Bridging Data Lifecycles: Tracking Data Use via Data Citations Workshop Report

January, 2013

Matthew S. Mayernik

NCAR Library

NCAR Technical Notes

National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado
80307-3000
www.ucar.edu

NCAR/TN-494 +PROC

NCAR TECHNICAL NOTES

http://www.ucar.edu/library/collections/technotes/technotes.jsp

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 yet at a point of a formal journal, monograph or book publication. Reports in this series are issued by the NCAR scientific divisions, published by the NCAR Library. 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 maybe 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.

National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado 80307-3000
ISSN Print Edition 2153-2397
ISSN Electronic Edition 2153-2400

Bridging Data Lifecycles: Tracking Data Use via Data Citations Workshop Report

Matthew S. Mayernik
(mayernik@ucar.edu)

NCAR Library (NCARLIB)

NCAR/TN-494+PROC
NCAR Technical Note
Published By: NCAR Library
January, 2013

Contents

Acknowledgements

Funding for this workshop was provided by the National Oceanic and Atmospheric Administration (NOAA) through the UCAR Joint Office for Science Support (JOSS). Thanks to Karon Kelly and Mary Marlino for their expertise in initiating and organizing the workshop, Pat Steinkamp for workshop logistical support, and Jose Castilleja for technical support. Jennifer Phillips and Matt Ramey took great notes during the workshop sessions. Thanks to Steve Williams for reviewing this report. This workshop grew out of the efforts of the NCAR/UCAR data citation working group. The NCAR technical note, "Data Citations within NCAR/UCP," which is a product of that working group, provided direction for this report (Mayernik, et al., 2012). And most importantly, thanks to the workshop speakers and participants for their ideas, interest, and active engagement.

Portions of sections 2, 4.2, and 5.6 were first published in:
Mayernik, M.S. (2012). Data Citations: Initiatives, Issues, and First Steps. *Bulletin of American Society for Information Science and Technology*, 8(5): 23-28.
http://www.asis.org/Bulletin/Jun-12/JunJul12_MayernikDataCitation.pdf

Abstract

Digital technologies for identifying and linking to resources on the internet promise to make connections between scholarly publications and their underlying data more transparent and traceable. "Data citations" are formal citations included in reference lists of published articles to data resources that led to a given research result. The workshop "Bridging Data Lifecycles: Tracking Data Use via Data Citations," held by the University Corporation for Atmospheric Research (UCAR) in April, 2012, brought together 80 people from 30+ organizations to discuss many aspects of data citations. This report outlines the important activities, tools, challenges, and impediments to data citation initiatives that were identified during the workshop. The report also outlines a set of recommendations on how to get started on the processes of assigning citations and actionable identifiers to data sets without having solved every issue. By making it easy for users, providing openness and transparency in how data citation tools are being applied, and leveraging the interest and expertise of the multiple communities of stakeholders, organizations can promote, enable, and embed data citations as regular components of the scholarly communication infrastructure.

1. Executive Summary

As digital research data collections grow in size and usage, the desire to understand the ways that data are used to produce new scholarly products also grows. In parallel, digital technologies for identifying and linking to resources on the internet promise to make connections between scholarly publications and their underlying data more transparent and traceable. “Data citations” are formal citations included in reference lists of published articles to data resources that led to a given research result. The workshop “Bridging Data Lifecycles: Tracking Data Use via Data Citations,” held by the University Corporation for Atmospheric Research (UCAR) in April, 2012, brought together 80 people from 30+ organizations to discuss many aspects of data citations. The goal of the Bridging Data Lifecycles workshop was to develop a cohesive view of the important challenges and impediments to data citation initiatives, as well as of the tools, techniques, and practices available to make such initiatives successful.

Themes

The workshop identified a number of common themes, ranging from conceptual debates about data publication to the practical challenges of tracking data use. Data citation initiatives are often tied to the idea that data sets should be published just like other kinds of scholarly products. The idea of publishing data sets, however, becomes problematic when looking at the similarities and differences between traditional scholarly publications and digital data sets. The multiple perspectives present at the Bridging Data Lifecycles workshop proved valuable in highlighting the ways that data citations can draw on the expertise of different stakeholders. Funding agencies, research organizations, libraries, research communities, and data archives/repositories contribute to data citation initiatives in different ways. A number of organizations are developing data citation tools and recommendations for data repositories and data user. For example, numerous approaches have been developed for identifying and locating digital objects, with most of them making use of actionable identifiers like Uniform Resource Identifiers/Uniform Resource Locators (URI/URLs), Handles, Digital Object Identifiers (DOIs), and Archival Resource Keys (ARKs). Tools and methods for tracking data citations, however, are very time and effort intensive, or are in the very early stages of development.

Challenges

Several challenges specific to data citation practices and policies were identified and discussed during the Bridging Data Lifecycles workshop. First, data citations are currently uncommon within many research communities. Most data users note their use of particular data sets in either the research methods or acknowledgements sections of their papers, not as formal citations in a paper's bibliography. Second, it can be very difficult to formalize the identity of a citable object within data management and archiving systems. Decisions about granularity and version tracking have to be made before assigning actionable identifiers to data sets. Third, data citation initiatives raise many implications for long-term archiving and maintenance of digital data sets. Once a data set has been cited, it becomes part of the scholarly record, with all of the associated implications of permanence.

Recommendations

Organizations that are in the beginning stages of developing data citation initiatives face a number of considerations. The set of recommendations developed via the workshop identify ways to get started on the processes of assigning citations and actionable identifiers to data sets without having solved every detailed problem.

Recommendation 1: Identify what you want to achieve via data citations

Organizations need to decide which data citation motivations, purposes, and considerations are most important in their local contexts, and should prioritize data citation-enabling efforts accordingly.

Recommendation 2: Understand the options for actionable identifier schemes

Many actionable identifier schemes exist. Understanding the similarities and differences between the schemes is an important component of data citation efforts.

Recommendation 3: Engage stakeholders

Configurations of people and organizational units will be different in every data citation initiative. Within any particular organization, different capabilities and relationships exist in libraries, computing units, data management and archiving teams, administrative offices, and research teams. Stakeholder engagement is particularly important in building awareness and knowledge of data citation issues among the researchers who generate the data in the first place, and among data user communities.

Recommendation 4: Start with well-bounded cases

Establishing unique and consistent identities for data sets is a central challenge to the data citation idea. Focusing initial data citation efforts on well-bounded cases can help to prevent getting hung up on the details of the most complex cases. Starting with simple cases can also help an organization to build knowledge and experience in relation to applicable recommendations and tools, and to show examples of the benefits of data citations.

Recommendation 5: Plan for long-term implications

Data citations come with implications of persistence. Archiving and ensuring the availability of resources over time is a fundamental precept of any scholarly communication system. As such, data citation initiatives must be undertaken with long-term implications in mind. For data sets to be citable, they must be surrounded by reliable archival storage technologies, data and metadata curation procedures, and organizational long-term data policies.

Conclusion

By making it easy for users, providing openness and transparency in how data citation tools are being applied, and leveraging the interest and expertise of the multiple communities of stakeholders, organizations can promote, enable, and embed data citations as regular components of the scholarly communication infrastructure.

2. Introduction

"Data citations" are a topic of increasing interest among federal agencies, data archives, and research institutions. As digital research data collections grow in size and usage, the desire to understand the ways that data are used to produce new scholarly products also grows. In parallel, digital technologies for identifying and linking to resources on the internet promise to make connections between scholarly publications and their underlying data more transparent and traceable. "Data citations," as the term suggests, are formal citations included in reference lists of published articles to data resources that led to a given research result. On the surface, the interest in data citations emerges out of a simple idea, namely, that data users should formally acknowledge their data sources. Upon investigation, however, data citation initiatives reveal many complications, bringing up issues of data identity, data and metadata standardization, accountability, and long-term sustainability.

This range of data citation issues was the focus of the workshop "Bridging Data Lifecycles: Tracking Data Use via Data Citations," held by the University Corporation for Atmospheric Research (UCAR) on April 5-6, 2012. Funded by the National Oceanic and Atmospheric Administration (NOAA) through the UCAR Joint Office for Science Support (JOSS), the Bridging Data Lifecycles workshop brought together 80 people from 30+ organizations to discuss many aspects of data citations. Many relevant stakeholders were represented at the workshop, including research libraries, research centers, data centers, federal agencies, universities, as well as non-profit and for-profit organizations. The multiple perspectives present at the Bridging Data Lifecycles workshop proved valuable in highlighting the ways that data citations impact different stakeholders, and the ways that different stakeholders can contribute to data citation initiatives. The workshop agenda and participant list are shown in Appendix I and II respectively.

The Bridging Data Lifecycles workshop built on a number of recent data citation-related events. The US Committee on Data for Science and Technology (CODATA) and the National Academies Board on Research Data and Information collaborated with the CODATA-International Council for Scientific and Technical Information Task Group on Data Citation Standards and Practices to sponsor a workshop in August, 2011, on data citation and attribution (see http://sites.nationalacademies.org/PGA/brdi/PGA_064019 for workshop presentations and Uhlir, 2012, for the workshop report). In addition, the National Science Foundation Directorate for Geosciences (NSF GEO) funded a workshop in March, 2011, titled "Geo-Data Informatics: Exploring the Life Cycle, Citation and Integration of Geo-Data." The report that resulted from the NSF GEO workshop provides a range of recommendations related to data citation, including recommendations relating to journals, professional societies, standards organizations, data repositories, educators, and research funders (NSF, 2011). The NSF GEO also released a data citation "Dear Colleague Letter" on March 29, 2012, which promotes data citations as necessary to ensure the transparency and openness of research data (Killeen, 2012).

A number of organizations are contributing to the growing data citation movement by developing tools and recommendations for data repositories and data users. The DataCite organization is an international federation of libraries and data centers that is promoting the

practice of citing data (http://www.datacite.org). DataCite provides Digital Object Identifier (DOI) registration services specifically designed for data sets, and is developing a community of libraries and research organizations around the world with the goal of promoting citations to data. The Federation for Earth Science Information Partners (ESIP) in the US and the Digital Curation Centre (DCC) in the UK both released data citation guidelines in 2011. The ESIP guidelines are targeted to data repositories in the geosciences (ESIP, 2012), while the DCC guidelines are more discipline-agnostic and are targeted toward multiple stakeholders, including authors, repositories, and publishers (Ball & Duke, 2011).

With so much activity related to data citations already ongoing, the goal of the Bridging Data Lifecycles workshop was to develop a cohesive view of the important challenges and impediments to data citation initiatives, as well as of the tools, techniques, and practices available to make such initiatives successful.

3. Themes

The topics presented and discussed at the Bridging Data Lifecycles workshop ranged widely. Data citations are intimately tied to many other important data curation issues, including data identification, discovery, management, and preservation. As the workshop proceeded, a number of themes emerged. Some were introduced by speakers, some by attendees, and some manifested as an undercurrent throughout the discussions. This section outlines a few of the most salient themes.

3.1 Data "publication" and "citation"

Data citation initiatives are often tied to the idea that data sets should be published just as other kinds of scholarly products (Klump, et al., 2006; Green, 2009). Looking at the similarities and differences between traditional scholarly publications and digital data sets illustrates some important considerations for the act of citing data.

First, why do scholars cite a journal article, conference paper, or a book? Debates about scholars' motivations and reasons for citing particular resources in their scholarly works have been ongoing for more than 50 years. In a review of citation analyses, Nicolaisen (2007) notes that theories of citing behavior typically suggest that scholars cite prior works in order to acknowledge intellectual debts, provide the historical lineage of knowledge, and to guide readers to sources. These motivations are heavily based on normative expectations and behaviors, namely, that scholars cite resources because they are trained and expected to cite resources in particular ways within their academic communities.

The same motivations appear to hold for citing data sets. Data sets exist in the historical lineage of journal articles, conference papers, etc., and at the same time, a citation can serve as a direct way to guide readers to a data set. The most widely espoused theories of citation behavior, including those reviewed by Nicolaisen, do not consider whether the properties of digital resources, such as web sites and data sets, might have an effect on why or how scholars create

citations. This is a significant gap in our understanding of citation behavior, and is an impediment for data citation initiatives.

Mark Parsons, in his Bridging Data Lifecycles workshop presentation (and in a paper of the same name, in press, co-authored by Peter Fox), asked the question: "is data publication the right metaphor?" Debates about data "publications" and "citations" draw on metaphors established and embedded within the scholarly communication system. A journal article is the prototypical example of a scholarly publication. The publication of a journal article involves making a non-changing print or electronic document available to the public (or some subset of the public) at a particular time. Transferring this notion of publication over to digital data sets is problematic. Unlike journal articles, data sets often change after they are first made available, and, as a result, are often released and re-released multiple times. The idea of a data "publication" involves stretching a metaphor that is not a perfect fit.

The imperfect fit of the publication metaphor for data sets has direct implications for the notion of data citation. The data citation initiatives noted in the introduction identify a number of purposes and benefits of data citations. Data citations can:

- Directly link scholarly articles to their underlying data
- Increase the reproducibility and transparency of data and science, thereby increasing the accountability of researchers and data stewards for their research products
- Help identify data use
- Help to make the impacts of data sets more traceable
- Help to assign professional credit to data authors and stewards

Each of these benefits holds true for traditional publications. Citations to journal articles can link articles together, make transparent the use of previous literature, indicate the impact of individual articles, and be used as metrics when assigning professional rewards to article authors.

Revisiting the imperfect fit of the publication metaphor for data sets, however, sheds light on data citation challenges. As discussed more in section 4.2, developing consistent ways to create persistent web links to digital data sets can be difficult because data sets, unlike traditional scholarly publications, are extremely variable in structure, representation schemes, and access mechanisms. In addition, section 3.5 notes how tracking data use is difficult because data sets are inconsistently cited by data users, and inconsistently registered in citation indices when data sets are cited. In contrast, citation practices and indices for traditional academic publications are much more standardized.

In his workshop address, Mark Parsons also noted that other metaphors might be as useful, or more useful than the "publication" metaphor, when thinking about developing data access and citation systems. Parsons and Fox (in press) present two alternate metaphors: data "ecosystems" and data "infrastructures." "Ecosystems" have the implication of many inter-related entities that grow and change over time. "Infrastructures," on the other hand, such as the interstate highway system, water and sewage systems, and electrical grids, provide substrates for the development of wide-ranging socio-technical systems. How would the mechanisms for linking to data, tracking data use, assigning credit for data, etc., differ for data "ecosystems" or "infrastructures"

than for data "publications"? Perhaps more to Parsons and Fox's point, how can multiple metaphors be used together to enable approaches to data management, curation, and citation challenges that are not limited by any single view?

While no perfect answers to these questions were proposed, the many implications of the data "publication" metaphor were a recurring theme in the Bridging Data Lifecycles workshop.

3.2 Identifiers and/or locators

The difference between identifiers and locators can be readily explained from a conceptual point of view (Chun, Lee, & Choi, 2011). Identifiers are a mechanism that establishes the uniqueness of an object and can be used to reliably identify an object. With digital resources, identifiers typically take the form of a unique character string, such as "ucar.cgd.ccsm4.c40.t62x1.verif.01." Identifiers are typically embedded into the object in a special metadata field or as the name of a file (Wynholds, 2011). Locators, on the other hand, are a mechanism to indicate where an object can be found. With web-based digital resources, the most common location mechanism is the Uniform Resource Locator (URL), such as http://www.google.com.

In practice, however, mechanisms for the identification and location of digital resources are often intertwined. Numerous schemes have been developed for identifying and locating digital objects, including URLs and URIs (Uniform Resource Identifiers), Persistent URLs (PURLs), Handles, Digital Object Identifiers (DOIs), and Archival Resource Keys (ARKs), among others. These schemes are often referred to as "persistent identifiers" or "unique identifiers," but many of them function primarily as locators (Duerr, et al., 2011). For the purposes of citing a digital resource, typically what is desired is an "actionable identifier." An actionable identifier is one that delivers the object that it identifies (Kunze, 2003). It should be noted, however, that data citation recommendations typically recommend that a data citation point to a "landing page" for a data set, not to a data resource itself. A landing page is a web page that provides metadata about a resource (or set of resources), as well as links to the resource itself. Thus, for citation purposes, an actionable identifier can be defined as an identifier that delivers the landing page for the object that it identifies. Landing pages are discussed further in Section 3.4 below.

This section outlines DOIs and ARKs. During the Bridging Data Lifecycles workshop, these were the most commonly discussed identifier/locator schemes. For comparative information about more than just these two schemes, see Duerr, et al., (2011) and Tonkin (2008).

Digital Object Identifiers (DOIs)

DOIs are the most common type of actionable identifier used within our current global scholarly communication systems. DOIs are most commonly assigned to journal articles, but are seeing a growing use for data. The DOI system provides a unique identifier scheme and an associated URL resolution service (Paskin, 2010). For example, the DOI "10.5065/D6RN35ST" is assigned to the dataset of the North American Regional Climate Change Assessment Program (NARCCAP). When entered into a web browser in URL form, http://dx.doi.org/10.5065/D6RN35ST, the DOI resolves to the current web location of that resource, which, as of this writing, is http://www.earthsystemgrid.org/project/NARCCAP.html.

Organizations that are interested in assigning DOIs to scholarly resources must 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 are both intermediaries between digital resource publishers and the DOI system. They provide DOI resolution services, 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.

While providing a similar functionality to ARKs, PURLs, and Handles, DOIs are more familiar and accepted among scientific communities and scholarly publishers, and as such are more prominant in data citation initiatives (Duerr, et al., 2011).

Archival Resource Keys (ARKs)
ARKs are actionable identifiers developed by the California Digital Library (CDL). ARKs are similar to DOIs in that they are identifiers that can be used in URL form to provide persistent locators for web resources. For example, the ARK "85065/d6rn35st", which is also assigned to the NARCCAP data set, can be used as an actionable identifier in URL form as http://n2t.net/ark:/b5065/d6rn35st. The CDL has developed the EZID service (http://n2t.net/ezid) and partnered with the DataCite organization to enable organizations to register both DOIs and ARKs.

Joan Starr's Bridging Data Lifecycles workshop presentation noted that while ARKs and DOIs provide similar capabilities as actionable identifiers, ARKs differ from DOIs in a number of important ways:

- ARKs can be deleted, DOIs cannot.
- ARKs can be used to access metadata about the digital resource, through the use of "inflections". For example, adding a "?" to the end of an ARK in URL form will return a metadata description of the resource instead of the resource itself. Similarly, adding a "??" to the end of an ARK in URL form will return a statement of the host organization's commitment to that resource.
- ARKs support "suffix passthrough," allowing one ARK registration to identify many thousands of extended ARKs in order to represent the many constituent parts of a digital resource.

Identifiers vs. Locators Revisited
DOIs and ARKs can serve as identifiers or locators, but their main use in the data citation context is as actionable identifiers, that is, as identifiers and locators together. In other contexts, non-actionable identifiers might be more appropriate. Digital archives assign identifiers to digital objects for many reasons, including for management, tracking, and preservation purposes. Many non-actionable identifier schemes also exist, and archives often develop their own internal identifier schemes. For example, the Universal Unique Identifier (UUID) is one of the more well-known identifier types (Leach, Mealling, & Salz, 2005). The UUID system was developed as a globally unique identifier scheme that has no central coordinating agency or service. Any

individual or organization can create UUIDs using everyday programming tools, such as C, C++, Java, MySQL, Python, and many others. The universal uniqueness of UUIDs is ensured by the structure of the UUIDs themselves. UUIDs consist of 32 hexadecimal digits, displayed in five groups separated by hyphens. For example, "0d340500-ac11-11e1-b003-00188b0bd8d1" is a properly formed UUID.

Depending on the identifier and/or locator scheme used, the identifier and the locator for a data set may be the same thing, or they may be two separate pieces of information. UUIDs are purely identifiers. They are not embedded within a larger location service like DOIs and ARKs. UUIDs, or any other non-actionable identifier, can be included in citations if appropriate, but their function should not be confused with the citable locator function of actionable identifiers.

3.3 Data citation stakeholders and responsibilities

A wide ranging set of stakeholders participated in the Bridging Data Lifecycles workshop. Different stakeholders have different roles, responsibilities, and potential contributions with respect to data citation initiatives. The following list outlines some of the stakeholders that were identified and discussed during the Bridging Data Lifecycles workshop, as well as their respective contributions to data citation initiatives. Some contributions are unique to particular stakeholders, while some cross-cut multiple stakeholders.

- Federal funding agencies
  - Promote and facilitate open access to data
  - Encourage data citation
  - Promote professional recognition and rewards for data management, archiving, and citation
- Research organizations
  - Promote and facilitate open access to data
  - Develop long term data policies that promote persistence and citation of data collections
  - Encourage data citation
  - Promote professional recognition and rewards for data management, archiving, and citation
- Data archives/centers/repositories
  - Provide data curation services
  - Provide recommended data citations for data sets
  - Work with data creators/collectors and users to raise awareness and use of data citations
- Libraries
  - Coordinate institutional units around data citation issues
  - Liaise between scientists and data archives
  - Provide expertise with actionable identifiers and bibliometric techniques
  - (In particular cases) serve as data repository (with the corresponding additional roles shown above)
- Academic publishers
  - Set and enforce data citation policies and standards

    - Work with data repositories to enable cross-linking between data sets and publications
    - Promote and enable the persistence of data collections and data citations
- Data collectors/creators
    - Ensure that data are archived in a citable form by submitting data to an archive/repository with appropriate metadata.
    - Encourage colleagues and students to cite data in their publications.
    - List archived data sets on CVs as scholarly products/accomplishments.
- Data Users
    - Cite data in publications!
    - Inform data providers about publications that led from data use

As the NSF GeoScience Directorate's data citation "Dear Colleague Letter" notes, establishing data citation as common practice will require an "evolutionary transformation" across institutions (Killeen, 2012). No one stakeholder can unilaterally develop expertise, educate researchers, and provide appropriate incentives for data citations. With coordination, collaboration, and complimentary initiatives, however, the stakeholders listed above can initiate an evolution of data citation practices within academic communities.

3.4 Citations as metadata

Many recommendations exist that detail the components of data citations (see for example ESIP 2012; Ball & Duke, 2011; Lawrence, et al., 2011; Mooney & Newton, 2012). The details of these data citation recommendations differ from case-to-case. For example, the ESIP recommendations are more specific than other recommendations, being targeted towards data curators and focused on earth science data. Among these recommendations, however, a few types of information are consistently called out as being essential components of a data citation, including an author, title, date of publication, publisher, and a data set identifier and/or locator. These commonly recommended elements closely follow the recommended citation formats that exist for scholarly works such as books and articles, such as those provided by the Chicago and American Psychological Association style guides.

Data citations are a kind of metadata. Citations provide brief descriptions of data sets in order to allow a reader to identify and locate a given resource. Citations, however, do not (and cannot) provide descriptions of a data set that are complete enough to allow a user to immediately understand and use the cited resource.

In his Bridging Data Lifecycles workshop presentation, Ted Habermann, from the NOAA National Geophysical Data Center, described data citations as "discovery metadata," and contrasted the information found in a citation to the suite of information required to understand and use data sets. To ensure that data sets remain understandable and usable at decadal time scales and for unanticipated users, data archives typically compile extensive metadata collections to document and complement the data sets themselves. A citation can and should contain only a small subset of the compiled metadata for a particular data set.

All data citation recommendations recommend the inclusion of a data set identifier and/or locator that points to the archival location for the data set being cited. As noted above in Section 3.2,

data citation recommendations typically recommend that a data citation point to a "landing page" for a data set, not to a data resource itself. As Ball and Duke (2011) note, a landing page should "enable readers to ensure they have located the right dataset, to (re-)familiarise themselves with the research context and supporting documentation, to consider licence terms prior to downloading and to switch to a more recent version (or otherwise-formatted representation) of the data if required" (pg. 10). Landing pages provide descriptive metadata about the resource(s) that they represent, and provide one or more mechanisms to access the data resources of interest. Landing pages often also provide additional links to secondary supplemental or documentary materials. The main purposes of landing pages are 1) to provide relevant information about resources, and 2) to ensure that a citation does not provide a direct link to a download of a prohibitively large data file.

3.5 Tracking data use

Data citations are intended to indicate the use and impact of data sets. In order to assess the use and impact of data sets, however, data citations must be tracked and compiled, just as citations to traditional publications are tracked and compiled by citation indexing services today.

Citation indexing services, such as the Thompson Reuters Web of Science database, the Scopus database, and the Google Scholar service, are potentially important participants in the process of tracking data citations. Currently, these indexing services do not track citations to data sets. Multiple Bridging Data Lifecycles workshop speakers noted that the DataCite organization has been in discussions with the Thomson Reuters corporation to ingest DataCite DOI metadata into the Web of Science index. About two months after the Bridging Data Lifecycles workshop took place, Thomson Reuters announced their development of a "Data Citation Index" (Thomson Reuters, 2012). While the Thomson Reuters product was not yet released at the time of this writing (nor the details for its scope or functionality), the development of such indices might significantly change the ways that data citations are tracked and counted.

Without established data citation indices, data archives need to track data use in other ways. Currently, many data archives use data download counts as one indicator of data use. Data download counts can be compiled via custom download logging processes, or via widely used web site analytics packages, such as Google Analytics. Data archives often require users to log in prior to downloading data sets. Log-ins allow data archives to get more detail about their users, as well as enable data archives to contact users if/when data sets have been updated. Data archives are then able to use download counts and any information provided about users through the registration process to show patterns of data access across a user community. For example, Jacobs and Worley (2009) show a couple of examples of how user download patterns for the NCAR Computational & Information Systems Laboratory's Research Data Archive (CISL RDA, http://rda.ucar.edu/) changed between 2006 and 2007. Steve Worley, in his Bridging Data Lifecycles workshop presentation, showed more recent examples of changing use patterns for the CISL RDA that were based on usage download counts.

Tracking data citations, however, requires different processes than tracking data downloads. Major (2011) conducted an intensive analysis of the use of NASA Earth Observing System (EOS) instrument data as manifested in the scholarly literature. To compile his study, Major searched through the Web of Science database, querying for specific instrument keywords. In

lieu of having an established indexing service for data citations, this kind of manual searching through article databases is the most accurate way of compiling data use information. Searching through databases, however, is time consuming, and is complicated by well known challenges and caveats (Smith, 1981). Citation practices vary widely in different scientific specializations, data sets might be cited or acknowledged in many ways, citations may contain errors, and article databases themselves vary in the types of sources that they index (see for example Meho & Yang, 2007, and Falagas, et al., 2008, for comparisons of the Web of Science, Scopus, PubMed, and Google Scholar databases).

Developing ways to reliably track data citations will be an important task as data citation initiatives proceed, but standard citation-based evaluation methods do not account for all kinds of scholarly work. Even conference papers, a widely cited type of scholarly resource, are poorly covered in the most comprehensive citation indices (Wainer, Goldenstein, & Billa, 2011). Thus, it should be expected that the integration of data citations into standard citation indices will be slow and piecemeal. Manual or semi-automated ways of compiling data citations will likely continue to be the norm for the immediate future.

4. Challenges

With a number of salient motivations and broad national and international backing, data citation initiatives are growing in visibility and scope. Several challenges, however, specific to data citation practices and policies were topics of discussion at the Bridging Data Lifecycles workshop. The three most salient challenges are discussed in this section: 1) the fact that data citations are currently uncommon, 2) the difficulty in formalizing a citable object, and 3) the uncertain long-term implications of data citations. Discussions around these challenges ranged from the conceptual boundaries of a data set to the mundane everyday reasons why a data user does or does not cite a data set.

4.1 Data citation, an uncommon event

The biggest challenge related to data citation is that data users currently do not cite data as a matter of common practice. A first impediment is that researchers often do not actively share data in the first place. Researchers might not share data for a number of reasons, including concerns about being scooped, fear of incorrect or improper use, difficulty of producing data in a sharable form, and many other reasons (Costello, 2009).

Even when data are widely shared, which is the case with data produced and provided by government agencies, data users do not commonly cite data sets in formal ways. Parsons, Duerr, and Minster (2010) showed how the users of a NASA remote sensing data set between 2002 and 2009 provided formal citations to the data set in only a small proportion of articles in which data use was acknowledged. This lack of formal data citation is common across physical and social sciences. Two different studies conducted 17 years apart showed that data citation was not, and is not, a common practice in the social sciences (Sieber & Trumbo, 1995; Mooney & Newton, 2012).

Instead of formally citing data sets, data users typically acknowledge data use in the text of a document, or in an article's acknowledgement section. Major (2011) showed how such data acknowledgements can be used to extrapolate data usage from published articles, but free-text acknowledgements are prone to incompleteness, and do not directly provide a means to find and access the data set being acknowledged (Parsons, Duerr, & Minster, 2010).

Why are data citations uncommon? One primary reason is that researchers receive professional rewards (such as promotion and tenure) for publishing books and articles, and for receiving citations to those products. The creation and citation of data sets are not yet activities that are part of professional evaluations for most research organizations. Thus, there are few career incentives for researchers to spend time documenting and packaging data sets so that they can be citable objects, and similarly, few career incentives exist for researchers to spend time and effort finding and including a citation to a data set in an article's bibliography. In fact, anecdotal evidence suggests that researchers have worries that citations to data sets will reduce the numbers of citations to articles that describe the findings derived from a data set. One question to this effect during the Bridging Data Lifecycles workshop asked about the impact of data citations on an author's citation rankings. If people start citing data sets, will they stop citing papers that describe the experiment or research project that led to the creation of a data set? This is an open question, though a citation to a data set should not preclude a citation to any other kind of relevant scholarly product.

Another reason why data citations are uncommon is that data users often do not know how to create data citations. As discussed above, a number of data citation recommendations exist, but these are still relatively new, and not well known outside of the small group of people who are developing and promoting data citation initiatives. As Tim Killeen noted in his presentation at the Bridging Data Lifecycles workshop, these recommendations can be difficult for non-experts to understand and use. Functional questions of what information to include in a citation are often unclear to typical data users, as are more conceptual questions about the boundaries of citable objects, which are discussed more in the next section. Traditional citation style guides, such as the Chicago, American Psychological Association (APA), and Modern Language Association (MLA) style guides are little help, as the few data citation recommendations that exist in such guides are highly variable (Newton, Mooney, & Witt, 2010). In addition, funding agencies, data repositories, and journals have few formal policies specific to data citation (Weber, Piwowar, & Vision, 2010).

In summary, data users are often 1) unaware that they can and should cite data sets, 2) unsure of how to cite data sets, and 3) lacking career motivations to forward data citations as a common activity.

4.2 What is the citable object?

Defining a "data set" is a difficult and highly situated task. Digital objects can be considered to be part of a data set if they fall in a particular grouping, contain representations of particular content, are considered to be related, or can be used for a particular purpose (Renear, Sacchi, & Wickett, 2010). Data sets are often combined into composite data sets, or pulled apart into sub-sets. In addition, many data sets, such as climate observations, stock prices, and social media feeds, are highly dynamic, changing on a daily or weekly basis.

The indistinct and dynamic boundaries around data sets complicate the process of recommending and creating data citations. Wynholds (2011) describes this boundary problem as the challenge of establishing distinct identities for data sets. To have a distinct identity, according to Wynholds, a data set must have at least four characteristics: 1) it must be a semantically concrete object, 2) it must have its identity embedded and/or inseparable from the data objects themselves, 3) it must have a stable notion of authorship, and 4) it must be translatable into a mechanism for retrieval and citation.

Each of these characteristics is problematic in relation to citing data sets.

1. Data sets are often highly dynamic, which means that semantic concreteness is an elusive characteristic. Duerr, et al., (2011) categorize this problem as the need to establish the "scientific uniqueness" of a data set, and note that it is a problem that no existing actionable identifier scheme (e.g. DOIs, Handles, etc.) can solve.
2. Embedding identity into data sets is also difficult. Books have title pages. Journal articles and conference proceedings have embedded identities in the form of standard headers on the first page that list the article title and author(s). Data sets have no standard form of embedded identity. Individual data repositories can develop standard practices for creating embedded headers, but such practices rarely cross institutional or organizational borders.
3. Authorship itself is a problematic notion in relation to data sets. Data sets are very commonly the product of collaborations. Different collaborators on the same project may have very different notions about data set (and metadata) authorship and responsibility (Wallis & Borgman, 2011). Properly attributing authorship credit must be negotiated amongst collaborators, and can be a fraught political issue.
4. From a data set retrieval point of view, data set identity is typically instantiated through persistent web identifiers such as DOIs, PURLs, and Handles. Assigning actionable locators, however, to objects that are not concrete, do not have embedded identities, and have varying forms of authorship is a challenge that is the subject of active research (Duerr, et al., 2011).

From a data citation implementation point of view, these data set boundary and identity questions manifest first and foremost in relation to granularity and version control decisions. What is the right granularity at which citations and actionable identifiers should be assigned? Many data collections are very hierarchical in organization. A given data set might contain hundreds or thousands of individual files, and users might request and use only particular sub-sets of those files. Most data citation recommendations suggest that a citations and actionable identifiers should be assigned to some level of data file aggregation, not to the individual files themselves. In part, this is to simplify the process for data users. When making granularity decisions, a number of considerations apply.

- How are data resources managed internally?
- How are data resources displayed to users?
- What does the user community consider to be a complete and sufficient citation in their field?

- How do users typically use particular data collections?
- What burdens are you putting on users to find and understand your recommended citations, and to compile the appropriate number of citations for the data that they used?
- Will the granularity support a reasonably accurate starting point for subsequent research?

Different answers to these questions might lead to different choices regarding the granularity of citation and identifier assignments.

Data set versioning raises another set of challenges for uniquely identifying and citing particular data sets. As noted above, highly dynamic data sets that change on a regular basis challenge data citation conventions. A couple of approaches have been suggested for citing and assigning actionable identifiers to continuously changing resources. The first approach is to assign citations and actionable identifiers to discrete sub-sets that do not change, such as year-long sub-sets of a growing multi-year data set (Ball & Duke, 2011; Callaghan, Lowry, & Walton, 2012). The second approach is to assign one citation and actionable identifier to the whole data set, even if it is growing, and request that users indicate in the citation that data are continuously updated, and indicate when the data set was downloaded (ESIP, 2012). For example, the citation might indicate that the data set is "updated daily" in the citation itself. A second data citation challenge related to versioning is when data sets are updated or re-released. Data updates take place for many reasons: re-calibration and re-processing, error correction, or file corruption. In these cases, the ESIP (2012) data citation recommendations suggest to assign citations and actionable identifiers to major versions of a data set, but not to minor versions. The definitions of "major" and "minor" versions are left to each data archive to decide, but a major version is considered to be an update to most or all of a data set, while a minor version is considered to be a change to a small sub-set of a data set, such as a few individual files out of many. ESIP's recommendations suggest that data providers request that users include a version indicator in the citation. The following example citation follows the ESIP recommendations to illustrate how the citation and the actionable identifier can be used in combination to provide versioning information (bolded):

- Doe, J. and R. Roe. **2001, updated daily**. The FOO Time Series Data Set. **Version 3.2.** The FOO Data Center. http://dx.doi.org/10.xxxx/1234567. **Accessed 1 May 2012**.

The challenge of identifying citable data objects manifests in a few other practical ways. At the Bridging Data Lifecycles workshop, Ruth Duerr and Lynn Yarmey both presented on newly developed data repository systems. Ruth presented about the Data Conservancy Instance (http://dataconservancy.org), and Lynn presented about the Advanced Cooperative Arctic Data and Information Service (ACADIS, http://aoncadis.org). Both of these systems allow users to create projects and upload data sets in a user-defined manner. From a data citation perspective, the question is whether the user will have the capability to define what the citable object(s) is within their projects. The alternate approach would be to program the data system to pre-define the citation granularity automatically.

Within systems that contain very diverse kinds of data, multiple approaches to assigning citations and actionable identifiers might be necessary.

Another challenge from a data citation perspective is that many data sets are distributed and replicated among multiple organizations and systems. In particular, data sets provided by government agencies in the US are copied, stored, and provided by multiple other organizations. For example, in the atmospheric sciences, the National Centers for Environmental Prediction (NCEP)/NCAR Reanalysis Data, a very large and widely used data set, can be accessed from the web sites of a number of different organizations, including:

- NOAA Earth System Research Laboratory, http://www.esrl.noaa.gov/psd/data/composites/hour/
- NCAR Computational and Information Systems Laboratory Research Data Archive, http://rda.ucar.edu/pub/reanalysis/
- NOAA National Weather Service Climate Prediction Center, http://www.cpc.ncep.noaa.gov/products/wesley/reanalysis.html

The data access mechanisms and the metadata provided vary among these three web sites, making it unclear from a citation and identity perspective if these data sets are all the same citable object.

If data are available from multiple data archives, which copy should a user cite? In the academic publishing world, this issue is known as the "problem of multiple copies," and has been a known issue for a number of years (Beit-Arie, et al., 2001). In the context of journal articles, however, there is still an assumption that even when multiple copies of an article might exist, they are copies of the same thing. With data, this is not an assumption that can be made. Data might be reformatted for a particular research community, reprocessed by a data archive to enable additional services to be provided, or supplemented by additional metadata. Thus, document-centric techniques for tracking multiple copies of web-based resources do not immediately transfer to research data curation systems.

In sum, unambiguously referring to particular data sets is difficult. Referring to anything on the web is an inherently ambiguous process (Hayes & Halpin, 2008), and the dynamic and boundary-challenging characteristics of data sets add to the difficulty. URL/URI-based schemes for actionable identifiers (including DOIs, ARKs, PURLs, etc), while a useful and necessary first step to enabling data citations, cannot in-and-of-themselves solve the problem. URLs and URIs are indexical by nature (Thompson, 2010). An indexical is a word or name that has different interpretations based on its use (Agre, 1997, pg. 230). For example, the word "here" refers to different places each time it is used. Similarly, URLs and URIs can only return whatever resource they refer to at the moment they are requested by a web browser or computer program. "In practice, web architecture does not determine what any names, including URIs, refer to. It only determines what they access. The relationship of reference is determined by the users of the URI" (Hayes & Halpin, 2008, n.p.). In other words, decisions about how to assign citations and actionable identifiers to web-based resources, including research data, cannot completely eliminate ambiguity. Data providers can minimize ambiguity, however, by using consistent approaches to citation and actionable identifier assignment, creating citation landing pages that provide easy to understand descriptions of data sets, and enabling data users to easily find and use recommended data citations.

4.3 Persistence and longevity

By promoting data sets as citable resources, data archives are moving into the territory traditionally held by academic publishers. The problems in applying the "publication" metaphor to data sets were introduced above in Section 3.1, but many of the processes and technologies being applied to the data citation issue do have root in the publication world. For example, the DOI technology initially was developed and implemented by members of the academic publishing industry (Paskin, 2010). Within the academic publication industry, there is an understanding that a resource, once published, will be available indefinitely. Assigning a DOI is tied to the idea of persistence, as it is a mechanism for ensuring web-based accessibility of published and cited resources. As Paskin notes, "DOI names are intended to be persistent identifiers: no time limit for the existence of a DOI name is assumed in any assignment, service, or DOI application" (Paskin, 2010, pg. 1590).

Academic publishers are explicit about the importance of persistence of cited resources. In his presentation at the Bridging Data Lifecycles workshop, Bill Cook, who recently served as director of publications for the American Geophysical Union (AGU), outlined AGU's data citation policy, which has been in place since 1996. This AGU policy makes clear that citations are embedded within larger data archiving institutions:

> "[D]ata 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" (AGU, 1996).

Persistence is a central consideration in AGU's policy. Resources held outside of persistent data archiving institutions, such as "data sets that are available only from the author," are not seen as fit for citing. The Geological Society of America (GSA), while not having an explicit data policy like AGU, indicates a similar expectation of persistence in its manuscript preparation template:

> "Papers must be formally accepted by a peer reviewed journal to be cited in a GSA publication. Unpublished data or papers in prep., submitted, in review, or in revision are not acceptable" (GSA, 2012, pg. 5).

Many publishers are not as uniform in their policies, instead allowing individual journals to formulate their own policies and guidelines for citing data sources. Weber, Piwowar, & Vision (2010) found that most journals do not have explicit data archiving or citation policies. Journal policies specific to data archiving or citation that do exist vary widely in many details, but formal publication and persistence of data sets are common requirements. One very detailed policy, from the Elsevier journal *Earth and Planetary Science Letters*, is as follows:

> "All data discussed in the text and presented in figures must either be presented in tabulated form in the paper or its supplemental files; be contained in cited, readily available, persistent sources; or be available in an accessible, persistent depository such as a public database or data archive, if it exists for the specific data type. … When citing

> published data, or showing compilations of published data in figures, sources must be explicitly provided" (Elsevier, 2012).

More typical are journal policies that detail how authors should (or should not) cite "unpublished data," such as the following policy from PLoS ONE.

> "Only published or accepted manuscripts should be included in the reference list. Papers that have been submitted but not yet accepted should not be cited. Limited citation of unpublished work should be included in the body of the text only as 'unpublished data'" (PloS, 2012).

While these examples are anecdotal in nature, they indicate the strong emphasis placed on the persistence of cited resources. They also often point to whether a resource is "published" as being an indicator of persistence. Thus, the issues pointed to earlier about whether data archives are becoming data "publishers" by promoting data citations are very relevant to the development and implementation of data citation recommendations.

The implication that emerges when looking at data citation initiatives from the perspective of persistence is that organizations assigning DOIs, or other actionable identifiers, are making a commitment to the long-term curation of the data resources themselves, as well as a commitment to maintain the identifiers that have been assigned. Data archives maintain and preserve data sets as part of their missions. Many universities and other research organizations are investigating whether (and how) to develop data curation services themselves. Assigning actionable identifiers to data sets, however, is a new process for most data providers. Once data providers start assigning actionable identifiers, maintenance of those identifiers becomes necessary. This involves keeping the identifier-to-landing page links up-to-date, as well as keeping the metadata associated with identifiers up-to-date. This also involves developing policies and processes for situations that arise in which data resources are to be removed from an archive, due to being deleted or deprecated for whatever reason. Monitoring and maintaining actionable identifiers might be a service provided by the registration agencies. EZID, for example, provides a "tombstone" service for identifiers that have stopped resolving to a live web site. As part of that service, EZID will put up a web page to which non-resolving identifiers will resolve. Other identifier registration agencies may not have such services. The long-term issue - that some data resources may be removed from a data collection at some point in the future - merits consideration and planning.

Citations included in a paper today will always be part of the scholarly record. Assigning actionable identifiers comes with a responsibility to ensure that those identifiers will continue to resolve to the appropriate web location over time. Many factors can complicate efforts to ensure persistence of digital resources, and the appropriate temporal horizon for such efforts is an open question, but considerations of persistence must be a part of any data citation initiative.

5. Recommendations and Conclusion

Data citation initiatives are in the beginning stages. Data citation recommendations have been developed by a few different groups, but those recommendations are themselves in the process of being evaluated and adopted by different communities and organizations. Those recommendations are also not very detailed in how to address the challenges identified in this report, partly because methods for resolving the challenges are not well developed, and partly because the challenges will manifest differently in different organizational contexts. Resolving those challenges within larger research and educational communities will require multi-faceted approaches at institutional, national, and international levels. Technologies must be designed to be flexible, multiple stakeholders must be engaged, and incentive structures need to align in ways that enable research infrastructures to support data citations as common practice (Edwards, et al., 2007).

This concluding section outlines some recommendations for organizations that are in the beginning stages of developing data citation initiatives. The goals of these recommendations are to identify ways to get started on the processes of assigning citations and actionable identifiers to data sets, without having solved every detailed problem. The recommendations derive from the Bridging Data Lifecycles workshop presentations, the workshop discussions, the related work identified in this report, and UCAR/NCAR's experiences in developing a data citation initiative during 2011 and 2012.

5.1 Recommendation 1: Identify what you want to achieve via data citations

As Mark Parsons noted in his opening keynote at the Bridging Data Lifecycles workshop, data citations have a number of purposes: 1) increase the reproducibility and transparency of data and scientific findings, 2) enable easier linking and access to data sources, 3) provide better ways to identify, track, and quantify data usage, and 4) provide credit to data creators/collectors and data curators. Not all of these purposes will be achievable immediately. Reproducibility and transparency are fundamental tenets of the scientific method, but in practice are often difficult to achieve, requiring concerted motivation, documentation, and effort by those who want to reproduce a particular result. Similarly, professional rewards, such as promotion and tenure, are assigned by universities and research agencies based on a wide range of criteria. The data archives that facilitate and track data citations can provide input to the professional reward process, but in most cases cannot directly provide those rewards themselves.

Particular organizations need to decide which data citation motivations, purposes, and considerations are most important in their local contexts, and should prioritize data citation-enabling efforts accordingly.

5.2 Recommendation 2: Understand the options for actionable identifier schemes

Section 3.2 outlined a few different actionable identifier schemes in detail, and how they are integral components of data citations. Understanding the similarities and differences between the different schemes is an important component of data citation efforts. When investigating and evaluating the different actionable identifier schemes, a few considerations are functionality,

availability, recognition, and ease of use. Duerr, et al., (2011) provide the most thorough comparative analysis of web-based identifier schemes, and provide evaluations of all of those factors. Duerr, et al., show that the core functionality of many of these schemes – DOIs, ARKs, PURLs, Handles, etc. – is similar, namely, to provide a persistent linking mechanism for online resources, but that some functionality differences do exist between the schemes. DOIs are the most widely used and recognized of the actionable identifier schemes, which is in-and-of-itself one of the reasons that they are most often promoted as the actionable identifier of choice for data citation initiatives. Particular use cases, however, might benefit from the different functionalities that other actionable identifier schemes provide.

5.3 Recommendation 3: Engage stakeholders

Section 3.3 illustrates the range of individuals and institutions who are data citation stakeholders. All of these stakeholders were represented at the Bridging Data Lifecycles workshop (see Appendix II for a list of the workshop attendees). The broad set of data citation-related initiatives that have taken place and are ongoing at national and international levels have been very effective at including many new stakeholders. Data citation initiatives at local scales need to continue the outreach and engagement efforts.

Within any particular organization, different capabilities and relationships exist in libraries, computing units, data management and archiving teams, administrative offices, and research teams. Configurations of people and organizational units will be different in every data citation initiative.

Stakeholder engagement is particularly important in building awareness and knowledge of data citation issues among the researchers who generate the data in the first place, and among data user communities. These two groups – data collectors/creators and data users – often overlap to a considerable degree, particularly within academic research. As noted in section 4.1, however, data citation is not yet common practice among most academic research communities. This indicates the many practical impediments to be overcome, such as confusion over how and when to cite data sets, but it also indicates the many interconnected incentives and disincentives for researchers to actually cite data.

Engaging researchers around the issue of data citation helps to raise awareness for data citation initiatives, and helps to identify potential sticking points. As Barb Losoff, from the University of Colorado, Boulder, noted in her Bridging Data Lifecycles workshop presentation, some researchers are more receptive to data services than others. Her presentation (which was based on Lage, Losoff, & Maness, 2011) focused on data services that might be developed by the university library, and discussed how particular disciplinary, personal, and institutional factors affect whether researchers are interested in such services, and the degree to which they would be willing to test new services out. The better such factors are understood at the beginning of a data citation initiative, the better efforts can be adjusted to meet the needs of the researchers who are most directly affected.

5.4 Recommendation 4: Start with well-bounded cases

Establishing unique and consistent identities for data sets is a central challenge to the data citation idea. Interconnectivities, complex relationships, and fuzzy boundaries abound when

looking closely at different kinds of data sets. Focusing initial data citation efforts on well-bounded cases can help to prevent getting hung up on the details of the most complex cases. Starting with simple cases can also help an organization to build knowledge and experience in relation to applicable recommendations and tools, and to show examples of the benefits of data citations.

What is a well-bounded data set? What characteristics of data sets lend themselves to being more clearly identified and cited as a unique entity? In many ways, this characterization will be different from case to case, but a few notable considerations are as follows:

- Is the data set well documented?
- Does the data set have a distinct start and finish? That is, is the data set open ended and continuously growing, or is it no longer including new data?
- Does the data set ever change? If it does change, does it have a clearly documented change process?
- Is there a clear point of contact, namely, someone responsible for upkeep of the data set, and able to answer questions?
- Is there a community of practice that recognizes a data set as a distinct resource?

The last two bullets in this list point back to the previous recommendation and the importance of having engaged stakeholders. With interested and engaged participants, the process of identifying data set boundaries and creating appropriate citations moves much more smoothly than otherwise.

5.5 Recommendation 5: Plan for long-term implications

Data citations come with implications of persistence. Recommending and enabling citations to data sets is equivalent to recommending that data sets become an accountable part of the scientific record, just as books, journal articles, conference proceedings, and research reports are now. Archiving and ensuring the availability of resources over time is a fundamental precept of any scholarly communication system (Van de Sompel, et al., 2004; Borgman, 2007). As such, data citation initiatives must be undertaken with long-term implications in mind. Is an organization committed to maintaining data resources over time? If so, what resources will be maintained, and over what temporal period can such a commitment be made? As implied by this last question, not all data resources can be, or need to be, archived permanently. This does not absolve a data archive from the responsibility to ensure the persistent ability to find information about a previously cited resource. In situations where data are promoted and cited as permanent resources, but are then removed from an archive or lost, the archive has the responsibility to ensure that the citation continues to provide useful information to someone who was interested in using or learning more about particular data sets.

For data sets to be citable, they must be surrounded by reliable archival storage technologies and organizational data retention and curation policies. Data citations are best promoted as one part of larger set of data management and curation policies and procedures, such as the practices recommended in the Open Archival Information System (OAIS) reference model (CCSDS, 2012). Such practices ensure the longevity of data, as well as metadata, software, and other

associated files. Without embedding data citations into wider data curation practices and institutions, data citations risk becoming another set of broken links to confound future scholars.

5.6 Conclusion: Transparency

To promote data citation as common practice, data providers must be crystal clear in their data citation recommendations and policies. Data users must know exactly what they are being asked to do. They must know what they are supposed to cite, when to do it, and how to do it. Data repositories must interpret data citation recommendations, such as the aforementioned ESIP and DCC recommendations, as appropriate for their local collections, and must present suggested citations to data users with little ambiguity. This will require data providers to make decisions about the data set identity challenges noted above. In addition, any tools that make it easier for users to find and access the proper citations for data sets should be built into data repositories.

By making it easy for users, providing openness and transparency in how tools are being applied, and leveraging the interest and expertise of the multiple communities of stakeholders, organizations can promote, enable, and embed data citations as regular components of the scholarly communication infrastructure.

6. Bibliography

Agre, P.E. (1997). *Computation and human experience*. New York: Cambridge University Press.

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

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/webfm_send/525

Beit-Arie, O., et al. (2001). Linking to the Appropriate Copy: Report of a DOI-Based Prototype. *D-Lib Magazine*, 7(9). http://www.dlib.org/dlib/september01/caplan/09caplan.html

Borgman, C.L. (2007). *Scholarship in the digital age: information, infrastructure, and the internet*. Cambridge, MA: MIT Press.

Callaghan, S., Lowry, R., & Walton, D. (2012). Data Citation and Publication by NERC's Environmental Data Centres. *Ariadne*, (68). http://www.ariadne.ac.uk/issue68/callaghan-et-al

Chun, W., Lee, T.-H., & Choi, T. (2011). YANAIL: yet another definition on names, addresses, identifiers, and locators. *Proceedings of the 6th International Conference on Future Internet Technologies* (pp. 8–12). Seoul, Republic of Korea: ACM. http://dx.doi.org/10.1145/2002396.2002399

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

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

Duerr, R., Downs, R., Tilmes, C., Barkstrom, B., Lenhardt, W., Glassy, J., Bermudez, L., et al. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, 4(3): 1-22. http://dx.doi.org/10.1007/s12145-011-0083-6

Edwards, P.N., Jackson, S.J., Bowker, G.C. & Knobel, C.P. (2007). *Understanding infrastructure: dynamics, tensions, and design*. Final report of the workshop, "History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures." http://hdl.handle.net/2027.42/49353

Elsevier. (2012). *Earth and Planetary Science Letters Guide for Authors*. http://www.elsevier.com/wps/find/journaldescription.cws_home/503328/authorinstructions

Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB Journal*. 22: 338-342. http://dx.doi.org/10.1096/fj.07-9492LSF

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

Geological Society of America (GSA). (2012). *GSA Manuscript Template*. http://www.geosociety.org/pubs/GSA_Manuscript_Template.doc

Green, T. (2009). *We Need Publishing Standards for Datasets and Data Tables*. OECD Publishing White Paper, OECD Publishing. http://dx.doi.org/10.1787/603233448430

Hayes, P.J. & Halpin, H. (2008). In Defense of Ambiguity. *International Journal on Semantic Web and Information Systems*, 4(2): 1-18. http://dx.doi.org/10.4018/jswis.2008040101

Jacobs, C.A. & Worley, S.J. (2009). Data curation in climate and weather: Transforming our ability to improve predictions through global knowledge sharing. *International Journal of Digital Curation*, 4(2). http://www.ijdc.net/index.php/ijdc/article/viewFile/119/122

Killeen, T. (2012). *Dear Colleague Letter - Data Citation*. NSF 12-058, March 29, 2012. http://www.nsf.gov/pubs/2012/nsf12058/nsf12058.jsp?WT.mc_id=USNSF_25&WT.mc_ev=click

Klump, J., et al. (2006). Data publication in the open access initiative. *Data Science Journal*, 5(0): 79-83. http://www.jstage.jst.go.jp/article/dsj/5/0/79/_pdf

Kunze, J. (2003). *Towards Electronic Persistence Using ARK Identifiers*. http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf

Lage, K., Losoff, B., & Maness, J. (2011). Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder. *portal: Libraries and the Academy*, 11(4): 915–937. http://dx.doi.org/10.1353/pla.2011.0049

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.

Leach, P., Mealling, M., & Salz, R. (2005). *A Universally Unique IDentifier (UUID) URN Namespace*. Internet Engineering Task Force, RFC 4122. http://datatracker.ietf.org/doc/rfc4122/

Major, G.R. (2011). Impact of NASA EOS Instrument Data on the Scientific Literature: 10 Years of Published Research Results from Terra, Aqua, and Aura. *Issues in Science and Technology Librarianship*, 67. http://dx.doi.org/10.5062/F4CC0XMJ

Mayernik, M.S. (2012). Data Citations: Initiatives, Issues, and First Steps. *Bulletin of American Society for Information Science and Technology*, 8(5): 23-28. http://www.asis.org/Bulletin/Jun-12/JunJul12_MayernikDataCitation.pdf

Mayernik, M.S., Daniels, M.D., Dattore, R.E., Davis, E.R., Ginger, K., Kelly, K.M., Marlino, M., Middleton, D.E., Phillips, J., Strand, G., Williams, S.F., Worley, S.J., & Wright, M.J. (2012). *Data citations within NCAR/UCP*. NCAR Technical Note, NCAR/TN-492+STR. Boulder, CO: National Center for Atmospheric Research (NCAR). http://dx.doi.org/10.5065/D6ZC80VN

Meho, L.I. & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google scholar. *Journal of the American Society for Information Science and Technology*, 58(13): 2105–2125. http://dx.doi.org/10.1002/asi.20677

Mooney, H. & Newton, M.P. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication*, 1(1): eP1035. http://jlsc-pub.org/jlsc/vol1/iss1/6

National Science Foundation (NSF). (2011). *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

Newton, M.P., Mooney, H., & Witt, M. (2010). A Description of Data Citation Instructions in Style Guides [poster]. *The 6th International Digital Curation Conference*, Chicago, Illinois. http://docs.lib.purdue.edu/lib_research/121/

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, 41(1): 609–641. http://dx.doi.org/10.1002/aris.2007.1440410120

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

Parsons, M. & Fox, P. (in press). Is data publication the right metaphor? *Data Science Journal.* https://dl.dropbox.com/u/546900/parsons_fox_metaphor_dsj_revised_submitted.pdf

Paskin, N. (2010). Digital Object Identifier (DOI®) System, In M.J. Bates & M. Niles Maack (Eds.), *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 1586-1592). http://www.doi.org/overview/DOI_article_ELIS3.pdf

Public Library of Science (PLoS). (2012). *PLoS ONE Manuscript Guidelines*. http://www.plosone.org/static/guidelines.action

Renear, A., Sacchi, S., & Wickett, K. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, 47(1): 1-4. http://dx.doi.org/10.1002/meet.14504701240

Sieber, J.E. & Trumbo, B.E. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics*, 1(1): 11-20. http://dx.doi.org/10.1007/BF02628694

Smith, L.C. (1981). Citation analysis. *Library Trends*, 30(1): 83-106. https://www.ideals.illinois.edu/bitstream/handle/2142/7189/librarytrendsv30i1_opt.pdf?sequence=3#page=88

Thomson Reuters. (2012). *Thomson Reuters Unveils Data Citation Index for Discovering Global Data Sets.* Thomson Reuters Press Release, June 22, 2012. http://thomsonreuters.com/content/press_room/science/686112

Thompson, H.S. (2010). What is a URI and Why Does It Matter? *Ariadne*, Issue 65. Http://www.ariadne.ac.uk/issue65/thompson-hs/

Tonkin, E. (2008). Persistent Identifiers: Considering the Options. *Ariadne*, Issue 56. http://www.ariadne.ac.uk/issue56/tonkin/

Uhlir, P.F. (Rapporteur). (2012). *For Attribution — Developing Data Attribution andCitation Practices and Standards: Summary of an International Workshop*. Board on Research Data and Information, Policy and Global Affairs, National Research Council of the National Academies. Washington, D.C.: The National Academies Press. http://www.nap.edu/catalog.php?record_id=13564

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

Wainer, J., Goldenstein, S., & Billa, C. (2011). Invisible work in standard bibliometric evaluation of computer science. *Communications of the ACM*, 54(5): 141-146. http://doi.acm.org/10.1145/1941487.1941517

Wallis, J.C. & Borgman, C.L. (2011). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. *Proceedings of the American Society for Information Science and Technology*, 48(1): 1-10. http://dx.doi.org/10.1002/meet.2011.14504801188

Weber, N.M., Piwowar, H.A., & Vision, T.J. (2010). Evaluating data citation and sharing policies in the environmental sciences. *Proceedings of the American Society for Information Science and Technology*, 47(1): 1-2. http://dx.doi.org/10.1002/meet.14504701445

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

7. Appendix I - Workshop Agenda

Speaker presentations archived at http://library.ucar.edu/data_workshop/

Thursday, April 5th , 2012

9:00am - 9:10am – Introduction, Mary Marlino (Director, NCAR Library)

9:10am - 9:20am – Welcome, Maura Hagan (Deputy Director, NCAR)

9:20am - 10:20am – Data publication and citation

Speaker: Mark Parsons (NSIDC - National Snow and Ice Data Center)

10:30am - 11:30am – Data citations and identifiers

Speaker: Joan Starr (California Digital Library, EZID, DataCite)

11:30am - 12:00pm - Publisher perspective on data publication and citation

Speaker: Bill Cook (American Geophysical Union)

1:30pm - 3:15pm – Data citation and identifier implementation issues

Speakers:
Matt Mayernik (NCAR)
Mike Daniels (NCAR)
Gary Strand (NCAR)
Nicole Kaplan (Colorado State University/Long Term Ecological Research network)
Mike Wright (NCAR) - Moderator

3:30pm - 5:00pm - Data Curation Service Models

Speakers:
Barbara Losoff (University of Colorado, Boulder)
Ruth Duerr (NSIDC)
Lynn Yarmey (NSIDC)
Ted Habermann (NOAA)
Matt Mayernik (NCAR) - Moderator

Friday, April 6, 2012

9:00am - 9:15am – Introduction, Mary Marlino (Director, NCAR Library)

9:15am - 10:15am – Exploring the Life Cycle, Citation and Integration of Geo-Data

Speaker: Dr. Tim Killeen (Assistant Director NSF GEO)

10:30am - 11:45am – Tracking data use: Current practices

Speakers:
Steve Worley (NCAR)
Dan Kowal (NOAA)
Leonard Sitongia (NCAR)
Matt Mayernik (NCAR) - Moderator

11:45am - 12:15pm – Closing, "How to get started without solving every detailed problem"

Speaker: Matt Mayernik (NCAR)

8. Appendix II – Workshop Participants

Last Name	First Name	Institution	City	State
Aquino	Janine	NCAR - EOL - RAF	Lafayette	CO
Assefa	Shimelis	University of Denver	Denver	CO
Barry	Roger	National Snow and Ice Data Center (NSIDC), University of Colorado	Boulder	CO
Bergstrom	Mary Linn	UC San Diego	La Jolla	CA
Bresnahan	Megan	University of Colorado Boulder	Boulder	CO
Brown-Sica	Margaret	Auraria Library University of Colorado Denver	Denver	CO
Burke	Ian	University of Denver	Denver	CO
Collins	Julia	National Snow and Ice Data Center, CIRES, University of Colorado	Boulder	CO
Cook	Bill	American Geophysical Union	Washington	DC
Daniels	Mike	NCAR - EOL	Boulder	CO
Davis	Ethan	UCAR - Unidata	Boulder	CO
Dean	Vince	NCAR	Boulder	CO
Dean	Robin	Colorado Alliance of Research Libraries	Denver	CO
Delserone	Leslie	University of Nebraska-Lincoln	Lincoln	NE
Duerr	Ruth	National Snow and Ice Data Center, CIRES, University of Colorado	Boulder	CO
Dwyer	Diana	USDA National Wildlife Research Center	Fort Collins	CO
Flemer	James	NDP, LLC.	Boulder	CO
Ginger	Katy	UCAR	Boulder	CO

Guy	Laura	Colorado school of Mines	Golden	CO
Habermann	Ted	NOAA - National Geophysical Data Center	Boulder	CO
Hamilton	Donna	Hamilton Information Services	Niwot	CO
Hart	David	NCAR/CISL	Boulder	CO
Hu	Xiao	University of Denver	Denver	CO
Humphries	Hope	INSTAAR, University of Colorado	Boulder	CO
Hunter	Nancy	Colorado State University Libraries	Fort Collins	CO
Johnson	Andrew	University of Colorado Boulder Libraries	Boulder	CO
Jones	Lance	NCAR	Boulder	CO
Jones	Karry	Equity Engineering	Strongsville	OH
Kaplan	Nicole	Natural Resource Ecology Lab Colorado State University	Fort Collins	CO
Keane	Ann	NOAA - ESRL - CSD	Boulder	CO
Kelly	Karon	UCAR - Integrated Information Services	Boulder	CO
Kiesel	Bruce	Thomson Reuters	Philadelphia	PA
Killeen	Tim	National Science Foundation	Washington	DC
Knudsen	Kay	Colorado Parks and Wildlife	Ft. Collins	CO
Kowal	Dan	NOAA - NESDIS - NGDC	Boulder	CO
Kraus	Joseph	University of Denver	Denver	CO
Lage	Katie	University of Colorado Boulder	Boulder	CO
Larsen	Suzanne	University of Colorado Boulder	Boulder	CO

Level	Allison	Colorado State University	Fort Collins	CO
Li	Jun	University of Wyoming	Laramie	WY
Losoff	Barb	University of Colorado Boulder	Boulder	CO
Maness	Jack	University of Colorado Boulder	Boulder	CO
Marlino	Mary	UCAR - Integrated Information Services	Boulder	CO
Mayernik	Matt	UCAR - Integrated Information Services	Boulder	CO
McCarthy	Deborah	University of Wyoming	Laramie	WY
McCullough	Heather	NOAA - National Geophysical Data Center	Boulder	CO
Middleton	Don	UCAR - CISL	Boulder	CO
Milan	Anna	National Geophysical Data Center - NOAA	Boulder	CO
Morse	Tami	University of Wyoming	Laramie	WY
Obeidat	Khawla	University of Colorado Denver	Denver	CO
Obenchain	Michael	University of Wyoming	Laramie	WY
Parsons	Mark	National Snow and Ice Data Center (NSIDC) University of Colorado	Boulder	CO
Phillips	Jennifer	NCAR - Library	Boulder	CO
Rilling	Robert	NCAR	Boulder	CO
Robinson	Erin	Foundation for Earth Science Information Partners (ESIP)	Raleigh	NC
Rosati	Antonia	NCAR	Boulder	CO
Schlagel	Elizabeth	National Snow and Ice Data Center (NSIDC), University of Colorado	Boulder	CO

Schmidt	Larry	University of Wyoming	Laramie	WY
Segal	Joan	NIST Boulder Labs Library - NOAA	Boulder	CO
Sitongia	Leonard	NCAR	Boulder	CO
Smallwood	Eamon	Alliance Digital Respository Regis University	Denver	CO
Smith	Cathy	NOAA - ESRL PSD	Boulder	CO
Stacey	Kimberly	Research Computing University of Colorado Boulder	Boulder	CO
Stansbury	Mary	University of Denver	Denver	CO
Starr	Joan	California Digital Library	Oakland	CA
Strand	Gary	NCAR - NESL - CGD	Boulder	CO
Strasser	Carly	University of California Curation Center California Digital Library	Oakland	CA
Van Cleave	Keith	U.S. Geological Survey, Denver Library	Denver	CO
Visnak	Kelly	University of Wyoming	Laramie	WY
Westra	Brian	University of Oregon	Eugene	OR
White	James	NCAR	Boulder	CO
Wilson	Anne	Wilson Laboratory for Atmospheric and Space Science (LASP) - CU Boulder	Boulder	CO
Worley	Steven	NCAR	Boulder	CO
Wright	Mike	NCAR Library	Boulder	CO
Yarmey	Lynn	National Snow and Ice Data Center (NSIDC), University of Colorado	Boulder	CO

9. Appendix III – List of Acronyms

ACADIS - Advanced Cooperative Arctic Data and Information Service
AGU - American Geophysical Union
APA - American Psychological Association
ARK - Archival Resource Key
CDL - California Digital Library
CISL RDA - NCAR Computational & Information Systems Laboratory's Research Data Archive
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
GSA - Geological Society of America
JOSS - UCAR Joint Office for Science Support
MLA - Modern Language Association
NARCCAP - North American Regional Climate Change Assessment Program
NASA - National Aeronautics and Space Administration
NASA EOS - NASA Earth Observing System
NCEP - National Centers for Environmental Prediction
NOAA - National Oceanic and Atmospheric Administration
NCAR - National Center for Atmospheric Research
NSF - National Science Foundation
NSF GEO - National Science Foundation Directorate for Geosciences
NSIDC - National Snow and Ice Data Center
OAIS - Open Archival Information System
PURL - Persistent URL
UCAR - University Corporation for Atmospheric Research
UCP - UCAR Community Programs
URI - Uniform Resource Locator
URL - Uniform Resource Locator
UUID - Universal Unique Identifier