadata held in the registration service. This way the DOI remains fixed and valid over time, and therefore ideal for citation in publications.

DOIs are most commonly assigned to journal articles, but are seeing a growing use for data [12, 26, 27]. A number of other digital identification systems provide similar functionalities as DOIs, including Archival Resource Keys (ARKs, discussed further in section 4.4), Persistent URLs (PURLs), and Handles, but DOIs are generally recommended for use in citing scholarly materials because of their familiarity and acceptance among scientific communities and scholarly publishers [16].

To assign DOIs to NCAR/UCP resources, it is necessary to partner with a DOI registration agency. The two DOI registration agencies that are most active in working with data sets are CrossRef and DataCite (www.crossref.org/ and http://datacite.org/). CrossRef and DataCite provide similar services. They are intermediaries between digital resource publishers and the DOI system. They provide an entry point into the database of DOI-to-digital object connections, and collect metadata for resources that are assigned DOIs. The salient difference between them is that CrossRef's DOI registration system was created and is customized for journal articles and books. DataCite, on the other hand, was founded to promote the assignment of DOIs to data, and is designing DOI services specifically for data sets, although they can be used for non-data resources as well.

Other research organizations are also instituting data citations. We compared the practices of four organizations: 1) Oak Ridge National Lab (ORNL), 2) National Snow and Ice Data Center (NSIDC), 3) Woods Hole Oceanographic Institute (WHOI), and 4) British Atmospheric Data Centre (BADC). For ORNL, NSIDC, and WHOI, we used a structured set of questions to interview one individual whom we knew to be actively working on data citations. With BADC, our information gathering was more informal, but consisted of a number of discussions with a member of the BADC data citation team at professional meetings. Key takeaways from our discussions are:

- ORNL was assigning DOIs to data sets via CrossRef until early 2012. They have since started assigning DOIs via DataCite.
- WHOI is assigning DOIs to data sets via CrossRef. Their partnership with CrossRef was set up before DataCite had formed.
- BADC is assigning DOIs to data sets via DataCite.
- NSIDC is a member of EZID, and will subsequently have the DOIs registered with DataCite.

Based on this assessment and testing, noted above, DataCite was selected as the registration agency and EZID the registration service most appropriate for NCAR/UCP. The EZID service allows users to assign DOIs or ARKs to digital resources through a GUI or an API, which is very effective for data collection with numerous entities. EZID also has a scalable price model, allowing users to create an unlimited number of identifiers for a flat yearly membership fee. NCAR's membership with EZID began Sept. 1, 2011.

The chain of persistence for DOIs involves a number of organizations. In addition to DataCite and EZID, the International DOI Foundation and Corporation for National Research Initiatives (CNRI) provide technical and organizational services that ensure that the DOI system is a globally stable way to link to digital resources. In summary:

- NCAR/UCP ensures persistence of digital resources and their resolving URLs (discussed below)
- EZID ensures persistence of the DOI registration service that connects to DataCite agency and services
- DataCite ensures persistence and wide accessibility of DOI registration service and associated metadata store for many institutions
- The International DOI Foundation ensures persistence of DOI resolution services
- CNRI ensures persistence of Handle technology that underlies the DOI system.

5. Citation and DOI policies and procedures

Once a relationship with a DOI registration agency is established, EZID in our case, NCAR/UCP can begin assigning DOIs. Prior to registering DOIs, however, it is necessary to create a coherent organizational framework through which DOI assignments can be coordinated across individual NCAR/UCP units. This framework will allow individual units to create DOI policies and procedures that work within their teams' workflows and meet the overarching expectation for the

organization as a whole. In P1 – P13 below, we outline the components of a recommended citation policy and procedure framework.

5.1 The Who, What, When, and How of assigning DOIs

Citations are embedded within larger archival institutions, as illustrated by the AGU policy for referencing data: "data cited in AGU publications must be permanently archived in a data center or centers that meet the following conditions: a) are open to scientists throughout the world, b) are committed to archiving data sets indefinitely, c) provide services at reasonable costs. … Data sets that are available only from the author, through miscellaneous public network services, or academic, government or commercial institutions not chartered specifically for archiving data, may not be cited in AGU publications" [1]. AMS has similar concerns for the longevity of resources that are cited in article reference lists (personal communication). Assigning DOIs must be supported by archival practices that insure complete integrity of digital resource collections, including files, metadata, documentation, and software, such as the practices recommended in the Open Archival Information System reference model [11]. These practices guarantee that the resources will be understandable and usable by future scientific generations.

P1. Guidance policy for NCAR regarding who is able to assign DOIs

DOIs and data citations are intended to indicate that data are maintained in stable data archives [21]. Thus, the working policy is that only NCAR/UCP groups that have plans for ensuring the availability of resources over time will be able to register and assign DOIs.

P2. Guidance policy regarding what resources should be assigned DOIs

DOIs can be assigned to a range of resources. The work that led to this white paper began by focusing on assigning DOIs to data sets, but most of the procedures and recommendations are applicable to other kinds of resources. DataCite's metadata schema (discussed more in section 4.4) states that DOIs can be assigned to twelve different resource types [15, pg. 14]. The resource types most relevant to NCAR/UCP usage are data sets, software, services, and texts. Groups will be free to assign DOIs to resources for which NCAR/UCP makes accessible to research communities and/or the general public. A few notes specific to particular resource types are:

- *Data sets:* NCAR/UCP data archives will be free to assign DOIs to data held within their archives. DOIs can be assigned to data sets that are available online, as well as to data that are offline and available by request. If data archives would like to assign DOIs to data sets that are not publicly accessible, or before they are publicly accessible, the DOI URL should resolve to a webpage that informs the public about when the data will be available, or the process through which the data can be accessed.
- *Software*: Software is central to many NCAR/UCP research efforts. Open and accessible software is important to ensure transparent science and data [20]. Software that is likely to be cited in a scientific article might benefit from being assigned a DOI. As with data sets, however, software packages should only be assigned DOIs if they are managed and maintained for use over time. If a DOI is

assigned to software, a corresponding citation should be provided to the software users.

- *Services*: Many NCAR/UCP groups provide data services, such as data sub-setting, regridding, or visualization services. Data services could be considered to be citable resources if they affect the data sets that they use. For example, a data service might create a new data set or modify an existing data set. Providing a citation and a DOI for a data service might be useful to allow these services to be acknowledged in a formal way.
- *Texts*: NCAR/UCP-produced textual documents might benefit from having DOIs. Many scientific articles are already assigned DOIs by academic publishers, but NCAR/UCP groups produce other types of documents that are managed and published within the organization. The key consideration is whether a document might be cited in a scholarly article. DOIs might be appropriate for technical reports, software or equipment manuals, or workshop reports.

P3. Guidance policy for the timeline of DOI assignment

When assigning DOIs to resources being made available for the first time, DOIs should be assigned when the resources are posted to a public web site. When back-populating existing publicly available resources with DOIs, the timeline of DOI assignment will vary with the resource collections. A collection can be assigned DOIs in a single batch, or collections can be assigned DOIs in phases as discrete batches, with the prioritization based on 1) simplicity (e.g. choosing well-bounded resources first as pilot DOI assignments), or 2) known resource use patterns (e.g. assigning DOIs to more highly used resources first). If possible and necessary, DOIs can be assigned before resources are made publicly available in order to provide the principal investigators with citable DOIs that can be used when publishing the initial results from a project. If DOIs are assigned to resources that are not yet publicly available, archives can direct the DOIs to a web page that provides a brief overview of the resource, with a statement such as "Data are currently only available to principal investigators, but will be publicly available in the future," that indicates when the resource(s) will be posted. When creating DOIs via the EZID API, DOIs can be declared to be "reserved." An identifier created with the "reserved" status is fully functional within EZID, but knowledge of its existence is withheld from external services (e.g., from the DOI system and indexing systems) until the identifier is made public. A "reserved" identifier could therefore be used internally until a resource is made public, at which time the DOI status can be changed to "public" via the EZID API. "Reserved" DOIs can also be deleted at a later date if their associated resource is never officially made public.

5.2 Challenges of assigning DOIs

Assigning DOIs to data sets and other digital resources is not as straightforward as assigning DOIs to a journal article. Journal articles are fixed and singular objects, that is, they do not change and they have one definitive published form. Data sets, on the other hand, often have indistinct identities [33]. Data sets might consist of many individual units (such as files or database tables) or might themselves be subsets of larger data collections. In addition, many data sets change on a daily or weekly basis, as, for example, when new measurements are

continuously added to an existing data set. The following set of policies addresses some of the salient challenges with assigning DOIs to digital resources.

P4. Guidance policy for addressing the granularity of DOI assignment

The resource managers within each NCAR/UCP archiving team should make the decision about the granularity at which resources should be assigned DOIs. This decision should be based on their experience with the ways that resources from their archives are typically used. The decisions regarding the DOI-to- resource granularity should be documented and remain consistent. Some questions that can provide guidance on making citation granularity decisions include:

- What resources do users have access to?
- What does the user community consider a complete and sufficient citation in their field?
- Will the granularity support a reasonably accurate starting point for subsequent research?
- What do you want users to cite?
- What do you want to get citation metrics on?
- How do you manage resources internally?
- How do you display resources to users?

P5. Guidance policy for assigning DOIs to versions or changing resources

DOIs may be assigned to resources that are released in different versions or that change over time. The ESIP data citation guidelines suggest a few approaches to this issue [18]. One approach is to use “major” and “minor” versions. In this approach, DOIs are assigned to "major versions" of resources, where a major version is considered to be a significant change to the full resources, such as a reprocessing or recalibration of all points in a data set. "Minor versions" are not assigned DOIs. “Minor” versions may indicate fixes to individual points or the addition of new values to an already existing data set. Resource managers should determine what “major” and “minor” means in their systems. The important task is to document and track how DOIs are assigned, and to insure as well as possible that follow-on research can accurately begin from a cited starting reference.

P6. Guidance policy for assigning DOIs in collaborative and inter-institutional projects

NCAR/UCP units should assign DOIs to resources for which they have long-term curatorial responsibility. For resources that were acquired from another research organization, units should check to see if a DOI is already assigned to those data, and contact the original data providers if questions exist regarding the propriety of assigning DOIs to resources brought in from the outside. Depending on whether or not the acquired resources have DOIs, NCAR/UCP units have a few options:

- *If the acquired resources do not have a DOI,* NCAR/UCP units can assign a DOI. In this case, the NCAR/UCP unit should inform the original resource owners that a DOI has been assigned.
- *If the acquired resources already have a DOI assigned,* NCAR/UCP units can 1) use the original DOI (for example, if the original DOI is maintained by a trusted repository, is an exact copy, and users are assured of receiving an exact copy

transparently through the services at NCAR/UCP), 2) assign a new DOI if any changes/additions are made to the resources at NCAR/UCP that merit the declaration of a new version (for example, sub-setting or compiling data, adding metadata, or providing additional services), or 3) ask that the ownership/responsibility for the original DOI be transferred to NCAR along with the resources (for example, if the original DOI owner is no longer maintaining DOIs).

If a new DOI is assigned by NCAR/UCP to resources that have existing DOIs from another organization, this relationship should be indicated in metadata for the new DOI. The DataCite metadata schema has a "RelatedIdentifier" field, which should be used to indicate that two DOIs are related. See the DataCite metadata documentation for further details [15].

Assigning a DOI is a statement about the long-term availability of a digital resource. As such, NCAR/UCP has an institutional obligation to maintain resources that have been assigned DOIs, as well as an obligation to ensure that the resolving URL for a DOI is always kept up to date. In addition, situations do occur in which resources are removed from archives. NCAR/UCP has an obligation to ensure that DOIs assigned to deprecated resources still resolve to a web page that informs users about the resources' status.

P7. Guidance policy for maintaining DOIs registered by NCAR

The EZID system supports identifier persistence by doing identifier checks to find outdated resolution URLs. We will make use of this independent service to conduct annual checks of the institution's DOIs (and ARKs if applicable) to ensure that they resolve to the designated data resource.

P8. Guidance policy for maintaining DOIs that were assigned to deprecated resources

If DOIs have been assigned to resources that are deleted or removed from public accessibility, the DOIs should resolve to a page that indicates that those resources are no longer available. If resources have been superseded by another version, the DOI resolution page should indicate where the new version can be found. If resources have been removed from a public archive and are not replaced, their DOIs should resolve to a page that describes when and why the removal took place. When permanently deleting resources, NCAR/UCP units should provide a notice period in which an announcement is made that the resources will be deleted at a certain point in the future. At minimum, this announcement should go to the users of those resources. The announcement could also be sent, for example, to other data archives that may be interested in taking responsibility for data sets. In some cases, it might be necessary to send the announcement to the relevant NCAR/UCP administration committees. The intent with this announcement is to provide a broad opportunity for anybody to access resources one last time, and/or make the case for why the resources should be preserved.

5.3 Recommended DOI and citation syntaxes

DOIs consist of a number of concatenated alpha-numeric segments. The first segment is the DOI "prefix" and is always in the form of "10.XXXX," where the XXXX indicates the agency (such as CrossRef or DataCite) through which the DOI is being registered. EZID has assigned NCAR

to use the prefix "10.5065." This prefix is not unique to NCAR/UCP. The second (optional) segment of a DOI is two characters long, and is called a "shoulder." EZID has assigned NCAR the shoulder "D6," which is unique to NCAR. All other subsequent characters in a DOI are collectively called the DOI "suffix," and can be designated at the discretion of the data provider. Thus, DOIs assigned by NCAR/UCP will have the form "10.5065/D6_____".

P9. Guidance policy for the syntax of NCAR DOI suffixes

No best practice guides exist yet for how DOIs themselves should be constructed. Our recommendation is that DOIs should be generated as random strings of characters. Within EZID, this is called "minting" DOIs. Minting a DOI via EZID returns a randomly generated DOI suffix. Randomly generated DOI suffixes are desirable for two reasons. First, NCAR/UCP's structure often changes and digital resources may be moved around between NCAR/UCP units. Thus, organizational designations in DOI suffixes may quickly become obsolete. For example, data from the Earth Observing Laboratory could be transferred into the Computational & Information Systems Laboratory Research Data Archive. Second, any human or machine-readable intelligence built into DOI syntax is another thing that needs to be managed and maintained over the long-term. Thus, random identifiers are easier to maintain and manage than non-random identifiers.

DOIs are one element of a citation. In addition to a DOI, citations to digital resources should include information - title, author, publisher, etc. - that allow a human to identify the resource and understand where it can be found. A number of recommendations exist regarding the elements that should be included in a data citation [7, 12, 18, 21]. Our recommended citation syntaxes are flexible because different NCAR/UCP resources will have different citation requirements (e.g. model vs. observational data vs. software). In addition, different journals have different required citation styles, even within a single publisher such as Springer or Elsevier. See [3] and [5] for citation formats recommended by professional society publishers.

P10. Guidance policy for recommended citation syntaxes

Recommended citation syntaxes should be provided on the resolving page for a DOI. The ESIP recommendations provide a reasonable starting point to shape the NCAR/UCP requirements [18]. Suggested required and optional elements are:

- *Required:* Author. Release date. Title. Archive/Provider. Locator/Identifier. Resource access date (date that the user accesses the resource).
- *Optional:* Version; Subset Used; Editor, Compiler, or other important role; Distributor, Associate Archive, or other Institutional Role.

P11. Guidance policy for authorship designation for resources generated by large distributed projects

For digital resources that are generated by large-scale distributed projects, such as the output from climate models or field projects, the authorship designation might be unclear. In those cases, two options exist. First, the project as a whole can be designated as the author, and any listing of individual contributions (if desired) can be maintained and made available on a project web site. Second, the project can compile an authorship list alphabetically, or in another fashion as desired, and ask that the data citation list all authors.

5.4 Working with EZID

With many NCAR/UCP units working to implement citations in different systems and for different kinds of resources, we have an opportunity to share knowledge and technical tools amongst each other. In particular, tools for interacting with EZID's API should be directly applicable from one NCAR/UCP unit to another. The EZID API documentation is found at: http://n2t.net/ezid/doc/apidoc.html. Some initial guidelines for working with EZID include:

EZID Testing

EZID provides a test account (contact Matt Mayernik for the test account name and password) and a test DOI prefix, "doi:10.5072/FK2". Test DOIs are fully functional in that they are recognized by the http://dx.doi.org resolver, but test identifiers are deleted by EZID after 14 days. When using the test account to test the EZID API, you need to specify the "doi:10.5072/FK2" prefix to mint or create test DOIs.

EZID Metadata

DOIs should be created with associated metadata that describe associated resource. EZID does not require metadata to register a DOI, but supports a number of metadata schemes. If possible, DOIs should be associated with metadata that complies with the DataCite metadata schema. This will enable NCAR/UCP resources to be integrated into the growing DataCite metadata store, which will increase the likelihood that they are discovered, accessed, and used. Version 2.2 of the DataCite metadata schema has five required elements and twelve optional elements [15, 31]. Version 2.2 of the DataCite metadata schema is provided in the Appendix.

Using Archival Resource Keys (ARKs)

EZID allows users to assign DOIs and ARKs. ARKs are assigned via EZID using the same methods available for assigning DOIs. The NCAR/UCP ARK prefix (assigned by EZID) is "ark:/85065/d7". Reasons why data archives might use ARKs include:

- ARKs can be deleted, DOIs cannot. Thus, if archives desire to use an identifier for temporary purposes, ARKs are more appropriate.
- ARKs and DOIs can work together. For example, archives may choose to use ARKs for materials such as documentation, programs/scripts, or web pages that are associated with data that have a DOI.
- ARKs support "suffix passthrough," allowing one ARK registration to identify many thousands of extended ARKS in order to represent the many parts of a dataset. As of this writing, DOIs do not support suffix passthrough.

5.5 Developing vocabularies

To ensure consistency across NCAR/UCP units, we should develop vocabularies that specify how NCAR/UCP units and people should be named and described. In addition, we should specify vocabularies for other metadata elements that we think necessary. These vocabularies should work with the metadata vocabularies that groups currently use, such as the Global Change Master Directory (GCMD).

6. Outreach to NCAR/UCAR community

Our citation work will have minimal impact on the NCAR/UCP community if it does not include efforts to promote such citations within our scientific communities. Anecdotal evidence shows that while scientists do formally cite data in some cases, this is not yet a regular practice in the earth and space sciences [13, 26]. Our outreach efforts will focus on raising the profile of citations to digital resources by informing scientists of our DOI work and being proactive in providing scientists with recommended citations. We will initially use the existing relationships that data archives have with scientific groups to raise awareness our efforts, and as our work proceeds perform formal presentations to labs who have not yet been involved with the initial data citation working group. The NCAR Library will coordinate these outreach efforts.

P12. Suggestion for promoting data citations within scientific groups

When an NCAR/UCP unit publishes a digital resource, i.e. makes the resource publicly accessible online, the archive should send an email to the resource's authors with a congratulatory message, as illustrated by the following fictional example, , "Congratulations, your data set 'Weather Data' has been published in the NCAR Data Archive. It has been assigned the DOI 10.9999/12345, and can thus be accessed at the persistent URL http://dx.doi.org/10.9999/12345. The internal NCAR/UCP URL for this data is http://www.ucar.edu/WeatherData. When publishing papers that make use of this data, please formally cite the data using the following recommended citation:"

P13. Suggestion for making citation information easily findable and usable

To help users to find and use citations and DOIs, citation information should be displayed on the web page of a resource. If possible, the citation information should be displayed on the target URL of the resource's DOI. Another recommendation is to enable users to download citation information in RIS and/or Bibtex formats, so that users can easily import data citations into their citation management software like EndNote, Zotero, RefWorks, or document creation software like LaTeX. DataCite also provides basic citation information in RIS, Bibtex, and other formats on the DataCite page for each DOI.

7. Understanding the impact of data and data citations

To evaluate the impact of enabling and promoting citations to data and other digital resources, we need to be able to count citations accredited to NCAR/UCP data over time. Assigning DOIs to digital resources will simplify this task, as DOIs provide a unique character string that can be searched for in databases and on the internet. A prerequisite to any citation counting is keeping an up-to-date list of DOIs registered by NCAR.

P14. Guidance policy for tracking DOIs registered by NCAR

The EZID service will allow us to compile lists of DOIs registered under the NCAR account. Thus, we do not need to develop a process to do this. We can also set up multiple usernames under our NCAR EZID account, which will allow us to compile separate lists for DOIs registered by different NCAR groups. With these lists, an accounting of each archives' DOI assignments can be made.

Counting citations is an inherently difficult process. Different citation indexes for journal articles, such as the Web of Science, Scopus, and Google Scholar, will give different citation counts for the same article. Counting citations to digital resources currently is even more difficult because there is no citation index for data, software, or other digital resources. DataCite is working with Thompson-Reuters to get DataCite DOIs indexed in the Web of Science, but this service, if developed, is likely still a few years off. Citations can be compiled manually by searching through databases and internet search engines for DOIs or other data set identification information, such as titles. This process is very time consuming, but is the default citation chasing method for the time being.

Thus, developing methods for counting and tracking citations to NCAR/UCP digital resources is an open research area. We might investigate a combined human/automated approach by developing scripts that reduce the manual effort required to find DOIs online. We should also investigate methods to parse articles in OpenSky for such citations, leveraging other similar work, such as that of Sanderson, Phillips, and Van de Sompel [29].

As our methods for assessing the impact of digital resource citations develop, we can use these impact assessments to promote increased rewards for scientists who produce data, software, and other digital resources. Digital resource archives will also benefit by understanding how the use of their resources, potentially allowing them to consider these usage workflows when designing improvements for user services.

8. NCAR/UCP progress

As mentioned above, NCAR/UCP has a membership with EZID to use their DOI registration services. We are sharing knowledge and technical tools related to EZID primarily through the UCAR wiki, and will develop other knowledge and tool sharing methods, such as public web pages, if appropriate. An additional benefit of working with EZID is that the EZID managers are part of DataCite leadership committees, including the DataCite metadata committee, which will make it possible for us to directly influence the DataCite services and community as they develop.

We are organizing pilot projects within NCAR/UCP data archives to begin assigning DOIs. As of September 16, 2012, we have assigned DOIs to two NCAR/UCP-hosted resources: the NCAR Command Language (NCL) software package, and the North American Regional Climate Change Assessment Program (NARCCAP) data set. These two resources provided well-bounded cases with which to test the EZID API and the DataCite metadata schema. We have also assigned an ARK to the Yellowstone supercomputing facility, which is managed by NCAR/UCP. Yellowstone was assigned an ARK instead of a DOI because it is a deliberate experiment to promote citation and attribution to the facility, and as such we wished to avoid the implications of permanence and persistence that come with DOIs.

The following bullets illustrate the identifiers for those three resources, and the associated metadata for each. Note that the same metadata elements are displayed on both the EZID and DataCite metadata pages for the two DOIs. This is because the metadata for these two DOIs were submitted in the DataCite metadata schema.

- NCL DOI - http://dx.doi.org/10.5065/D6WD3XH5
  - DOI target URL - http://www.ncl.ucar.edu/
  - EZID metadata page - http://n2t.net/ezid/id/doi:10.5065/D6WD3XH5
  - DataCite metadata page - http://data.datacite.org/10.5065/D6WD3XH5
- NARCCAP DOI - http://dx.doi.org/10.5065/D6RN35ST
  - DOI target URL - http://www.earthsystemgrid.org/project/NARCCAP.html
  - EZID metadata page - http://n2t.net/ezid/id/doi:10.5065/D6RN35ST
  - DataCite metadata page - http://data.datacite.org/10.5065/D6RN35ST
- Yellowstone supercomputing facility ARK - http://n2t.net/ark:/85065/d7wd3xhc
  - ARK target URL - http://www2.cisl.ucar.edu/resources/yellowstone
  - EZID metadata page - http://n2t.net/ezid/id/ark:/85065/d7wd3xhc

Other pilot projects will be developed within different teams around NCAR/UCP as our work progresses.

9. Cost analysis

Assigning DOIs to digital resources and maintaining those resources (and DOIs) over time do come with some costs. The pilot projects will be used to assess cost details, but some initial considerations are as follows. In initiating DOI and citation services, human time and effort will be necessary to incorporate the DOI registration services into existing digital archiving systems. Back-populating DOIs for existing resources will also require up-front effort. As mentioned above, we hope to minimize these up-front costs by sharing tools and knowledge across projects, systems, and NCAR/UCP units. The central cost considerations over time, however, are the costs required to meet the goal of having sustainable archives. Guidance policy P8 above describes procedures for if/when resources are to be removed from an archive. The expectation when assigning DOIs, however, is that the resources and associated DOIs will be maintained over time, and that removing resources from archives will be a rare occurrence.

10.Next Steps

This white paper is open for comments from any NCAR/UCP project or unit. After we have assigned DOIs within different NCAR/UCP units, and to different kinds of resources, we will evaluate our progress and identify any challenges to be addressed. At that point, we will report to the NCAR Directors and UCAR Exec. Committees with recommended policy steps that are necessary in order to make digital resource citation and DOI assignment institutionally supported practices.

Our citation work is opening many questions that require input beyond our group. For example, what is the process by which a data archive can show sustainability? Should only archives that have base funding be allowed to assign DOIs? Similarly, NCAR/UCAR do not have a Human Resources job classification, such as “data manager,” that recognize data work. If data management work (including the responsibility to assign DOIs) is invisible at an administrative level, assigning responsibility for the sustainability of data archives (and all associated services) is problematic. These issues require discussions of policy that are beyond the capacity of our

working group to resolve. UCAR's policy on Publication & Information Dissemination (http://www.fin.ucar.edu/polpro/section3/3-5.html) "supports an open exchange of data and scholarly information derived from our research". Neither the policy nor associated procedures, however, are explicit about the roles, responsibilities, or mechanisms to ensure the effective management, preservation, and access of digital resources. Such a policy would be required if these questions are to be answered at an institutional level. This white paper provides some initial directions on what such a policy might include, but the precise parameters require considerable thought and debate. We encourage and welcome discussion on these topics.

11.References

1. American Geophysical Union (AGU). (1996). *Policy on Referencing Data in and Archiving Data for AGU Publications.* http://www.agu.org/pubs/authors/policies/data_policy.shtml

2. American Geophysical Union (AGU). (2009a). *AGU Position Statement: The Importance of Long-term Preservation and Accessibility of Geophysical Data*. http://www.agu.org/sci_pol/positions/geodata.shtml

3. American Geophysical Union (AGU). (2009b). *AGU Reference Style*. http://www.agu.org/pubs/authors/manuscript_tools/journals/pdf/AGU_reference_style.pdf

4. American Meteorological Society (AMS). (2009). *AMS Ad Hoc Committee on Data Stewardship Prospectus.* http://www.unidata.ucar.edu/staff/mohan/Data%20Stewardship%20Prospectus.pdf

5. American Meteorological Society (AMS). (2012). *Author Reference/Citation Guide.* http://www.ametsoc.org/pubs/journals/author_reference_guide.pdf

6. Arzberger, P, P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir, & P. Wouters. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal* 3: 135-152. http://www.jstage.jst.go.jp/article/dsj/3/0/135/_pdf

7. Ball, A. & Duke, M. (2011). How to Cite Datasets and Link to Publications. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides

8. Bechhofer, S., et al. (2012). Why linked data is not enough for scientists. *Future Generation Computer Systems*, http://dx.doi.org/10.1016/j.future.2011.08.004

9. Bizer, C. (2009). The Emerging Web of Linked Data. *IEEE Intelligent Systems*, 24(5): 87-92. http://dx.doi.org/10.1109/MIS.2009.102

10. Brase, J. (2004). Using Digital Library Techniques – Registration of Scientific Primary Data. *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science* (Vol. 3232, pp. 488-494). Springer Berlin / Heidelberg. http://dx.doi.org/10.1007/978-3-540-30230-8_44

11. Consultative Committee for Space Data Systems (CCSDS). (2012). *Reference Model for an Open Archival Information System (OAIS).* Recommendation for spac*e data system standards, CCSDS 650.0-M-2.* http://public.ccsds.org/publications/archive/650x0m2.pdf

12. Cook, R. (2008). Editorial: Citations to Published Data Sets. *FluxLetter: the Newsletter of FluxNet*, 1(4): 4-5. http://bwc.berkeley.edu/FluxLetter/FluxLetter-Vol1-No4.pdf

13. Cook, R. (2011). Archiving Earth Science Data: Experiences of the ORNL Distributed Active Archive Center. Presentation at *DataCite 2011 Summer Meeting*, Berkeley, CA, August 25, 2011. http://datacite.org/slides/DataCite2011/DataCite0502-Cook.ppt

14. Costello, M.J. (2009). Motivating Online Publication of Data. *BioScience*, 59(5), 418-427. http://www.jstor.org/stable/10.1525/bio.2009.59.5.9

15. DataCite. (2011). *DataCite Metadata Schema for the Publication and Citation of Research Data*. Version 2.2. http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

16. Duerr, R., Downs, R., et al. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, 1-22. http://dx.doi.org/10.1007/s12145-011-0083-6

17. Federation of Earth Science Information Partners (ESIP). (2012a). Welcome to ESIP Federation. http://www.esipfed.org/

18. Federation of Earth Science Information Partners (ESIP). (2012b). Interagency Data Stewardship/Citations/provider guidelines. http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines

19. Heffernan, O. (2010). Saluting scrutiny. *Nature Reports Climate Change*, 2 March 2010. http://dx.doi.org/10.1038/climate.2010.20

20. Ince, D.C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, 482(7386): 485–488. http://dx.doi.org/10.1038/nature10836

21. Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: moving towards formal data publication. *International Journal of Digital Curation*, 6(2): 4-37. http://www.ijdc.net/index.php/ijdc/article/view/181

22. National Academy of Sciences (NAS). (2011). *Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop.* http://sites.nationalacademies.org/PGA/brdi/PGA_064019

23. National Science Foundation (NSF). (2010). *Scientists seeking NSF funding will soon be required to submit data management plans*. http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

24. National Science Foundation (NSF). (2011). *Changing the Conduct of Science in the Information Age*. http://www.nsf.gov/pubs/2011/oise11003/index.jsp?WT.mc_id=USNSF_25&WT.mc_ev=click

25. *NSF Geo-Data Informatics: Exploring the Life Cycle, Citation and Integration of Geo-Data: Workshop report.* (2011). http://tw.rpi.edu/media/latest/WorkshopReport_GeoData2011.pdf

26. Parsons, M.A., Duerr, R., & Minster, J.-B. (2010). Data Citation and Peer Review. *Eos Transactions, AGU*, 91(34). http://dx.doi.org/10.1029/2010EO340001

27. Paskin, N. (2005). Digital Object Identifiers for scientific data. *Data Science Journal*, 4(0), 12-20. http://www.doi.org/topics/050210CODATAarticleDSJ.pdf

28. Pepe, A., Mayernik, M.S., Borgman, C.L., & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the American Society for Information Science and Technology*, 61(3): 567-582. http://dx.doi.org/10.1002/asi.21263

29. Sanderson, R., Phillips, M., & Van de Sompel, H. (2011). Analyzing the Persistence of Referenced Web Resources with Memento. *Open Repositories 2011 Conference*. http://arxiv.org/abs/1105.3459

30. Science Staff. (2011). Challenges and Opportunities. *Science*, 331(6018): 692-693. http://dx.doi.org/10.1126/science.331.6018.692

31. Starr, J. & Gastl, A. (2011). isCitedBy: A Metadata Scheme for DataCite. *D-Lib Magazine*, 17(1/2). http://www.dlib.org/dlib/january11/starr/01starr.html

32. Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking scholarly communication: Building the system that scholars deserve. *D-Lib Magazine*, 10(9). http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

33. Wynholds, L. (2011). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. *International Journal of Digital Curation*, 6(1). http://www.ijdc.net/index.php/ijdc/article/viewFile/174/242

12. Appendix – DataCite Metadata Schema, Version 2.2

The properties listed in Table 1 *must be* supplied when submitting DataCite metadata. The optional properties listed in Table 2 *may be* supplied when submitting DataCite metadata.

Tables taken from:

http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

Table 1: DataCite Mandatory Properties

ID	*Property*
1	Identifier (with type attribute)
2	Creator (with name identifier attributes)
3	Title (with optional type attribute)
4	Publisher
5	PublicationYear

Table 2: DataCite Optional Properties

ID	*Property*
6	Subject (with schema attribute)
7	Contributor (with type and name identifier attributes)
8	Date (with type attribute)
9	Language
10	ResourceType (with description attribute)
11	AlternateIdentifier (with type attribute)
12	RelatedIdentifier (with type and relation type attributes)
13	Size
14	Format
15	Version
16	Rights
17	Description (with type attribute)