CHAPTER 4

SHARING AND CITING RESEARCH DATA: A REPOSITORY'S PERSPECTIVE

▪ ▪ ▪

Elizabeth Moss, Christin Cave, and Jared Lyle[*]

I. THE IMPORTANCE OF SHARING AND ENHANCING RESEARCH DATA

Making research data accessible and usable is increasingly viewed as important—even crucial—to advancing scientific knowledge.[1] There are tangible benefits to sharing data.[2] It spawns new work through alternative analysis and updates; reduces overall research costs through reusing and repurposing existing data; and enables verification, replication, and validation.[3] Perhaps the most motivating reason to share data is that both funding agencies and scholarly journals recognize the value of publishing supporting data and are requiring authors to make their data publicly available.[4]

While it is increasingly easy to share data—through personal websites, cloud-based storage systems, email, and even removable media—informal sharing provides largely temporary and bespoke access to the data. It does

[*] Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan.

[1] Memorandum from John P. Holdren, Director of the Office of Science and Technology Policy, to the Heads of Executive Departments and Agencies, Increasing Access to the Results of Federally Funded Scientific Research (Feb. 22, 2013), *available at* http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

[2] *See* JENNY FRY ET AL., IDENTIFYING BENEFITS ARISING FROM THE CURATION AND OPEN SHARING OF RESEARCH DATA PRODUCED BY UK HIGHER EDUCATION AND RESEARCH INSTITUTES (2008), *available at* http://repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf.

[3] *See* NATIONAL RESEARCH COUNCIL, SHARING RESEARCH DATA (Stephen E. Fienberg et al. eds., 1985); Gary King, *Replication, Replication*, 28 PS: POL. SCI. & POL. 444 (1995); Margaret Law, *Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data*, 2005 IASSIST Q. 5, *available at* http://www.iassistdata.org/iq/reduce-reuse-recycle-issues-secondary-use-research-data.

[4] *See* National Science Foundation, *Dissemination and Sharing of Research Results*, http://www.nsf.gov/bfa/dias/policy/dmp.jsp (last visited June 30, 2015) (stating that "[i]nvestigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants"); Carly Strasser, *Thanks in Advance for Sharing Your Data*, DATA PUB (Nov. 20, 2012), *available at* http://datapub.cdlib.org/2012/11/20/thanks-in-advance-for-sharing-your-data/.

not address making enhancements for long-term access, preservation, and usability, including:

- metadata to adequately describe the context of the data;
- systems for the protection of confidential and sensitive data;
- unique, persistent identification and access to the files;
- long-term preservation of formats and files; and
- tools to find the data through search engines and online catalogs.

These enhancements make it possible to find and effectively reuse research data. This applies equally in the emerging "Big Data" environment of large-scale collections of heterogeneous, unstructured, messy, and partially described collections.

II. REPOSITORIES AND REPOSITORY REGISTRIES

To preserve usable research datasets, best practice is to deposit them in a data repository. Research data repositories help mediate the sharing and enhancing of data by offering trusted services that make data discoverable, usable, persistent, and citable. Below are the three most popular types of repositories.

- **Publication repositories,** like the Dryad Digital Repository,[5] are typically hosted by a journal or a scholarly press. These repositories link published works to the supporting data and thereby provide long-term digital access to the data used by that publishing house.

- **Institutional repositories (IRs),** such as the Johns Hopkins University Data Archive,[6] are typically based at universities. Once focused primarily on the storage and retrieval of faculty publications, IRs are beginning to offer data curation services.

- **Domain repositories**, such as the Inter-university Consortium for Political and Social Research (ICPSR),[7] are devoted to specific topics or disciplines. For example, ICPSR is focused on the social and behavioral sciences and can

[5] *Dryad Digital Repository*, DRYAD, http://datadryad.org/ (last visited June 30, 2015).

[6] *About Storing & Archiving Your Research Data*, JOHNS HOPKINS UNIVERSITY DATA MANAGEMENT SERVICES, http://dmp.data.jhu.edu/preserve-share-research-data/preserve-archive/ (last visited June 30, 2015).

[7] Carol Ember et al., Sustaining Domain Repositories for Digital Data: A White Paper (2013), *available at* http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf.

address data issues unique to those scientific communities.[8] (Some of the enhancements that ICPSR provides will be described in detail later in this chapter.)

These are just three examples of the many types of repositories, all of which offer varying levels and kinds of service. Subject librarians expert in relevant fields of study or in data management can provide nuanced recommendations that are helpful in steering researchers to trusted, long-lived repositories. Curated online registries of data repositories can also help. These registries attempt to catalog the expanding data repository landscape, although the listings can be minimally descriptive. Two such registries are Databib[9] and the Registry of Research Data Repositories,[10] which combined in 2014 to provide researchers with a unified discovery tool that searches one international collection of repositories.[11]

III. THE ROLE OF DATA CITATION IN SHARING AND ENHANCING DATA

Data citation is a key element of the growing data-sharing infrastructure and facilitates sharing, discovery, and proper use. It also enables data impact tracking, allowing researchers to receive credit for their contributions. Research data often form the basis for the claims made in scholarly publications, but there is no single tradition or standard common to the various fields of science demanding that these data be identified by name, let alone be associated with full attribution and location information in formal citations.

However, with properly formatted citations, data can be discovered and accessed in their archival state, allowing reuse based on metadata crucial for accurate analysis. Data sharing and archiving advocates like those in the UK's Digital Curation Centre state that formal data citation "must echo the role that traditional, journal citation has played in ensuring longevity of the scholarly record, acting as a bridge to permanent access and enabling reward systems."[12]

In practice, citing research data requires minimal effort compared to its potentially high dividends. Although the types of research data differ depending on the discipline, the information needed to cite most data is

[8] *ICPSR: A Partner in Social Science*, INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH, http://www.icpsr.umich.edu/icpsrweb/landing.jsp (last visited June 30, 2015).

[9] Databib, *About Databib*, http://databib.org/about.php (last visited June 30, 2015).

[10] Registry of Research Data Repositories, http://www.re3data.org/ (last visited June 30, 2015).

[11] re3data.org team, *DataCite, re3data.org, and Databib Announce Collaboration* (Mar. 25, 2014), http://www.re3data.org/2014/03/datacite-re3data-org-databib-collaboration/.

[12] JONATHAN RANS ET AL., ENABLING THE CITATION OF DATASETS GENERATED THROUGH PUBLIC HEALTH RESEARCH (2013), *available at* http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp051762.PDF.

surprisingly similar. The essential elements of a data citation are not unlike those for printed matter like books and journal articles, and they are listed in Box 1.

Box 1. Elements of a Good Data Citation, for Use in a Publication's References Section

- Study author/data collector;
- Study title;
- Year of publication;
- Publisher and/or distributor;
- Edition or version; and
- Unique identifier.

Of these elements, perhaps the most important is the dataset's unique identifier. Registering and assigning data with such identifiers is best practice for any trustworthy repository.[13] The University of California Digital Library notes that there are three requirements for a unique identifier. Each must be:

- actionable, meaning that it can be linked to a specific online location such as a website;
- globally unique across the Internet; and
- persistent, at least for the period in which the data have relevant research value.[14]

The type of unique identifier used is up to the repository and depends on a variety of factors, including the repository's infrastructure.

Since these short names or character strings assigned to a dataset guarantee that the data can be permanently identified independent of the data's location,[15] access to the data over time can be ensured despite any subsequent changes in hardware and software. The identifier should direct anyone who clicks on it to the latest available version of the data, or at least to their metadata, enabling the user to access the correct version and format.

Because they are machine-actionable, the identifiers also make data digitally discoverable. Unique identifiers are by design easy to find with a

[13] ALYSSA GOODMAN ET AL., 10 SIMPLE RULES FOR THE CARE AND FEEDING OF SCIENTIFIC DATA (2014), at 3, *available at* http://arxiv.org/pdf/1401.2134.pdf.

[14] John Katz, *Data Citation Developments*, DATA PUB, http://datapub.cdlib.org/2013/10/11/data-citation-developments (last visited June 30, 2015).

[15] Micah Altman & Gary King, *A Proposed Standard for the Scholarly Citation of Quantitative Data*, 13 D-LIB MAG., Mar.–Apr. 2007, *available at* doi: 10.1045/march2007-altman, *also available at* http://www.dlib.org/dlib/march07/altman/03altman.html.

search engine. Commercial bibliographic databases can collect and index these standardized citations. For example, datasets are now discoverable through highly used online bibliographic databases, such as Thomson Reuters' *Web of Science*, Elsevier's *Scopus*, or Google's *Scholar*. This leads to better quantification of data use and enables data creators to be recognized for their primary research output.

To maximize the ability of others to find a dataset, authors should use a standardized data citation with a well-recognized unique identifier. Ideally, the citation should be placed in a journal article's references section, as this portion of the article is often freely available to the public—outside of journal paywalls, where it is easier for search engines to find and harvest them. Also, formally listing datasets in the references section acknowledges them as equally important as any other source material.

IV. THE MOVEMENT TOWARD EFFECTIVE DATA CITATION IN SCHOLARLY JOURNAL PUBLICATION

A worldwide data citation infrastructure is emerging, helped along by groups attempting to harmonize principles and requirements. These include global scientific organizations like the International Council for Science Committee on Data for Science and Technology (CODATA),[16] the Research Data Alliance (RDA) Data Citation Working Group,[17] DataCite,[18] and FORCE11.[19] The *Joint Declaration of Data Citation Principles* (see Box 2), created by FORCE11 in 2013, advocates a basic set of principles that authors and publishers should consider when taking journal publication into the era of open access and digitally enhanced publishing.

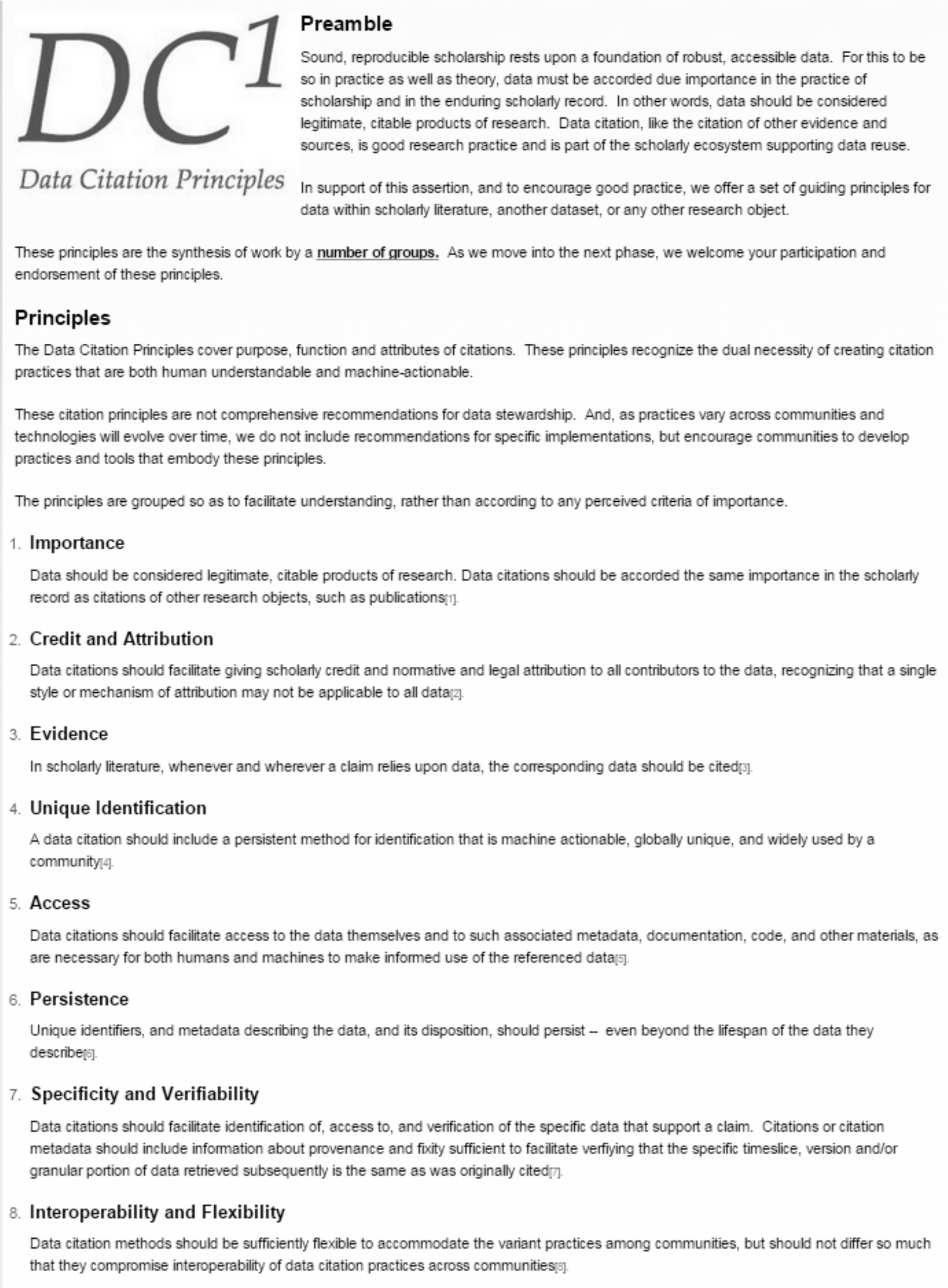
[16] CODATA, *Data Citation Standards and Practices*, http://www.codata.org/task-groups/data-citation-standards-and-practices (last visited June 30, 2015).

[17] Research Data Alliance, *Data Citation WG*, https://rd-alliance.org/groups/data-citation-wg.html (last visited June 30, 2015).

[18] DataCite, *What Do We Do?*, http://www.datacite.org/about-datacite/what-do-we-do (last visited June 30, 2015).

[19] FORCE11, *FORCE11: The Future of Research Communications and e-Scholarship*, http://www.force11.org/ (last visited June 30, 2015).

Box 2: FORCE11 Joint Declaration of Data Citation Principles[20]



Preamble

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

In support of this assertion, and to encourage good practice, we offer a set of guiding principles for data within scholarly literature, another dataset, or any other research object.

These principles are the synthesis of work by a **number of groups.** As we move into the next phase, we welcome your participation and endorsement of these principles.

Principles

The Data Citation Principles cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable.

These citation principles are not comprehensive recommendations for data stewardship. And, as practices vary across communities and technologies will evolve over time, we do not include recommendations for specific implementations, but encourage communities to develop practices and tools that embody these principles.

The principles are grouped so as to facilitate understanding, rather than according to any perceived criteria of importance.

1. **Importance**

   Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications[1].

2. **Credit and Attribution**

   Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data[2].

3. **Evidence**

   In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited[3].

4. **Unique Identification**

   A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community[4].

5. **Access**

   Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data[5].

6. **Persistence**

   Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe[6].

7. **Specificity and Verifiability**

   Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited[7].

8. **Interoperability and Flexibility**

   Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities[8].

[20] *Joint Declaration of Data Citation Principles—Final*, FORCE11, http://www.force11.org/datacitation (last visited June 30, 2015).

Beginning in 2000, journal publishers, having expanded from print-only to online publication, were able to start registering their digitized journal articles with digital object identifiers (DOIs) through the DOI registration agency CrossRef. This has made "reference linking throughout online scholarly literature efficient and reliable."[21] Partly to take advantage of the DOI infrastructure that was already well-established and widely used for identifying research articles, a DOI registration agency was created in 2009 to register DOIs for data. It is called DataCite.[22] Both individual researchers and data repositories now register their data with DOIs from DataCite's member organizations, which are located both in the United States and internationally. As it becomes common practice to store research data in repositories that provide data DOIs, the hope is that a culture will be created that rewards both data creators (by receiving credit for their work) and data reusers (enabling them to generate new publications from existing data).

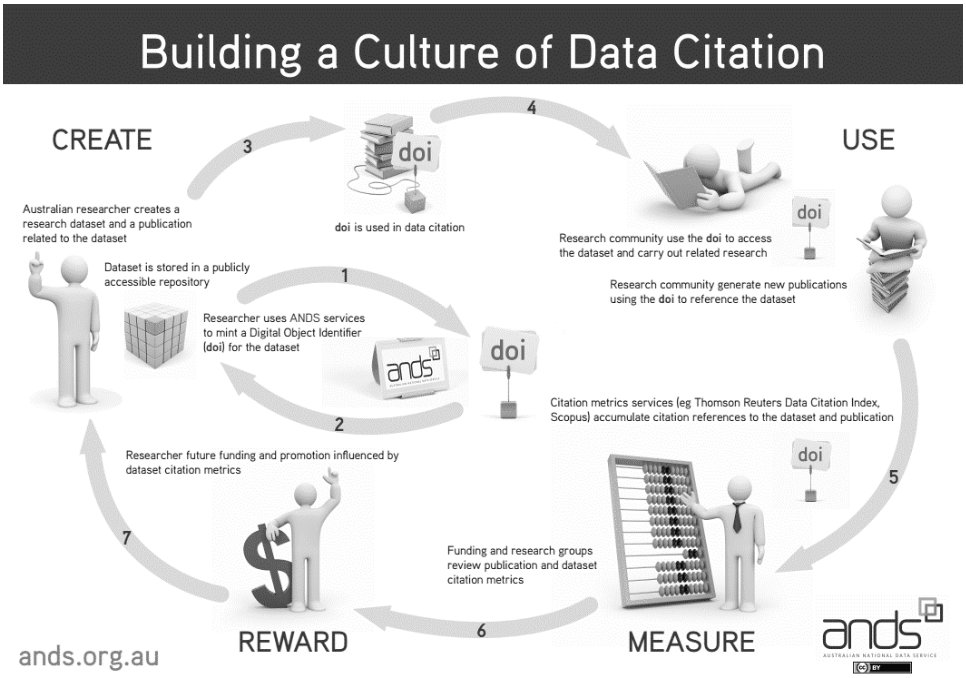
The Australian National Data Service (ANDS), a member of DataCite, illustrated how "building a culture of data citation" might work, in their diagram shown in Box 3. It depicts how archives, researchers, publishers, and funders would be able to maximize their efforts and reap the resulting rewards. ANDS provides funding to Australian institutions to assist them in developing infrastructure and guidance in support of this data-citing culture, through collaboration and outreach with librarians and researchers.[23]

[21] CrossRef, *FastFacts*, http://www.crossref.org/01company/16fastfacts.html (last visited June 30, 2015).

[22] DataCite, *Frequently Asked Questions*, http://www.datacite.org/faq (last visited June 30, 2015).

[23] Natasha Simons et al., *Growing Institutional Support for Data Citation: Results of a Partnership Between Griffith University and the Australian National Data Service*, 19 D-LIB MAG., Nov. 2013, *available at* doi:10.1045/november2013-simons.

Box 3: Building a Culture of Data Citation[24]



Building a Culture of Data Citation

CREATE

Australian researcher creates a research dataset and a publication related to the dataset

Dataset is stored in a publicly accessible repository

1 Researcher uses ANDS services to mint a Digital Object Identifier (doi) for the dataset

2

3 doi is used in data citation

4 Research community use the doi to access the dataset and carry out related research

USE

Research community generate new publications using the doi to reference the dataset

5 Citation metrics services (eg Thomson Reuters Data Citation Index, Scopus) accumulate citation references to the dataset and publication

MEASURE

6 Funding and research groups review publication and dataset citation metrics

REWARD

7 Researcher future funding and promotion influenced by dataset citation metrics

ands.org.au

As effective, formal data citation practice becomes the cultural norm, one might expect these outcomes:

- Tracking data use is similar to the tracking that takes place in bibliographic citation indices. This applies to both bibliometrics and other tracking measures such as altmetrics. "While citation metrics track formal citations, altmetrics tools such as ImpactStory use DOIs to track mentions in the social media and non-traditional scholarly communications across the Web."[25]
- Scholars who work hard to gather and describe data for others to reuse are recognized for the impact revealed.
- Formal data citation improves academic transparency, thereby enabling potential reusers to feel confident in data quality.
- Scientific innovation and evidence-based policy development improve through increased verification, refinement, and even refutation of existing data.

[24] Australian National Data Service, *Building a Culture of Data Citation*, http://www.ands.org.au/guides/data_citation_poster.pdf.

[25] Simons et al., *supra* note 23.

Fortunately, practical applications of data-citing standards are beginning to appear in scholarly publishing. There is a "growing need and eagerness among those involved in scholarly communication to agree [upon] new conventions that are practical for all."[26] This is especially so among academic journals that have begun to acknowledge the importance and value of data citation by accommodating—even requiring—its use. In some cases, journals go so far as to provide a recommended citation format and to stipulate citation policy and citation location. The journal *Science* requires that "all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*."[27] Professional associations and the journals they publish are beginning to require that authors cite data formally. For example, the *American Economic Review* (and all journals published by the American Economic Association) recommends that data citations be added to the reference list and that they "include the author name or name of the provider hosting the data, the year the data were collected or posted, the name or title of the dataset, the name of the database if applicable, and any other information necessary for one to retrieve the data."[28] The American Sociological Association's flagship journal, *The American Sociological Review*, now instructs authors to include data citations in the reference list, utilizing DOIs.[29]

Other key players in the research cycle and publishing infrastructure are seeing the value of data citation to increase data discovery, sharing, and credit. By 2011, the journal-associated repository Dryad could already estimate an "impressive scientific return" for an "ongoing financial investment in data archiving infrastructure" because of the rate of reuse of certain datasets.[30] In 2012, Thomson Reuters rolled out the *Data Citation Index* (DCI) as an additional service of the larger scholarly publishing platform known as the *Web of Science*. The DCI was advertised as "providing a comprehensive picture of research output to understand data in context and maximize research efforts."[31] According to Thomson Reuters, the DCI "provides subscription-only access to metrics associated

[26] RACHAEL KOTARSKI ET AL, REPORT ON BEST PRACTICES FOR CITABILITY OF DATA AND ON EVOLVING ROLES IN SCHOLARLY COMMUNICATION 5 (2012), *available at* http://www.stm-assoc.org/2012_07_10_STM_Research_Data_Group_Data_Citation_and_Evolving_Roles_ODE_Report.pdf.

[27] Science, *General Information for Authors*, http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#submission (last visited June 30, 2015).

[28] American Economic Association, *AEA Publications Sample References*, http://www.aeaweb.org/sample_references.pdf.

[29] American Sociological Review, *Manuscript Submissions*, http://www.sagepub.com/journalsProdDesc.nav?ct_p=manuscriptSubmission&prodId=Journal201969&currTree=Subjects&level1=N00 (last visited June 30, 2015).

[30] Heather A. Piwowar, *Data Archiving Is a Good Investment*, 473 NATURE 285 (2011) *available at* doi: 10.1038/473285a.

[31] Thomson Reuters, *The Data Citation Index: Connecting the Data to the Research It Informs*, http://wokinfo.com/products_tools/multidisciplinary/dci/about/ (last visited June 30, 2015).

with research data from global repositories covering multiple disciplines."[32] It links the data used in journal articles to repositories where those data can be found. Currently, since authors rarely cite datasets, let alone according to a common standard, Thomson Reuters cannot automate the collection of these links. Instead, it depends upon repositories to supply the DCI with lists of journal articles that they happen to know have used their data. This is an incomplete record of data use. Since properly cited data includes machine-discoverable identifiers that can be found by Web crawlers, one day the DCI and other tools may be able to automate the tracking of data references, providing a more accurate picture of how often data are used and reused in scholarship. Only then will we be able to measure what poor data citation practice has partially obscured—the vital importance of shared data to the advancement of science.

V. THE ROLE OF DISCIPLINARY DATA REPOSITORIES IN SHARING AND ENHANCING DATA

The Inter-university Consortium for Political and Social Research (ICPSR) is an archive of social and behavioral science research data dedicated to enhancing these data in order to make them discoverable, meaningful and usable, persistent, trustworthy, confidential (as needed), and citable.[33] Founded in 1962, ICPSR originated as a vehicle for sharing the much-sought-after American National Election Studies (ANES).[34] At the time, "the concept of giving access to all interested scholars to one's basic (micro) data was so foreign as to be considered 'revolutionary.' "[35] When contemplating how to build an archive for distributing his data, ANES' principal investigator, Warren Miller, and his colleagues "recognized that data preservation and dissemination were not cost-free," and so "conceived of a collegial ('consortial') mechanism for facilitating this type of data sharing."[36]

Today, ICPSR's consortium of research institutions is international, with over 750 members. ICPSR disseminates and preserves curated and enhanced data from over 8,000 studies spanning the breadth of the social and behavioral sciences. Data DOIs and citations are provided for each

[32] Simons et al., *supra* note 23.

[33] INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH, ICPSR COMMENTS ON PUBLIC ACCESS TO FEDERALLY-SUPPORTED RESEARCH AND DEVELOPMENT DATA (2013), *at* http://www.icpsr.umich.edu/files/ICPSR/ICPSRComments.pdf.

[34] Inter-university Consortium for Political and Social Research, *American National Election Study (ANES) Series*, *at* http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/3 (last visited June 30, 2015).

[35] Inter-university Consortium for Political and Social Research, *ICPSR: The Founding and Early Years*, http://www.icpsr.umich.edu/icpsrweb/content/membership/history/early-years.html (last visited June 30, 2015).

[36] *Id.*

study. In addition to the data, ICPSR provides supporting documentation in the form of codebooks, questionnaires, and publication lists to assist users in analyzing and understanding the archived data.

ICPSR is interested in helping preserve and share data even outside the consortium of members. For example, for a nominal deposit fee, anyone may immediately distribute their social or behavioral research data and supporting materials in openICPSR,[37] a public-access research data-sharing service. Although these objects are not curated by ICPSR staff, the organization issues each object a DOI and a citation. ICPSR also indexes and distributes each object through ICPSR's data catalog. Once distributed, the public can access the objects at no charge.

ICPSR also collaborates with a number of federal agencies and foundations to create and fund topical archives of data on areas such as aging, demography, criminal justice, education, and race/ethnicity. As the need for data repositories grows, this relationship between ICPSR and research funding agencies serves as a model for best practice in data management. A good example is the National Archive of Criminal Justice Data (NACJD), housed within ICPSR. NACJD takes a holistic approach by enabling researchers to collect, use, preserve, and distribute data for reuse. See Box 4 for more information about the procedures and policies established by the federal agencies funding NACJD.

Box 4: Case Study

A Model for Sharing and Citing Data Funded by United States Federal Agencies: The National Institute of Justice's Data Resources Program[38]

Housed within ICPSR, the National Archive of Criminal Justice Data is a topical archive focused on the collection, processing, and preservation of crime and justice studies. Established in 1977, NACJD is supported by the National Institute of Justice (NIJ), the Bureau of Justice Statistics (BJS), and the Office of Juvenile Justice and Delinquency Prevention (OJJDP) and maintains a collection of over 2,200 studies with over 9,000 accompanying publications of data-related literature. Each of the three federal agencies supports different types of data collection, as well as the growth of the ICPSR Bibliography of Data-related Literature (described in detail later in this chapter). Populating this citations database with crime and criminal justice references enables their data users to associate published materials with the actual

[37] Inter-university Consortium for Political and Social Research, *openICPSR*, http://openicpsr.org/ (last visited June 30, 2015).

[38] National Institute of Justice, *Data Archiving Plans for NIJ Funding Applicants*, http://www.nij.gov/funding/data-resources-program/applying/Pages/data-archiving-strategies.aspx (last visited June 30, 2015).

data used in analysis, and the funders are able to determine the extent to which their sponsored research is reused.

Since 1978, the National Institute of Justice has been accumulating an archive of hundreds of datasets resulting from projects funded through research grant programs. In 2007, NIJ added the condition that grant awardees must submit data and documentation at the completion of the research project. With few exceptions, these datasets are archived and disseminated via NACJD. Historically, this has meant that upon grant termination, the grantee archives at NACJD their computer-readable data and corresponding documentation related to NIJ, which could then be used by other federal agencies or distributed at the discretion of NIJ for other research purposes.

The latest data submission requirements encourage grant applicants to plan for data archiving at the conception of a research project. To ensure usability and avert disclosure risk, all NIJ funding applicants must submit a data archiving strategy with their research grant application. The plan should be composed of the following elements:

- Data description;
- Data collection procedures;
- Software used to collect, store, and analyze data;
- Data formats (quantitative, qualitative, geospatial);
- Direct identifiers; and
- Technical documentation (data dictionary or codebook).

And to ensure compliance with the grant requirement, NIJ withholds some amount of the grant award funds until the final report, codebook, technical documentation, and research data are submitted to NACJD.

Through the Data Resources Program, NIJ not only encourages data sharing, archiving, and secondary data use by requiring grantees to archive data at NACJD, but also, through an extension of the program, provides funding to conduct original research using existing data. NIJ gives funding priority to research projects that utilize datasets available at NACJD resulting from projects previously supported by NIJ, BJS, and OJJDP.

Recognizing the value of archiving and disseminating data, and tracking its use, NIJ provides an excellent model for how other federal agencies can encourage data to be archived at a specialized repository.

VI. THE ICPSR BIBLIOGRAPHY OF DATA-RELATED LITERATURE

Specialized or discipline-focused data repositories are keenly aware of the essential role data citation plays. Many such repositories now embed citations within metadata. Some expend much human effort to find and link data to the corresponding articles in bibliographic databases and to track data reuse through citation indexes and bibliographies. ICPSR has been a leader in implementing and promoting the use of data citations by providing a recommended data citation with its datasets since 1990 and by registering and providing DOIs with each data citation since 2008.

In 2000, ICPSR received a four-year National Science Foundation grant to enhance the discovery and reuse of data archived at ICPSR. The grant included funding for ICPSR staff to locate and link ICPSR data to the scholarly publications in which those data are used to form analyses. The outcome of the grant was the creation of the ICPSR Bibliography of Data-related Literature,[39] a continually updated database that now contains over 70,000 citations to articles, reports, and other publications. The Bibliography continues to exist today with the financial support of both the ICPSR membership and several of its topical archives. The primary criterion for inclusion in the Bibliography is that a publication must contain new analysis or extensive discussion of the original research data archived by ICPSR.

ICPSR's metadata catalog,[40] where archived studies are both described and available for download, is linked to the database where the Bibliography's citations are held. The links are two-way, in that they enable users to (1) identify a data collection of interest and link to its related literature or (2) locate a publication and link to the underlying data. Through OpenURL linking resolvers,[41] the full text of many publications is delivered straight to the user's desktop, enhancing the Bibliography's utility.

All types of data users investigate previous research based on ICPSR data with the help of the Bibliography. Instructors often direct students there to begin data-related research projects by reading some of the major works based on the data. Advanced researchers find reading the existing literature to be an effective way to begin using a dataset or to learn more about the study methodology. Reporters seeking interpreted statistics look for reports explaining particular studies. Funders want to quantify the

[39] Inter-university Consortium for Political and Social Research, *Find Publications*, http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/ (last visited June 30, 2015).

[40] Inter-university Consortium for Political and Social Research, *ICPSR: Find Data*, http://www.icpsr.umich.edu/icpsrweb/ICPSR/ (last visited June 30, 2015).

[41] Library of Congress Portals Applications Issues Group, *OpenURL Resolver Products & Vendors*, http://www.loc.gov/catdir/lcpaig/openurl.html (last visited June 30, 2015).

degree of secondary use that their sponsored research receives in order to determine whether the initial investment was valuable and whether it should be sustained. And principal investigators (PIs) check to see how much and in what way their data are being used.

As data creation becomes more valued as a legitimate product of research, PIs will become more interested in measuring the amount of use their data receive, which will require those data to be formally cited with vital information about access and attribution. ICPSR has become an advocate for making it common practice to cite data in a standard format that includes the title of the dataset, the author or creator of the data, and the year the data were collected. Along with this information, ICPSR advises that a machine-readable, unique identifier be included in the citation and listed in the references section of a publication. A model data citation is included with each study description listed on the ICPSR website (see Box 5).

Box 5. Example ICPSR Data Citation

> United States Department of Justice. Bureau of Justice Statistics. Survey of Inmates in State and Federal Correctional Facilities, 2004. ICPSR04572-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2007–02–28. doi:10.3886/ICPSR04572.v1

Although ICPSR is working with many groups in their efforts to standardize citation practices, it is still the case that "data citation is an emergent practice rather than a norm of scholarly attribution."[42] While searching the scholarly literature for articles that analyze data distributed from ICPSR, the ICPSR staff have found that the lack of formal data citation is typical, not the exception, unlike the standard practices used for citing legal cases, statutes, and journal articles. And when data are acknowledged, they rarely appear in the references section or bibliography of a publication, making it difficult to detect efficiently. Vague descriptions of the data in the abstract or methodology sections are what is more commonly found. Authors, whether they are the data creators or secondary users, often do not mention the study title, let alone cite the ICPSR-assigned citation or the DOI. Analyses by researchers like Mooney (2011)[43] and Sieber & Trumbo (1985)[44] reach similar conclusions about the inadequate practices used to acknowledge what data were analyzed in publications. This makes the job of finding data use highly labor-intensive, requiring costly human judgment and guesswork. Another de facto

[42] Simons et al., *supra* note 23.

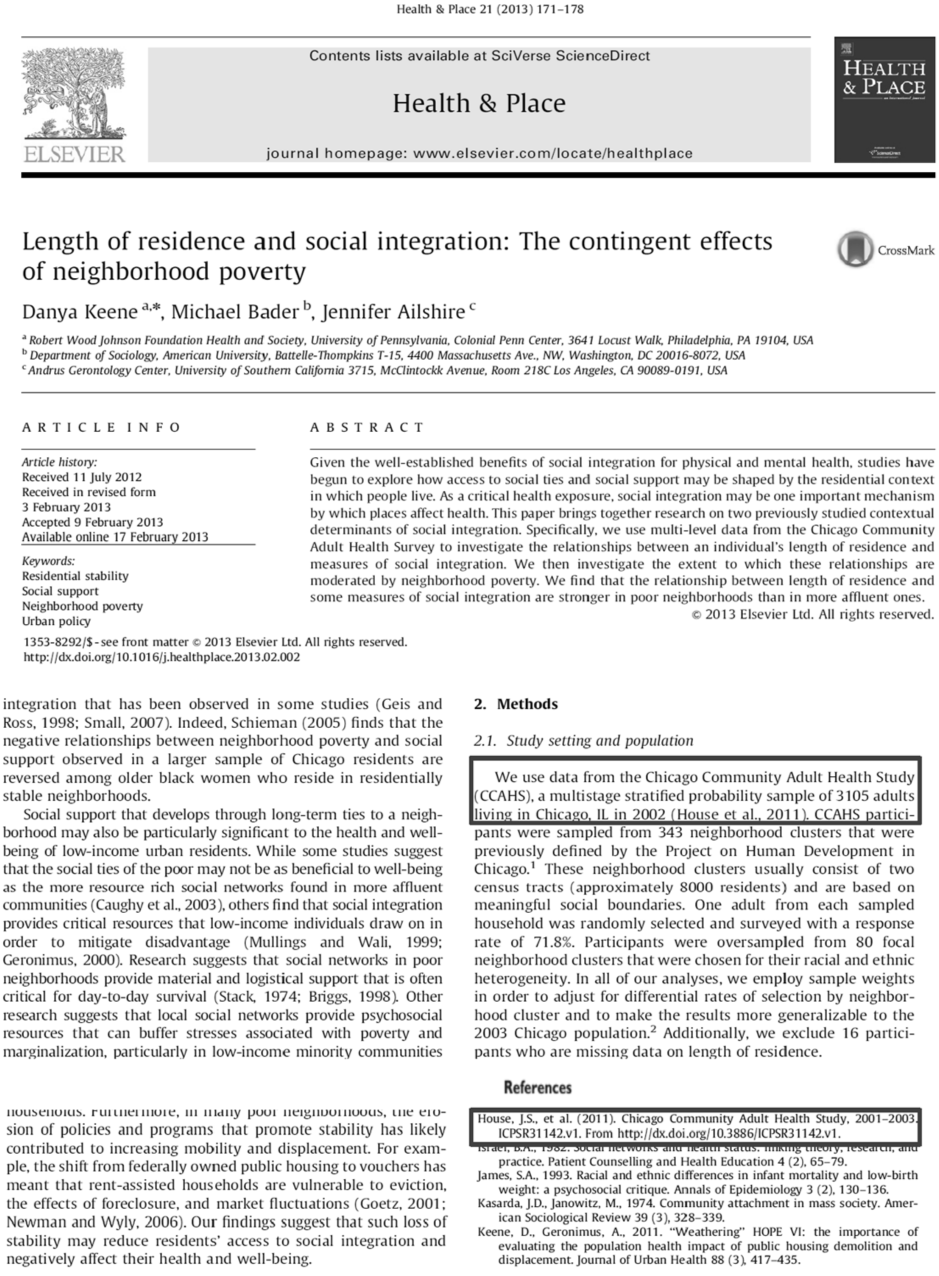
[43] Hailey Mooney, *Citing Data Sources in the Social Sciences: Do Authors Do It?*, LEARNED PUB. 24(2):99–108 (2011), *available at* doi:10.1087/20110204.

[44] Joan E. Sieber & Bruce E. Trumbo, *(Not) Giving Credit Where Credit Is Due: Citation of Data Sets*, SCI. & ENGINEERING ETHICS 1(1):11–20 (1995), *available at* doi: 10.1007/BF2628694.

outcome of this lack of citation is that readers are forced to try to infer which data were utilized. The effect is that data producers are not always receiving proper credit, and much data are effectively hidden from those who would replicate analyses or reuse data.

Some authors do cite data effectively. Box 6 shows how the authors of a journal article make sure to describe in that article's Methods section the dataset they reused in their article, parenthetically citing the authors. They then formally cite the data in the article's list of references. With this information, any reader (or Web crawler looking for data DOIs) is able to link to the data cited in the article. And the original data creators receive acknowledgement and credit for their work.

Box 6. Effective Formal Data Citation in a Journal Article



Health & Place 21 (2013) 171–178

Contents lists available at SciVerse ScienceDirect

Health & Place

journal homepage: www.elsevier.com/locate/healthplace

Length of residence and social integration: The contingent effects of neighborhood poverty

Danya Keene[a,*], Michael Bader[b], Jennifer Ailshire[c]

[a] Robert Wood Johnson Foundation Health and Society, University of Pennsylvania, Colonial Penn Center, 3641 Locust Walk, Philadelphia, PA 19104, USA
[b] Department of Sociology, American University, Battelle-Thompkins T-15, 4400 Massachusetts Ave., NW, Washington, DC 20016-8072, USA
[c] Andrus Gerontology Center, University of Southern California 3715, McClintockk Avenue, Room 218C Los Angeles, CA 90089-0191, USA

ARTICLE INFO

Article history:
Received 11 July 2012
Received in revised form
3 February 2013
Accepted 9 February 2013
Available online 17 February 2013

Keywords:
Residential stability
Social support
Neighborhood poverty
Urban policy

ABSTRACT

Given the well-established benefits of social integration for physical and mental health, studies have begun to explore how access to social ties and social support may be shaped by the residential context in which people live. As a critical health exposure, social integration may be one important mechanism by which places affect health. This paper brings together research on two previously studied contextual determinants of social integration. Specifically, we use multi-level data from the Chicago Community Adult Health Survey to investigate the relationships between an individual's length of residence and measures of social integration. We then investigate the extent to which these relationships are moderated by neighborhood poverty. We find that the relationship between length of residence and some measures of social integration are stronger in poor neighborhoods than in more affluent ones.

1353-8292/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.healthplace.2013.02.002

integration that has been observed in some studies (Geis and Ross, 1998; Small, 2007). Indeed, Schieman (2005) finds that the negative relationships between neighborhood poverty and social support observed in a larger sample of Chicago residents are reversed among older black women who reside in residentially stable neighborhoods.

Social support that develops through long-term ties to a neighborhood may also be particularly significant to the health and well-being of low-income urban residents. While some studies suggest that the social ties of the poor may not be as beneficial to well-being as the more resource rich social networks found in more affluent communities (Caughy et al., 2003), others find that social integration provides critical resources that low-income individuals draw on in order to mitigate disadvantage (Mullings and Wali, 1999; Geronimus, 2000). Research suggests that social networks in poor neighborhoods provide material and logistical support that is often critical for day-to-day survival (Stack, 1974; Briggs, 1998). Other research suggests that local social networks provide psychosocial resources that can buffer stresses associated with poverty and marginalization, particularly in low-income minority communities

households. Furthermore, in many poor neighborhoods, the erosion of policies and programs that promote stability has likely contributed to increasing mobility and displacement. For example, the shift from federally owned public housing to vouchers has meant that rent-assisted households are vulnerable to eviction, the effects of foreclosure, and market fluctuations (Goetz, 2001; Newman and Wyly, 2006). Our findings suggest that such loss of stability may reduce residents' access to social integration and negatively affect their health and well-being.

2. Methods

2.1. Study setting and population

We use data from the Chicago Community Adult Health Study (CCAHS), a multistage stratified probability sample of 3105 adults living in Chicago, IL in 2002 (House et al., 2011). CCAHS participants were sampled from 343 neighborhood clusters that were previously defined by the Project on Human Development in Chicago.[1] These neighborhood clusters usually consist of two census tracts (approximately 8000 residents) and are based on meaningful social boundaries. One adult from each sampled household was randomly selected and surveyed with a response rate of 71.8%. Participants were oversampled from 80 focal neighborhood clusters that were chosen for their racial and ethnic heterogeneity. In all of our analyses, we employ sample weights in order to adjust for differential rates of selection by neighborhood cluster and to make the results more generalizable to the 2003 Chicago population.[2] Additionally, we exclude 16 participants who are missing data on length of residence.

References

House, J.S., et al. (2011). Chicago Community Adult Health Study, 2001–2003. ICPSR31142.v1. From http://dx.doi.org/10.3886/ICPSR31142.v1.
Israel, B.A., 1982. Social networks and health status: linking theory, research, and practice. Patient Counselling and Health Education 4 (2), 65–79.
James, S.A., 1993. Racial and ethnic differences in infant mortality and low-birth weight: a psychosocial critique. Annals of Epidemiology 3 (2), 130–136.
Kasarda, J.D., Janowitz, M., 1974. Community attachment in mass society. American Sociological Review 39 (3), 328–339.
Keene, D., Geronimus, A., 2011. "Weathering" HOPE VI: the importance of evaluating the population health impact of public housing demolition and displacement. Journal of Urban Health 88 (3), 417–435.

VII. CONCLUSION

Formal data citation is a key element of the growing data-sharing infrastructure, not only facilitating sharing, discovery, and proper use, but also enabling data impact tracking that allows researchers to receive credit for their contributions. Specialized data repositories, such as ICPSR, integrate data citations within study metadata to enhance access and encourage data sharing. National and international efforts are underway to encourage adoption of these types of practices. The eventual result should be that more data creators will benefit from citations by receiving credit for their work. More researchers will benefit by readily finding reproducible research. And more funding agencies will benefit by tracking supported projects' usage and gauging impact beyond the initial funding.

ICPSR and its topical archives, like NACJD, provide an example of how data citation can encourage data archiving and secondary use. They support the growth of the ICPSR Bibliography of Data-related Literature and see the collection as evidence of new scientific findings for consideration in shaping public policy. The Bibliography's two-way linkages between data and data-associated publications have improved the discovery and the chances of good secondary use of ICPSR data. Due to inconsistent and inadequate data-citing practices in the scholarly literature, tracking data reuse is costly and labor-intensive. Despite this, ICPSR continues to value and invest in the collection of data-related publications, while promoting the creation and use of standards for citing and sharing research data according to best practices.