A correction article relating to this paper has been published and can be found at http://datascience.codata.org/articles/10.5334/dsj-2021-021/

Data Science Journal, Volume 12, 13 September 2013

OUT OF CITE, OUT OF MIND:

THE CURRENT STATE OF PRACTICE, POLICY, AND TECHNOLOGY FOR THE CITATION OF DATA

CODATA-ICSTI Task Group on Data Citation Standards and Practices

Edited by Yvonne M. Socha

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/articles/10.5334/dsj-2021-021/

OUT OF CITE, OUT OF MIND: THE CURRENT STATE OF PRACTICE, POLICY, AND TECHNOLOGY FOR THE CITATION OF DATA

CODATA-ICSTI Task Group on Data Citation Standards and Practices

Edited by Yvonne M. Socha

PREFACE

The growth in the capacity of the research community to collect and distribute data presents huge opportunities. It is already transforming old methods of scientific research and permitting the creation of new ones. However, the exploitation of these opportunities depends upon more than computing power, storage, and network connectivity. Among the promises of our growing universe of online digital data are the ability to integrate data into new forms of scholarly publishing to allow peer-examination and review of conclusions or analysis of experimental and observational data and the ability for subsequent researchers to make new analyses of the same data, including their combination with other data sets and uses that may have been unanticipated by the original producer or collector.

The use of published digital data, like the use of digitally published literature, depends upon the ability to identify, authenticate, locate, access, and interpret them. Data citations provide necessary support for these functions, as well as other functions such as attribution of credit and establishment of provenance. References to data, however, present challenges not encountered in references to literature. For example, how can one specify a particular subset of data in the absence of familiar conventions such as page numbers or chapters? The traditions and good practices for maintaining the scholarly record by proper references to a work are well established and understood in regard to journal articles and other literature, but attributing credit by bibliographic references to data are not yet so broadly implemented.

Recognizing the needs for better data referencing and citation practices and investing effort to address those needs has come at different rates in different fields and disciplines. As competing conventions and practices emerge in separate communities, inconsistencies and incompatibilities can interfere with promoting the sharing and use of research data. In order to reconcile this problem, sharing experiences across communities may be necessary, or at least helpful, to achieving the full potential of published data.

Practical and consistent data citation standards and practices are thus important for providing the incentives, recognition, and rewards that foster scientific progress. New requirements from funding agencies to develop data management plans emphasize the need to develop standards and data citation practices.

The CODATA-ICSTI Task Group on Data Citation Standards and Practices was first organized in 2010 jointly by the international and interdisciplinary Committee on Data for Science and Technology (CODATA) and the International Council for Scientific and Technical Information (ICSTI). Both CODATA and ICSTI adhere to the International Council for Science (ICSU), a nongovernmental umbrella scientific organization headquartered in Paris, France. Additional information about all three groups is available at www.codata.org, www.icsti.org, and www.icsu.org, respectively.

Together with representatives from several other organizations, the CODATA-ICSTI Task Group examines a number of key issues related to data identification, attribution, citation, and linking. Additionally, the Task Group helps coordinate international activities in this area and promotes common practices and standards in the scientific community. This report is part of that focused effort.

To address these challenges, the Task Group's first major activity was to collaborate with the U.S. National Academy of Sciences' Board on Research Data and Information (BRDI) and the U.S. National Committee for CODATA on an international workshop held in August 2011, in Berkeley, California (http://sites.nationalacademies.org/PGA/brdi/PGA_064019). The workshop culminated in the report, *For Attribution—Developing Data Attribution and Citation Practices and Standards*, National Academies Press, 2012 (available openly and freely for download at http://www.nap.edu/catalog.php?record_id=13564).

Since the 2011 workshop, the Task Group has undertaken a series of activities designed to build upon the international body of knowledge on data citation and attribution practices. The report presented here represents the next step identified by the Task Group: to document the current state of practice for data citation and attribution, noting emerging trends, successes, and challenges.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

The principal methods the Task Group used in writing this paper included the following:

- Literature Search and Compilation of Bibliography

From its inception, the Task Group assembled a bibliography on the topic of data citation and attribution. This activity continued throughout the completion of this report. We drew upon references provided by speakers and participants at the workshop, conducted library and web searches, monitored listservs and blogs of organizations working in the field of data publication and related topics, and received many submissions from Task Group members and their colleagues. The resulting bibliography is posted online and explained in Appendix B of this report.

- Stakeholder Interviews

Recognizing that different stakeholder communities might have different interests or concerns regarding data citation and attribution practices, the Task Group identified those communities likely to have the greatest potential impact upon the development of citation and attribution practices: managers at data repositories and academic libraries, scholarly journals, research institutions, and research funding organizations. While individual researchers were also identified as an important stakeholder community, in the interest of efficiency, the Task Group chose to focus its primary attention upon the institutional stakeholders with whom individual researchers would necessarily interact. Members of the Task Group then conducted telephone interviews with representatives of those stakeholder communities in which the selected representatives were asked questions tailored specifically to each community. The interviews made no effort to achieve statistical validity but rather were designed to support the Task Group's effort to assess the progress of those communities in their efforts to recognize and address issues regarding data citation as well as their perceptions of its importance. The list of interviewees is presented in Appendix C.

- The Writing Process

Task Group members developed an outline based upon discussions conducted primarily by email and teleconference. The members then volunteered to focus on certain topics based upon their respective interests and expertise. In addition to monthly teleconferences, the writing teams met in person several times for drafting sessions to elaborate upon and refine the chapters. Two-day writing sessions were held in Copenhagen in June of 2012 and in Taipei in October of 2012.

The Task Group engaged the services of a technical writer with expertise in Library and Information Science who had previously worked on compiling the bibliography to refine the output of the various chapter teams into a document in a more consistent voice. The technical writer also met with several members of the writing team in November of 2012 to further refine the draft based upon inputs developed at the Taipei writing sessions. The Task Group continued to circulate drafts of the revised chapters to the writing teams working on other chapters and to all the Task Group members for internal review and comment.

Finally, the Task Group identified external peer reviewers with the appropriate expertise to critique the paper. The writing team then responded to reviewer comments and made appropriate revisions to the manuscript prior to publication. The Acknowledgement section that follows the body of the report contains the names of all the people who were involved in the production of this report, including the funders.

Keywords: Data citation, Data management, Data policy, Data publishing, Data centers, Data access, Reuse of data, Metadata, Digital preservation, Information standards, Information infrastructure, Information technologies, Internet, Libraries, Scientific organizations, STM publishers, Research funders, Information metrics

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Table of Contents

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Executive Summary

Traditionally, scientific findings have been shared by means of publication and citation, in which papers are published, read, and critiqued, with the links between papers established through a formal process of bibliographic referencing and citation. So well established is this practice that most journal *Instructions to Authors* provide the details of what information should be provided and how the references should be structured . A data citation is a reference to data for the purpose of credit attribution and facilitation of access to the data. As data sets have become larger and more complex, however, it is often no longer possible to publish them as part of a paper; the ability of scientific assertions in the paper to withstand scrutiny demands that the link between the data and the publication be maintained.

The relatively new practice of making bibliographic references to data sets with formal citations begins to address long-standing problems limiting our collective ability to locate data and to reuse them effectively in advancing science. References made and citations received support a research infrastructure to provide the necessary recognition and reward of data work, in addition to providing attribution detail, facilitating future access, and fostering cross-collaboration and investigation. They are the links between the data and the published research results needed to maintain the integrity of the scientific method.

Some research funders have begun to require that publicly funded research data be deposited with various data centers. As these practices become better established, the ability to detect, locate, obtain, and understand the data from prior research will be circumscribed by our ability to have a sufficient description of those data: a citation.

Based on a review of emerging practices and analysis of existing literature on citation practices, we have identified the following set of "first principles" for data citation:

1. Status of Data: Data citations should be accorded the same importance in the scholarly record as the citation of other objects.

2. Attribution: Citations should facilitate giving scholarly credit and legal attribution to all parties responsible for those data.

3. Persistence: Citations should be as durable as the cited objects.

4. Access: Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data.

5. Discovery: Citations should support the discovery of data and their documentation.

6. Provenance: Citations should facilitate the establishment of provenance of data.

7. Granularity: Citations should support the finest-grained description necessary to identify the data.

8. Verifiability: Citations should contain information sufficient to identify the data unambiguously.

9. Metadata Standards: Citations should employ widely accepted metadata standards.

10. Flexibility: 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 across communities.

These principles are offered as guides to implementers, not as an endorsement of any particular implementation.

Nonetheless, many disciplines, institutions, and countries have begun to develop and implement data citation practices and protocols. Several groups have been formed to study and promote various aspects of data citation, including the DataCite organization. While such bottom-up actions have been useful and encouraging, it is essential at this relatively early juncture that these approaches be coordinated as much as possible and that they learn from each other. As data-intensive research becomes broadly adopted, there is a concomitant need for good data management practices, including the implementation of a data citation infrastructure. It is thus the responsibility of the major stakeholders in the research enterprise—the research institutions, funders, data centers and libraries, publishers and scholarly societies, and the researchers themselves—to assess and assert their respective roles in data citation policy and practice.

Moreover, as technology evolves and becomes more sophisticated, it facilitates the generation, management, analysis, and dissemination of increasing amounts of research data. Because the continuity of scientific progress builds on prior knowledge, the challenge to technologists is to balance innovation and the connectedness of new and

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

historical research results. Standards such as protocols for data exchange, best practices for publishing, and tools for cataloging need to be implemented by the communities of interest in parallel with technological progress. The best methods of generating, managing, analyzing, and disseminating research data must come from within the scientific communities themselves. However, data production should be supported by a social infrastructure and cyberinfrastructure that integrates data production with publication, citation, attribution, and curation.

Adoption of good data citation practices can be expected to have a high benefit-to-cost ratio, both from a general, systemic standpoint and from the perspective of major stakeholder groups. The latter will vary according to discipline, time, and institution. Examples of successful data citation implementation approaches can help make the case to the different stakeholder communities.

While data citation offers broad potential benefits for domain and interdisciplinary research and an improved understanding of science, a robust and effective data citation infrastructure requires further research itself, not just coordination and resources. The research community needs to build broad and coherent theoretical models and to translate and apply them to enable evaluation of the semantic equivalence of data in diverse formats; to strengthen theoretical and practical applications of data provenance; and to enable standardized approaches to managing the legal rights and responsibilities that adhere to research data. Research also will be needed to assess the effectiveness and impact of the implementation of data citation and to guide development of the infrastructure.

For all these reasons, a broadly implemented data citation infrastructure, coupled with wider and more uniform data access, would stimulate and support the progress of science.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Chapter 1 THE IMPORTANCE OF DATA CITATION TO THE RESEARCH ENTERPRISE

1.1 Introduction

In the last decade, the amount of data created by large scientific facilities, sensors, new observation instruments, and supercomputing has outpaced our ability to process, store, and analyze the data. As technological factors, such as faster processors, better storage, and increased bandwidth, have enabled the much greater production and capture of data, the creation of standards to manage these data has not kept pace. Nor are data management issues solely limited to the data produced by high performance computing (HPC) and scientific computing (SC); in fact, the aggregated data produced by individual researchers or small research groups may well dwarf that created by HPC or SC. By its nature, these "long tail" data are hard to find, standardize, and account for but still deserve proper data management.

While some organizations have recognized the need for policies regarding data management (Helly, 1998) and have implemented policies that begin to address the issues (Helly, Elvins, Sutton, & Martinez, 1999; Helly et al., 2002; Staudigel et al., 2003), there is still a lack of overall consensus regarding the treatment of data, especially in ways to reference data sets and maintain the scholarly record.

This report focuses on the process of data reference itself—the citation of data. It is important to note that in bibliometrics parlance, references are made and citations are received.[1]

Traditionally, scientific findings are shared by a mechanism of publication and citation, in which papers are published, read, and critiqued while the links between papers are made through a formal process of citation. So well established is this practice that well-defined *Instructions to Authors* provide the details of what information should be provided and how the citations should be structured for almost every journal. As data sets have become larger and more complex, however, it is often no longer possible to publish them as part of a paper; however, the ability of scientific assertions in the paper to withstand scrutiny demands that the link between the data and the publication be maintained.

Data and citations to them are critical to verify research conclusions and enable reuse of data. The relatively new practice of the citation to data sets begins to address long-standing problems limiting our collective ability to locate data and use them effectively in advancing science. Citations support a research infrastructure to provide the necessary recognition and reward of data work, in addition to providing attribution detail, facilitating future access, and fostering cross-collaboration and investigation (Berns, Bond, & Manning, 1996; Committee on the Preservation of Geoscience Data & Collections & Committee on Earth Resources, 2002; Committee on Responsibilities of Authorship in the Biological Sciences, 2003; Feinberg, Martin, & Straff, 1985; Seiber, 1991; Towne, Wise, & Winters, 2004; Uhlir & Schröder, 2007).[2] It is for these and many other reasons that the more pervasive use of data citation is beneficial to the scientific community.

When data are captured as part of the publication, in the form of a graph, table, or image, for example, they are cited as a part of that article. Data sets have been included to some degree as supplementary materials related to published

---

[1] Another view (which is not in conflict with the above) is that a citation is the performative act linking two entities, which can be expressed as an RDF triple: _:EntityA cito:cites _:EntityB. Such a citation is manifested in the citing document in two ways: as an in-text reference pointer (e.g., "Jones et al., 2009") in the body text of a document and as a bibliographic reference (e.g., "Jones A, Bloggs R and Smith L (2009) Title. Journal Name 45:213-224.") in the reference list of the document.

This nomenclature and the relationship between terms are defined in BiRO, the Bibliographic Reference Ontology (http://purl.org/spar/biro/), and the C4O (Citation Counting and Context Characterization) Ontology (http://purl.org/spar/c4o/).

[2] Note that though matters such as integrity and provenance are relevant for determining if data are suitable for use in new research, these cannot be determined from the citation alone as they cannot be embedded in the citation itself. Instead, discovery of provenance and integrity information can be facilitated by citation through providing links to other relevant information.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

manuscripts. The data themselves reside on servers maintained by the publishers, accessible only to those with subscriptions.

More recently, some government funding agencies have required that publicly funded research data be deposited with national data centers. As these practices become better established, the ability to detect, locate, obtain, and understand the data from prior research becomes limited by our ability to have a sufficient description of those data: a citation. This is the link between the data and the published research results needed to maintain the integrity of the scientific method. Chapter 1 thus introduces the need for implementation of data citation, management, and sharing.

Data are produced in many settings, ranging from large-scale facilities with hundreds of staff and dedicated high-performance computing infrastructure, down to single researchers working by themselves on desktop PCs. According to Anna Gold (2007, Section 1.1), research infrastructure now takes the form of ". . . high performance computing (HPC) centers – located at most research universities; supercomputing capabilities at national laboratories; and a small number of distributed supercomputing nodes that provides supercomputing capabilities on demand to scientists . . ."

This use of HPC centers is true for only a small minority of researchers – a large and growing number of others working with "long tail" data primarily use the Internet (specifically the Web and email) as their research infrastructure. Data repositories on the web are becoming increasingly important, both for the archiving and dissemination of research data while email, wikis, and other web-based tools provide quick and convenient methods for sharing information and discussions. Many other scientific domains are exploring the potential of cloud environments for their data needs.

The capacity to collect and analyze multi-source data is transforming domains such as biology and physics and is now starting to change the social and health sciences as well (Lazer et al., 2009). However, a concerted effort to manage, share, and cite data is needed to ensure that these new resources are available to the public, to scientists working in the academic sphere, and to individuals and communities who can benefit from such data (World Economic Forum, 2012).

So-called "big data" (or at least data created by large organizations and hosted in specialized data centers) tend to have standardized processes for sharing and assigning credit to data producers because they are typically produced on an industrial scale, with many users (often all within the large project team) wishing to access them. For an example, see the Earth System Grid Federation portal for accessing climate model data and metadata resulting from the 5th Climate Model Intercomparison Project (CMIP5) (http://pcmdi9.llnl.gov/esgf-web-fe/). Note, however, that the "big" in "big data" is not exclusively about massive quantities, but it is also about complexity and diversity of sources. By contrast, the resources needed for data sharing and citation may not be available to the lone researcher. For all data producers, big and small, many issues are the same: the desire to receive credit for data production and to ensure the reuse and reproducibility of the data.

Besides scientific data themselves, there are many other modes of increased data production, such as mobile phone position data, crowd-sourced data entry, supermarket loyalty card databases, and other non-traditional forms of potential digital research content. These new forms of facts can be valuable beyond their intended purpose and represent an additional opportunity for data mining and analysis, such as faster disease outbreak tracking, mapping population trends, food production and distribution statistics, and improved studies of consumer knowledge and education.

Social media play a unique role as well. Many platforms, such as Twitter, Facebook, and YouTube, are becoming new modes of scholarly networking, communicating, and sharing of information that can easily cross research fields. Researchers and others members of different scientific communities, such as the Citizen Science Alliance (http://www.citizensciencealliance.org/), are beginning to cite data, providing views of new scientific content and promoting access to a new channel for sharing. Conversely, these new platforms add challenges to the creation of best practices in comprehensive digital data citation because citations of information published via social media tools become part of a dynamic communication process that may reflect on the impact of an article or data set. Data sharing is increasingly acknowledged as an important part of research (Borgman, 2012a).

1.2 The role of data in the research lifecycle

Data are essential products resulting from and useful to basic scientific tenets, such as reproducibility and transparency. Data should be labeled in ways that allow them to be reused; the new mode of data intensive science makes the use of existing data a central asset of future science (Hey, Tansley, & Tolle, 2009). Data have always

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

been the cornerstone of science as it is not possible to replicate experimental findings, perform observational research, or test assertions without them. Because data often have a longer lifecycle than the research projects that create them, understanding the role of data in the research lifecycle is vital. However, it is important to note that research lifecycles are as varied as the types of research performed so generalizations are not always helpful.

Data are integral to the modern practice of research. When data are not included as part of the dissemination of scientific research, the link from the published research results back to them is broken, and the provenance and trustworthiness of the results may be in question.

In a research cycle, the researcher creates various versions of data sets, which often are recorded in the same database or repository. The data set is therefore a composite object. The identifying descriptors of that object must include enough specificity about its constituent parts so that the citation can refer to one and only one, unambiguous, clearly defined data set. This requires versioning of records and identification of entities that have contributed to or changed them, such as original data author and interpreted data author. However, this is not simply a problem of assigning identifiers or metadata. For the purposes of aggregation, computation, verification, reproducibility, and replicability, the data set must be defined so that it can be referenced in a way that yields a concrete search result (Wynholds, 2011).

Currently, data sets exist mostly in the pre-publication phase and are only infrequently part of the scholarly dissemination process for research publications. Furthermore, data associated with manuscripts in the current nodes of dissemination are usually in highly reduced forms, such as tables or graphs. In order to fully realize the potential offered though data mining, discovery, collaboration, and reuse, data citation is a necessary endeavor though it is important to remember that data can and will be cited without their being made open to all users. Much work is being done by data centers and research funders to encourage researchers to include data curation and data management into their research practice, including identifying community needs for supportive facilities, such as data repository services, metadata services, data discovery and data mining services, and data preservation services.

1.3 Organization of this report

This report summarizes the current state of the art of data citation in research worldwide. In Chapter 2, we examine data citation terminology and concepts as they apply to general, archival, and infrastructural terms, and we discuss the distinctions among literature-to-data, data-to-literature, and data-to-data citations.

In Chapter 3, we discuss data citation principles that are a mix of those distilled from existing best practices, several of which were identified at a 2011 "Data Citations Principles Workshop" (http://projects.iq.harvard.edu/datacitation_workshop/) on data citation as well as those that resulted from an analysis of the existing literature. We present a set of ten principles that are intended to guide the development of precise instructions on "what to do" for those who wish to publish their data to the highest standards of scientific best practice. We do not, however, present an actual structure for data citation.

Chapter 4 describes the data citation institutional infrastructure and includes the key players and status of work that is being done. We survey current policy and operational experience, discuss the importance of good data management to different stakeholder communities, and provide examples of operational work currently being undertaken in data citation.

In Chapter 5, we examine standards for metadata used within data citation, which include the selection and structure of metadata elements as well as the context and determination of their placement and uses in the scientific record. We also include a discussion of incomplete practices and gaps regarding metadata and data citation and a discussion of the current and emerging tools and technology.

The methodological and technical aspects of developing data citation protocols are not technologically difficult. The major challenges are largely in overcoming socio-cultural, institutional, and economic barriers to the broad uptake of those protocols. Before surveying the nature of these barriers and suggesting ways to overcome them, Chapter 6 begins with a discussion of the benefits of surmounting these challenges. In Chapter 7, we discuss what research is needed to guide a maximally effective data citation system. We then discuss the research needed for effective data citation practices. Finally, we discuss the types of new metrics and domain research that data citation aims to enable.

The appendices at the end of the report include: (A) a list of the CODATA-ICSTI Task Group on Data Citation Standards and Practices members and links to their professional biographies, (B) a Task Group bibliography of data citation publications, and (C) organizations interviewed by Task Group members concerning data citation practices.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Chapter 2 DEFINING THE CONCEPTS AND CHARACTERISTICS OF DATA

2.1 Definitions

It is important to define the terms used in considering the data citation process and infrastructure. We therefore have compiled a number of key terms and their definitions from several authoritative sources, most notably from the Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest (1999). These terms may have other meanings in other contexts.

2.1.1 Terms for data objects

Establishing what does and does not constitute data and data sets is a contentious issue and varies from one research domain to another (Borgman, 2012a). This report is concerned primarily with digital data although a large portion of raw data was recorded historically as analog data, which also can be digitized.

> The term "data" as used in this document is meant to be broadly inclusive. In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, and simulations), ***digital data*** refers as well to forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socio-economic data; and other forms of data either generated or compiled by humans or machines (Uhlir & Cohen, 2011,[3] as reported in Borgman, 2012a, p. 1061).
>
> A ***data set*** is a collection of related data and information—generally numeric, word oriented, sound, and/or image—organized to permit search and retrieval or processing and reorganizing. Many data sets are resources from which specific data points, facts, or textual information is extracted for use in building a derivative data set or data product. A *derivative data set*, also called a *value-added* or *transformative* data set, is built from one or more preexisting data set(s) and frequently includes extractions from multiple data sets as well as original data (Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, 1999, p. 15).
>
> A ***database*** is a well-defined structure into which data can be placed, stored, and accessed.
>
> A ***data set producer*** acquires data in raw, reduced, or otherwise processed form—either directly, through experimentation or observation, or indirectly, from one or more organizations or preexisting data sets—for inclusion in a data set that the data set producer is generating. Such data set creators—sometimes known as data set publishers or originators but for the purpose of this report referred to as data set producers—traditionally are the *rights holders* of the intellectual property rights in the data sets (Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, 1999, p.15).

Metadata: According to Greenberg (2005, p. 20), metadata "addresses data attributes that describe, provide context, indicate the quality, or document other object (or data) characteristics." Metadata are often classified by their purpose, including *descriptive* metadata, *administrative* metadata, and *structural* metadata as the most common subclassifications. Rights management (terms and conditions), provenance, and preservation metadata are most often subcategorized under administrative metadata; however, some taxonomies promote these to first-class categories (Greenberg, 2005; National Information Standards Organization [NISO], 2004). Metadata may also contain detailed methodologies or protocols (for example, current practice in the ecological observation field), and as such, these cannot be viewed as administrative.

Data paper: A data paper is a method for collecting and publishing metadata about the data in a format that is easily human-readable. These metadata may consist of the reasons why the data was collected or created, how it was collected, and any other information the data producer thinks is important to record. Data journals exist in some domains, such as the Earth Systems Science Data Journal (http://www.earth-system-science-data.net/) and Geoscience Data Journal (http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060), to publish data papers and provide services, such as peer review of the paper and the underlying data set. There is room for the format to

[3] Uhlir, P. F. & Cohen, D. (2011, March 18). Internal document. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

increase in complexity with the incorporation of other valuable elements, both general purpose and discipline specific, to enrich discovery, reuse, and archiving (Kunze et al., 2011). A data paper is not a traditional journal paper as it allows the publication of the data set without the requirement for novelty or significant analysis or conclusions to be drawn. It is a way to provide discoverability and quality assurance mechanisms for data that may be of use or interest to others while at the same time providing credit for the researchers involved in creating the data set. Note that although data papers enable attribution and credit, they are not the only way of doing so.

A data paper minimally consists of a cover sheet and a set of links to archived artifacts. The cover sheet contains familiar elements, such as title, author(s), date, abstract, and persistent identifier—at a minimum, enough to permit basic exposure to and discovery of data by Internet search engines; also just enough to build a basic data citation, to instill confidence in the identifier's stability, and to be picked up by indexing services (Kunze et al., 2011). This will, in many instances, be the same as a human-readable rendering of the main metadata elements that map to the concept of a "citation."

Data integrated in publications: This occurs when data are integrated into the article via data viewers, interactive PDF files, and the like. Integrated data can be useful when the article readers wish to more closely examine the data used to create specific graphs or figures – "the data behind the graph." Note that the majority of publications at this time are text-only publications and publications without interactive viewings.

Enhanced publication: A technical term that refers to aggregations of literature and data building a compound, distributed object.

2.1.2 Terms for data preservation

Long-lived data: Data citation also implies the need for archiving and preservation so that others may enable and verify conclusions, provide attribution and future access, and foster cross-collaboration and investigation. Because data citations require long-term preservation, we adopt the definition of "long-lived" or "long-term" that is provided in the Open Archival Information System (OAIS) standards of the Consultative Committee for Space Data Systems (CCSDS):

> A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future (Consultative Committee for Space Data, 2002).

Collections: Digital collections are limited to those that can be accessed electronically, for example, via the Internet. An increasing number of organizations recognize ". . . the growing importance of these digital data collections for research and education, their potential for broadening participation in research at all levels, . . . and the rapid multiplication of collections with a potential for decades of curation" (National Science Board, 2005, p.4). Collections are sets of digital objects that may or may not reside within a database management system.

2.1.3 Terms for citation and metadata

Data Citation: The practice of providing a citation to data in the same way researchers routinely include a bibliographic reference to published resources (Australian National Data Service [ANDS], 2011). In traditional print publishing, a "bibliographic citation" refers to a formal structured reference to another scholarly published or unpublished work. Typically, in-text citation pointers to these structured references are either marked off with parentheses or brackets, such as "(Author, Year)," or indicated by superscript or square-bracketed numerals although in some research domains footnotes are used. Such citations of a single work may occur several times throughout the text. The full bibliographic reference to the work appears in the bibliography or reference list, often following the end of the main text, and is called a "reference" or "bibliographic reference." Traditional print citations include "pinpointing" information, typically in the form of a page range that identifies which part of the cited work is being referenced (Van Leunen, 1992).

The terminology commonly used for digital citation has come to differ from this older print usage. Possibly this variance is a result of the greater freedom available to digital work to separate presentation from meaning and thus to present the same information in multiple places. We adopt the more current usage in which "citation" is used to refer to the full bibliographic reference information for the object.

The current usage leaves open the issue of the terminology used to describe the more granular references to data, including subsets of observations, variables, or other components and subsets of a larger data set. These granular

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

references are often necessary in-text to describe the precise evidential support for a data table, figure, or analysis and are analogous to the "pin citation" used in the legal profession or the "page reference" used in citing journal articles. Altman and King (2007) and Buneman (2006) suggest the term "***deep citation***" be applied to more granular citation.

Relationship between metadata and citation: Citations and metadata are interdependent. Traditionally, scholarly citations are used for credit attribution, verification, and location of cited articles, books, or web pages. Each discipline has its own specific style, such as Chicago, The American Psychological Association (APA), or the Modern Language Association (MLA), though it is important to note that the style is a bibliographic convention and not the citation itself. A data citation is a reference to data for the purpose of credit attribution and facilitation of access to the data. Citations generally embed a limited number of metadata elements, such as a persistent identifier, descriptive title, and fixity information (for provenance verification). The data objects described by the citation are generally discoverable by this citation metadata. In addition, the data object is generally associated with richer metadata than is available in the citation alone. Thus, the citation is useful, indirectly, for accessing larger sets of metadata.

The differences between provenance metadata and descriptive metadata: While metadata describe an information resource, *provenance* metadata are considered a form of administrative metadata and are distinguished from metadata that describe the intellectual content. For example, the title, author, and abstract for a work are considered descriptive. On the other hand, information describing file fixity, such as a Universal Numeric Fingerprint (UNF) or cryptographic hash, or describing version and change history, such as the creation date, would be provenance metadata because they refer to the chain of control and modification of the object (Altman, 2008). UNFs may be more readily accepted in some disciplines than others, according to the disciplinary practice and requirements. Metadata about the creator of an object may be either descriptive, if the creator is also the author to whom the work is attributed, or provenance-related, if the creator is part of the formal chain of control.

Structural metadata: Data objects, unlike textual publications, are rarely self-describing. To understand the semantics of a data object generally requires additional information describing its format, structure, and other technical characteristics. Because this is a property of the current representation of the data object, and not of the intellectual work it represents, this information is generally categorized as "structural" or "technical" metadata. It is generally not embedded in the citation but kept as additional information in close proximity to the data object in repository and delivery systems.

Persistent Identifier: A unique and persistent web-compatible alphanumeric code that points to a resource (e.g., data set) that will be preserved for the long-term (i.e., over several hardware and software generations). The author or publisher should direct the user to the latest available version of the resource or to the metadata that enables acquisition of the desired version or format. Furthermore, a data citation needs an identifier that is both unique and persistent. Persistent identifiers are categorized as descriptive metadata. However, their role is so critical that they may be put in their own category.

It is possible to conceive of persistent identifiers that are not web-compatible but still useful – in some contexts. For example, one could create an icon or other symbol to represent an object (e.g., the symbol invented by the artist formerly known as Prince), but such an icon would be less useful for the purposes of citation we discuss. We argue that web compatibility is an essential part of our definition of a persistent identifier because it should allow for a widely available tool (such as a browser on a PC or other device) to dereference the identifier.

Attribution stacking: This issue arises when data are licensed for reuse on the proviso that the authors of the original data set are acknowledged in any derivative works created using the original data set. This can result in situations where all the authors of all the data sets used to create the derived data set must be acknowledged and attributed, notwithstanding the fact that the originating data sets may be many times removed from the derived data set.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Chapter 3 EMERGING PRINCIPLES FOR DATA CITATION

3.1 Introduction

Data citation principles have evolved considerably since early efforts in the 1990s to preserve digital scientific data. Increased awareness of the need for improved practices and methods for data publications has spurred such efforts as the standardization of the digital object identifier (DOI) syntax in 2000 through the National Information Standards Organization (NISO) and its subsequent approval as an ISO standard in 2010.[4] Many individual and collective efforts have led to the current recognition of the need for data citation standards and procedures as well as for something equivalent to "instructions to authors" guidelines used by publications to ensure standardized manuscript and citation preparation.

The primary purpose of citation has been to support an argument with evidence, though over the years it has also become a mechanism for attribution, discovery, quality assurance, and provenance. Altman and King (2007) state that data citations should contain information sufficient to locate the data set that is referred to as evidence for the claims made in the citing publication, to verify that the data set the user has located is semantically equivalent to the data set used by the original authors of the citing publication, and to correctly attribute the data set.

3.2 Principles

Based on observation of emerging practices and analysis in existing literature on citation practices, we have identified a set of ten "first principles" for data citation: status, attribution, persistence, access, discovery, provenance, granularity, verifiability, standards, and flexibility.

1. The Status Principle:

Data citations should be accorded the same importance in the scholarly record as the citation of other objects.

The scholarly record comprises the relationships among scholarly works and evidence. Traditional citations most often signify the relationship between statements in the article and statements in other published works to which the former are related.

The relationship between a statement in the article and the supporting data is neither less important nor representative of a different intellectual relation than the traditional citation of other published works. Thus, data citations should be treated in a similar manner as citations to other works.

Citations to data should be presented along with citations to other works, typically in a references or "literature cited" section (Altman, 2012; Callaghan et al., 2012). The former states the principle that data citations should be first-class objects for bibliographic purposes, and the latter made the argument that data themselves should be first-class objects for scientific purposes. In other words, publishers should not impose additional requirements for citing data, nor should they accept citations to data that do not meet the core requirements for citing other works (Altman, 2012).

2. The Attribution Principle:

Citations should facilitate giving scholarly credit and legal attribution to all parties responsible for those data.

Credit is the universally recognized currency for both quality and quantity of scientific activity. Citing data allows the use of metrics to evaluate use and impact factor of data sets, potentially encouraging data creators to make their data available for use by others. This fosters transparency and enables recognition of scholarly effort.

Current legal attribution requirements do not necessarily match expectations for receiving credit, nor do they perfectly map to accepted standards of citation. To avoid the problems of attribution stacking and the need to inject legal text into research publications, licenses should be used that recognize citations as legal attribution, such as the forthcoming version 4.0 of the Creative Commons licenses. Where additional information is required to satisfy legal

[4] It is important, however, not to conflate the DOI, which is a persistent ID and *part* of the citation, with the citation itself. DOIs do not by themselves address provenance, fixity, nor granularity, among other things.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

attribution, it should, if at all possible, be obtainable via the persistent identifier that is included in the citation, as per principle three.

Citations may function to provide credit directly through the text of the citation or indirectly through well-known and reliable linkages to other resources that provide more detailed credit and attribution. It is not necessary to embed all contributors within the citation itself.

Tracking each contributor's data and providing credit to the contributor's effort is a necessary task, similar to what is needed for a research manuscript. If many contributors contributed to a work, it may be more practical to list them elsewhere; this is acceptable as long as there is a reliable and systematic (ideally programmatic) way of looking up the full list of contributors based on the citation. The full list of contributors may be indexed in the metadata associated with the DOI, for example, rather than embedded in the citation. See Section 5.2.3 for an example of indirect mechanisms. In other words, data citation must support unambiguous assignment of credit to all contributors, possibly through the citation ecosystem.

3. The Persistence Principle:

Citations should be as durable as the cited objects.

The citations themselves have to be persistent and must be linked in some manner to the identity of the curator currently responsible for the referent data set (whether the curator is an institution or a named individual). Based on the findings of the 2011 "Data Citations Principles Workshop" (http://projects.iq.harvard.edu/datacitation_workshop/), citations should persist and enable resolution to the version of the data cited, at least as long as the cited work persists. If possible, data publishers should use some form of persistent resource identifier, such as DOIs, persistent URLs (PURL), or others (Altman & King, 2007; Green, 2009; Lawrence et al., 2007; Star & Gastl, 2011). It is strongly recommended that these persistent identifiers be in URL form or directly resolvable to URLs.

A persistent digital identifier enables unambiguous referencing, cross-referencing, authentication, and validation. It also provides a basis for practices such as citation counting in career merit reviews. Furthermore, as noted in the provenance and attribution principles, registries or related services are needed to guarantee the persistence of metadata associated with the citation, such as fixity information or the full list of contributors.

A central resolver and registry require maintenance and must be sufficiently robust. For example, resolver databases could be mirrored and updated in multiple locations around the world in a timely manner, given our current Internet architecture. To ensure reliability and availability, there should be redundant access points without single-point failures. Additional discussion of persistent identifiers can be found in Chapter 5.

4. The Access Principle:

Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data.

Access to data citations facilitates attribution, verification, reuse, and collaboration. Unlike literature, data are not generally self-describing, and meaningful access to data therefore requires access to metadata that describe the data sufficiently to permit accurate interpretation and use.

5. The Discovery Principle:

Citations should support the discovery of data and their documentation.

Discovery of data enabled by citations makes it easier to validate and build upon previous work. Facilitation of discovery through data citation ensures proper attribution and provides information about related methodology that allows the data to be put into context.

Data should include metadata that aid in their discovery. Part of such metadata should be the citations of all publications related to the data. In effect, this means bi-directional linking from data sets to publications and from publications to data sets (Borgman, 2007).

While data are discoverable only if they are archived, any embargo period specified for the data can influence their discoverability. Consideration will also vary by the discipline and type of data, such as raw data, observations, models, samples, and the like (Borgman, 2012a). Descriptive titles and content tagging in the metadata can also be aides to discovery.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

6. The Provenance Principle:

Citations should facilitate the establishment of provenance of data.

Citations should include sufficient fixity and other *administrative* provenance metadata to verify that a data object later accessed is equivalent to the data object originally cited. Provenance information related to data processing, including, for example, data versions and fixity checks, is used for a variety of purposes, such as understanding the process and auditing as well as for reproducibility (Tilmes, Yesha, & Halem, 2011). Accurate provenance facilitates chain-of-custody, accurate data, comparison data, and reproducibility.

Additionally, *methodological* provenance is challenging to include in the citation itself, but it is often required for a full understanding of the data: Many data sets are derived from measurements, simulations, or other data sets in ways that are not obvious from reading the paper. The metadata associated with a data set should include sufficient methodological provenance to evaluate the data set and to link it to other data sets. For derived or aggregated data, such metadata will generally include citations to the parent or aggregated data sets.

7. The Granularity Principle:

Citations should support the finest-grained description necessary to identify the data.

Where a more finely grained data subset is used to support a specific evidentiary claim (such as statement, figure, or analysis), that subset should be described in the text reference, preferably as part of a structured "deep citation," or as an unstructured note, if necessary. Note that this rule does not require that separate persistent identifiers, such as DOIs, are created for fine grained citation, merely that there be some way of unambiguously identifying the portion used within the data cited.

A data set may form part of a collection and be made up of several files, with each containing several tables and many data points. Abstract subsets, such as features and parameters, may also be used. It is not always obvious what constitutes a whole data set (Ball & Duke, 2012). For authors, the pragmatic solution is to list data sets at the finest level of granularity provided by the repository or institution and to clearly identify the subset of the data that underlies each figure and analysis. This is analogous to using the author, title, publisher, and date of publication to refer to a book but including page numbers in the in-text reference.

Providing explicit granularity minimizes additional information needed to find the source. The optimum level and nature of granularity, however, would vary with the kind of data. Further discussion on granularity can be found in Chapter 5.

8. The Verifiability Principle:

Citations should contain information sufficient to identify the data unambiguously.

Citations are used to associate published claims with the evidence supporting them and to verify that the evidence has not been altered. Thus, there must be sufficient fixity information embedded in or associated with a citation to verify that the data used match data originally cited.

For scientific literature, persistent identifiers such as DOIs have often resolved to a landing page containing at least bibliographic metadata. This practice arose partly because publishers had different views on how much of their content they were actually prepared to expose to a DOI lookup and also because there is reasonable homogeneity of type across the scientific literature. Data sets are more heterogeneous in terms of size, file type, and intended purpose.

Resolution of persistent identifiers consistent with the verifiability principle may be achieved by bringing the user to a landing page (data surrogate) where additional metadata, such as the data form, provenance, fixity, and content, are given explicitly. It is also permissible to link directly to a data set, provided that the combination of citation and data set together contains adequate metadata to verify that it is what is being sought, including fixity information in the citation, as described in Chapter 5. If possible, open protocols, such as Content Negotiation (http://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html) and Resource Description Framework in Attributes (RDFa) (http://www.w3.org/TR/xhtml-rdfa-primer/), should be used so that computational agents are given direct access to the metadata and data in machine-readable form when resolving identifiers, such as DOIs, rather than forcing a detour to a human-readable HTML page.

9. The Metadata Standards Principle:

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Citations should employ widely accepted metadata standards.

Some scientific disciplines have invested decades of effort in developing community metadata standards. Implementing metadata standards alleviates both redundancy and ambiguity that can diminish interoperability. Additionally, standards also provide legal clarity and certainty concerning the permitted uses of the data.

Unlike narrative publications, data are designed to be processed directly. Standardized metadata for data publications, as discussed in Chapter 5, would explicitly enhance current and future automation for the data ingestion, authentication, and quality control needed for reliable reuse of the data, both in verification of published results and repurposing of data.

It is important that interoperability requirements distinguish semantics from presentation and provide the necessary and sufficient information about the data independently of any presentation layer, format, or style. New metadata standards continue to be developed, and some will be appropriate for data citation. Notwithstanding, metadata used in citation must be open, platform-independent, and well-recognized by the community. There must always be a means of exporting and importing metadata across systems that are used within and between disciplines.

10. The Flexibility Principle:

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 across communities.

Widespread adoption of data citation practices requires that these practices serve the needs of a broad range of constituencies. The nature of data used in different disciplines varies in size, complexity, useful level of granularity, and other characteristics. This principle addresses the commonly stated maxim that one size does not fit all.

While broad acceptance requires that citation practices accommodate the needs of a diverse array of disciplines and communities of practice, there should be a minimum baseline set of elements upon which different disciplines and communities can build in order to meet their specific needs and to facilitate interoperability.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Chapter 4 THE EXISTING INSTITUTIONAL INFRASTRUCTURE FOR DATA CITATION

4.1 Introduction

Despite constraints upon the research expenditure of some countries since the financial crisis of 2008, there has been a sustained increase in the amount of research being undertaken, and the Internet lies at the core of an advanced scholarly information infrastructure to distribute data and promote information-intensive collaborative research (Finch, 2012, pp. 22-23; Borgman, 2007, preface). In *Cyberinfrastructure Vision for 21st Century Discovery* (2007), the National Science Foundation Cyberinfrastructure Council writes, "Worldwide, scientists and engineers are producing, accessing, analyzing, integrating, and storing terabytes of digital data daily through experimentation, observation and simulation. . . .The enormous growth in the availability and utility of scientific data is increasing scholarly research productivity, accelerating the transformation of research outcomes into products and services, and enhancing the effectiveness of learning across the spectrum of human endeavor" (p. 21). As an "infrastructure based upon distributed computer, information, and communication technology" (Atkins Report of the National Science Foundation, 2003, p.5), cyberinfrastructure enables a virtually integrated environment of mixed information and research resources that serves specific projects and research communities. It has a wide range of uses in both the public and private sectors that include, but are not limited to, technology transfer, inter-sector collaboration, public education, and commercial innovation.

Yet this technological promise, absent structured coordination, can also quickly inhibit scientific progress: Incompatible formats, tools, or structures can isolate scientific disciplines and communities; inadequate or antiquated preservation or curation processes can put good data in practical obscurity; investment in obsolete infrastructures can impose prohibitive costs; and latecomers in the technological adoption lifecycle may be effectively excluded from participation (Atkins Report, 2003, pp. 10-12; Wenger, White, and Smith, 2010). As Borgman (2007, preface) succinctly observes, "Every stage in the lifecycle of a research project now can be facilitated—or complicated—by information technologies".

As scientific research has become more data-intensive in the past few decades, there has been a growing perception by the policy community that the access and reuse or repurposing of data can promote more efficient and effective research and greatly improve the public returns of investment (European Commission, 2011a). Many researchers advocate the establishment of infrastructure for the systematic curation of data sets, and many have also contributed to a better understanding of infrastructural mechanisms, such as data policy, stewardship, provenance tracking, permanent identifiers, metadata, and citation protocols. Although the ability to cite data in some standard way is a building block of this infrastructure, data citation has yet to be broadly systematized or institutionally adopted, as discussed in Chapter 1. Complex challenges, both social and technical, have continued to impede the process (Wynholds, 2011).

Fortunately, research funders are increasingly putting pressure on institutions to establish this infrastructure, and new policies are being created as a result. For example, the United Kingdom's Engineering and Physical Sciences Research Council's (EPSRC) *Policy Framework on Research Data* puts the onus on funded research institutions to have data management policies and processes in place (2011).

Moreover, there is growing international and national support for open access systems. In a press release following the launch of the European Commission's Open Data Strategy for Europe, the EC Vice President Neelie Kroes stated, "[public sector and government] data is worth more if you give it away" (European Commission, 2011a). As a result, data will no longer only be used by a privileged segment of researchers; in an open access environment, they will finally be made useful for funding agencies, patent services, and the curious masses. On the national level, The Royal Society, the independent, scientific academy of the United Kingdom dedicated to promoting excellence in science, issued a report on *Science as an Open Enterprise* that proclaims the values of openness and transparency but cautions that any resulting infrastructure must meet necessary standards of intelligibility and accessibility (Boulton et al., 2012).

Adopting good practices regarding the sharing of data is important in the context of collaborative research and open access. Moreover, as part of this infrastructure, new collaborative research platforms will require active and professional management of the processes by which data are generated, organized, evaluated, and disseminated. It is

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

important to consider the legal and policy framework for data sharing, such as intellectual property, confidentiality, and privacy, as well as a range of other legislation that defines rights in research data (EPSRC, 2011).

The rest of this chapter describes some of the issues important to data citation infrastructure that are being addressed by different sectors and institutions in the research community. It then identifies specific examples of data citation approaches that have been developed by specific scientific data or discipline communities from the bottom-up. These communities not only highlight the recognized importance of data citation practices at the working level but also prove that such models can be effectively implemented.

4.2 International organizations with a role in data citation

National and international science policy is primarily developed at the governmental and intergovernmental levels through a variety of institutional channels—some well-established and others newly emerging. These policies are largely coordinating mechanisms that orient the rules and priorities governing various aspects of the research enterprise and manage the short and long-term budget allocations, which then are made available in the public sector. Currently, it is largely a top-down and institutionalized set of systems for decision-making, with any changes implemented slowly and incrementally, usually after they have been developed and tested at the working level by the client communities from the bottom-up. These priorities, however, must be coordinated with the academic and industrial research organizations within each country.

Established organizations that can play a role in the formation and implementation of scientific data citation include intergovernmental science policy and data management bodies, such as the Organization for Economic Co-ordination and Development (OECD) (http://www.oecd.org/), the United Nations Educational, Scientific, and Cultural Organization (UNESCO) (http://www.unesco.org/new/en/), and other specialized organizations of the United Nations. Discipline or sector specific entities include the Group on Earth Observations (GEO) (http://www.earthobservations.org/index.shtml), the Inter-governmental Oceanographic Data Exchange (IODE) (http://www.iode.org/), the Global Biodiversity Information Facility (GBIF) (http://www.gbif.org/), and many others. Intergovernmental coordination mechanisms for research policy, practice, and funding include the Commission of the European Communities (http://ec.europa.eu/index_en.htm) and the Belmont Forum (http://igfagcr.org/index.php/belmont-forum). Undoubtedly, national funders and policy makers for scientific research are also key organizational players in the formulation of policy that directly impacts data management and data citation practices.

At the international nongovernmental level, scientific cooperation organizations that can play an important role are DataCite (http://datacite.org/), the International Council for Science (ICSU) (http://www.icsu.org/), and groups that are under the ICSU umbrella, most significantly the World Data System (WDS) (http://www.icsu-wds.org/), the Committee on Data for Science and Technology (CODATA) (http://www.codata.org/), the International Council for Scientific and Technical Information (ICSTI) (http://www.icsti.org/), and the ISU Scientific Unions and the emerging Research Data Alliance (http://rd-alliance.org/), which is government supported but implemented through scientific communities. Data citation infrastructure resides within the Internet infrastructure, and the key standard setting body for the Internet is the W3C (http://www.w3.org/). The International Standards Organization (ISO) (http://www.iso.org/iso/home.html), is an international and national standard setting body that affects the production, dissemination, and use of scientific data while the National Information Standards Organization (NISO) (http://www.niso.org/home/) performs this role specifically in the United States. In Japan, the National Institute of Information and Communications Technology (NICT) (http://www.nict.go.jp/en/) conducts ICT research on the leverage of data citation for archiving, mining, and searching. NICT additionally hosts the International Program Office of the ICSU World Data System (WDS).

Also at the nongovernmental level are publisher organizations, such as the International Association of Scientific, Technical, and Medical Publishers (STM) Association (http://www.stm-assoc.org), the Open Access Scientific Publishers Association (OASPA) (http://oaspa.org/), and CrossRef (http://www.crossref.org/). There is also ORCID (http://about.orcid.org/), a newly emerging not-for-profit group that is working on the standardization of author names.

Lastly, parallel institutions and organizations exist at the national levels as well—intergovernmental and nongovernmental research policy and funding entities, data centers, publishers and publisher groups, and data and information technical and standards bodies—and these can and should play a role in the data citation infrastructure issues.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

4.3 Importance of good data management practice to research and the scholarly record

Data citation is a critical part of good data management; for it to become part of the research "business as usual," it must be effectively integrated into the processes of scholarly communication, research design, funding, tenure, and promotion. Several constituents from government, academic, scholarly publishing, and research funding communities are coming to realize that there is a critical need for a better way to manage data, and they are actively taking steps to recognize and implement good data management practices to research. Indeed, for some of these organizations, there is already specific recognition of the importance of data citation. We realize that this is a very active area, and what follow are selected examples germane to data citation policies and practices.

The sampling of data policy and research funding agencies in some countries through interviews conducted by this Data Citation Task Group has suggested that there is growing awareness of the importance of data citation practices in such larger governmental organizations. Several of them have indicated that they are now forming data citation policies for their agencies or ministries to apply both internally and, for their research grantees, externally (Uhlir, 2012). Documented descriptions of the specific activities related to data citation are included in the attached bibliography (Appendix B). Further information may be found there about some of the organizations and programs highlighted in the following sections. It also contains references to organizations that have published work in this area but are not specifically highlighted below.

4.3.1 International scientific organizations

In its 2004 report, *Scientific Data and Information: A Report of the CSPR Assessment Panel*, the International Council for Science (ICSU) states: "Scientific data and information management can no longer be viewed as a task for untrained amateurs or as part of the routine 'clean up' conducted hurriedly by scientists at the completion of a research project" (p. 9). It remains a responsibility of all scientists and should be valued accordingly, but it is also an increasingly important professional activity, one that is essential to the scientific enterprise. Data management should be considered a professional activity, and capacity should be built to train researchers and managers in this process and to provide career paths for data management professionals. This statement brings the socio-cultural dimension (discussed in Chapter 6) to the infrastructure needed for an adequate data management and publishing process in which data citation plays a key role for the recognition of the producers of any given data sets.

Similarly, in 2007, the Organization for Economic Co-Operation and Economic Development (OECD) issued *Principles and Guidelines for Access to Research Data from Public Funding*. This document adopted the basic principle that publicly funded research data are a public good that should be made as accessible as possible. These principles have contributed to an increased awareness of the need for sharing data, and the OECD encourages its member states to implement the recommended set of principles at a national level.

In December 2011(a), the European Commission launched an Open Data Strategy for Europe, which includes plans for an update of the 2003 Directive on the reuse of public sector information to:

- Make it a general rule that all documents made accessible by public sector bodies can be reused for any purpose, commercial or noncommercial, unless protected by copyright.
- Establish the principle that public bodies should not be allowed to charge more than the costs of the individual request for data (marginal costs). In practice, this means most data will be offered online at no cost, unless duly justified.
- Make it compulsory to provide data in commonly used, machine-readable formats, to ensure data can be effectively reused.
- Introduce regulatory oversight to enforce these principles.
- Significantly expand the reach of the Directive to include libraries, museums, and archives for the first time because the existing 2003 rules will apply to data from such institutions.

Finally, a Royal Society Report (Boulton et al., 2012, p. 8) emphatically argued for six major changes to both the hard and soft underpinnings in policies for data management and for elements of a data cyberinfrastructure:

1. A shift away from a research culture where data are viewed as a private preserve.
2. Expanding the criteria used to evaluate research to give credit for useful data communication and novel ways of collaborating.
3. The development of common standards for communicating data.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

4. Mandating intelligent openness for data relevant to published scientific papers.
5. Strengthening the cohort of data scientists needed to manage and support the use of digital data, which will also be crucial to the success of private sector data analysis and the government's Open Data strategy.
6. The development and use of new software tools to automate and simplify the creation and exploitation of data sets.

4.3.2 Researchers and research institutions

Many researchers and research organizations have been adopting best practices and legal frameworks to implement responsible data management and beneficial sharing of research outputs. It is difficult to document good practices at the individual researcher level because the documentation typically originates from institutions and data centers, both of which represent the researchers with whom they work. It is the hope and intent that promoting the data citation issue will elicit more involvement from the individual researcher as well as researchers in the data citation community of interest. Some examples of the work from institutions in United Kingdom include the University of Edinburgh's "Research Data Management Policy" (2011), the University of Oxford "Policy on the Management of Research Records" (2012), and the University of Oxford's Research Data Management web page (http://www.admin.ox.ac.uk/rdm/). These practices have included the comprehensive development of protocols for organizing, preserving, and enabling access to the reuse of research data. Furthermore, academic research institutions also depend on reliable records of scholarly accomplishments for key decisions about hiring, promotion, and tenure. Such institutions must demonstrate to their funders that the research funds have been spent wisely.

Research institutions are also increasingly recognizing the need to set high standards for research integrity of the published record and the adherence to rules, professional norms, and standards. For example, in the Netherlands, recent cases of fraud concerning fabricated and manipulated research data have provided an impetus to introduce better data management practices introduced at all levels (Carey, 2011). Data citation and attribution validate the integrity of modern research by facilitating the reusability of published results as well enabling researchers to trace the provenance of the data sets. The nascent field of citation metrics can also be useful in documenting the impact of data citation on research integrity. Nonetheless, metrics need time to be developed and collected, and work on them is still in early stages. This includes tools that must still be further developed, such as the Data Citation Index discussed in Section 5.3.2.

4.3.3 Publishers of scholarly journals

Over the past two decades in a world where authors can communicate directly with their readers, ". . . intermediaries—publishers, aggregators, abstract and indexing services, and libraries—have had continually to re-assess and redefine their roles" (Finch, 2012, p. 28). Publishers serving different scientists and domains are in an ideal position to share knowledge and ensure best practices among different fields. As a result, they have shifted investment into meeting ongoing needs for content quality assurance, digital preservation, and search optimization (Finch, 2012).

In recent years, the idea of including data sets within the scientific record has gained significant momentum among publishers. Indeed, the number of open access data repositories has continued to rise since they first began in the 1970s and then drew increased attention with information technology advances beginning in the 1990s, especially within the fields of medicine and the life sciences (Finch, 2012). Many publishers (see for example, "*Guide to publication policies of the Nature Journals*," 2013) encourage their authors to submit their data sets to community-endorsed databases and repositories with the aim to include the database accession numbers or DOIs in the manuscript. The Elsevier publishing company has even added actionable data viewers (http://www.elsevier.com/about/content-innovation/matlab) that enable online readers to consult relevant data directly within the context of the article. Other publishers, such as BioMed Central (http://www.biomedcentral.com/), advocate open access publishing to increase transparency in scientific research and scholarly communication. Furthermore, BioMed Central has facilitated the development of the Panton Principles (http://pantonprinciples.org/), which are a set of recommendations that address how to best make published data available for reuse.

In addition, publishers are recognizing that data are first-class research assets for publication with increased investment in information flow systems, searchable content, text mining, semantic publishing, and linked data collections (Finch, 2012). As forward-looking projects such as Elsevier's "Article of the Future" (http://www.articleofthefuture.com/) develop, the ability to cite data will be increasingly important. The Thomson Reuters Data Citation Index (http://wokinfo.com/products_tools/multidisciplinary/dci/) is another example of how

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

publishers are finding opportunities in the data deluge, and citations will have an important role in managing the data sets.

The push to establish infrastructure and make research data available in conjunction with journal publications has also been analyzed and critiqued in various journals. In February 2011, *Science* devoted a special online issue (http://www.sciencemag.org/site/special/data/) to dealing with data and the data deluge. In February 2012, the editors of *Nature Biotechnology* agreed to allow the first DOIs for data to be cited in the journal (Peng et al., 2012). Rather than conceiving of data and journals as separate entities, one can consider the article and its supporting data sets as one increasingly integrated product.

Thus, to ensure that data sets are seamlessly integrated into the published record of science, it is important that links between publications and data sets in repositories are bi-directional (Borgman, 2007; Bourne, 2005). The publications should cite related data sets properly, but also the data sets should provide citations of all publications related to the data set. Efficient collaborations between data repositories and the main journals in their area show how this can work; see for example Pangaea (http://www.pangaea.de/), Dryad (http://datadryad.org/), the Cambridge Crystallographic Data Centre (http://www.ccdc.cam.ac.uk/pages/Home.aspx), and others.

An interesting initiative from the Research Object for Scholarly Communication Community Group (ROSC) (http://www.w3.org/community/rosc/) aims to facilitate a community data model to encapsulate research data and methods with all the contextual information essential for interpreting and reusing them. The group believes conventional digital publications are inadequate for scientists to access, share, communicate, and enable the reuse of scientific outputs. They assert that research investigations are increasingly collaborative and require 'borrowing strength' from the outputs of other research. The group brings together interested parties to provide a platform for discussion of current development of various container models and their implementations.

In a recent feasibility study implemented by the Center for Research Communications, the Journal Research Data Policy Bank (JoRD) (http://crc.nottingham.ac.uk/projects/jord.php) project is examining the scope and shape of a sustainable service to collate and summarize journal policies on research data. The aim of this future service will be to provide researchers, managers of research data, and other stakeholders with an easy reference to help them understand and comply with differing research data policies. A summary of the project as of February 2013 ("A rather long post") explains that nearly half the journals examined thus far have no data sharing policy (para. 6), and that the lack of consolidated infrastructure for the sharing of data is a considerable obstacle to making data openly accessible (para. 8). The next phase of the project involves implementing a database of sharing policies, engagement with stakeholders, and third party application programming interface (API) development with the intention to build use to the level at which a second phase, a self-sustaining model, would then be possible (para. 9).

4.3.4 Academic research libraries

The academic research library community, too, is increasing its interest and investment in scientific data repositories, archiving, curation, and related services. Some national libraries offer DOI registration services for persistent identification of research. For example, in 2009, the British Library, in cooperation with the German National Library of Science and Technology (TIB) and others, signed a Memorandum of Understanding to improve access to research data on the Internet using DOIs. Establishing such a memorandum fosters global cooperation among scientists and within a non-profit agency that supports registering research data sets (Brase, 2009).

Scientific data repositories are used by scientists to support aspects of their research, product development, and educational activities. In some cases, journal editorial policies require that authors who submit manuscripts must also upload supporting experimental data to specific data repositories relevant to the subject of the manuscript (see for example, "Availability of supporting data," 2013). Data sets are then assigned accession numbers that become part of the final journal publication. This direct connection between research article publication and data deposited in repositories creates opportunities for science librarians to engage in providing guidance and support in locating, selecting, and using repositories for data preservation. Data citation indexes will help science librarians to monitor data repository growth trends (Kirlew, 2011).

Additionally, librarians may play an important role in assisting scientists with managing the transition to data citation management. Currently, few style guides are available that address data citation and attribution. Furthermore, librarians can act as data citation advocates by providing instruction and support in the management of data sets as required for publication. One such project, Databib (http://databib.org/), is currently underway at the libraries of Purdue University and The Pennsylvania State University. Databib is essentially a registry of research data repositories, maintained by librarians, with the focus of connecting data creators, researchers, and librarians. It helps

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

users find data and data producers locate appropriate repositories and funding agencies that mandate data management (Witt, 2012).

4.3.5 Research funding agencies

Funding agencies play a crucial role in providing data management policies and mandates. In general, they, as the supporters of public research, want to see value for the money that they allocate to research projects. Making data citations a common practice helps promote cutting edge interdisciplinary research, which in turn advances careers of researchers and makes their contributions to the public good more visible (Spengler, 2012).

A data management plan is a tool that has been shown to be useful in large data acquisition projects. In developing a plan, the researchers have to take into account how to handle the data, in what format, and under what user conditions. The plan is written before the project begins, with the long-term goal of ensuring well-managed data. Funding organizations can make the data management plan a mandatory deliverable. Services to facilitate the creation of data management plans tailored to the requirements of specific funding agencies have been developed by the Digital Curation Centre (https://dmponline.dcc.ac.uk/) in the United Kingdom and by the California Digital Library (https://dmp.cdlib.org/) and collaborators in the United States.

In the United States, a number of major federal agencies now require all grant awardees to include a data management plan with their grant proposals. Since 2006, the National Institutes of Health (NIH) data sharing policy (http://grants.nih.gov/grants/policy/data_sharing/), for example, requires scientists submitting applications for research grants of $500,000 or more to include a data management plan and cite the source of the data upon which the research and resulting project are based (http://grants.nih.gov/grants/policy/nihgps_2012/index.htm). As part of that policy, NIH recommends the use of archives and data repositories for the protection of grantee data sets. It is important to note that NIH has the necessary infrastructure to enable and support data citation, yet several of their programs fail to mention citation practices in their guidance documents. Incorporating data citation practices with current data management plans is a crucial step needed to further data citation in health sciences in the United States.

In 2011, the National Science Foundation (NSF) made the submission of data management plans a requirement for all proposals, and in 2012, a single word was substituted in grant applications that went from citing “publications” to citing “products” (as cited in Piwowar, 2013). This one small change now allows data set curation and citation to be a first-class object in proposal reviews. The National Oceanic and Atmospheric Administration (NOAA), Institute of Museum and Library Services (IMLS), National Endowment for the Humanities (NEH), and the National Aeronautics and Space Administration (NASA) also have data sharing and management policies for grantees. In addition, the Gordon and Betty Moore Foundation, and Wellcome Trust, both notable science funders, have adopted data management plan requirements for their grantees.

In the United Kingdom, the Research Councils (RCUK) issued in 2011 the “Common Principles on Data Policy,” which states (para. 1):

> Making research data available to users is a core part of the Research Councils’ remit and is undertaken in a variety of ways. We are committed to transparency and to a coherent approach across the research base. These RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy and include a number of tenets:
>
> - Publicly funded research is a public good that should be made openly available with as few restrictions as possible.
> - Institutional and project specific data management policies and plans should be in accordance with relevant standards and community practice.
> - Metadata should be recorded and made available.
> - Public funds can appropriately be used to support the management and sharing of publicly funded research data.
> - Published results should always include information on how to access the supporting data.

Furthermore, the European Commission (2001b) makes its own data public through a new "data portal." This portal does not explicitly refer to the data produced by academic research though the EU Framework Program 7 (FP7) OpenAIREplus project (“Paving the way,” 2011) endeavors to cross-link the data outputs of research work with the associated papers derived from them. The aims of the project are to enable researchers to deposit their funded research results into open access repositories.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

4.4 Existing data citation practices of operational organizations and disciplines

4.4.1 Introduction

The following sections describe the bottom-up activities in operational organizations with regard to data management, data publishing, and data citations. Numerous individual centers in many disciplines are working on practical problems in data citation. We have selected exemplary organizations that have either published on their activities or have presented on and are active participants in the data citation conversation. These organizations are grouped by overall disciplinary areas. Organizations that focus on a broad range of disciplines are discussed in Section 4.4.6.

It should be noted that this is a rapidly evolving field and new organizations and activities continued to be announced as the Task Group continued its work; this report constitutes a slice in time of the data citation landscape as of early 2013.

4.4.2 Earth sciences

The Publishing Network for Geoscientific and Environmental Data (PANGAEA)

In existence for nearly twenty years, PANGAEA (http://www.pangaea.de/) is an open access library and data publisher for the earth and environmental sciences. The library and repository is hosted by the Alfred Wegner Institute for Polar and Marine Research (AWI) and the World Data Center for Marine Environmental Sciences (WDC-MARE) in Germany. Although PANGAEA accepts data from all disciplines, it specializes in data from earth system research and is the designated archive for the Earth System Science Data (ESSD) journal (http://earth-system-science-data.net/). All data sets are registered though the German National Library (TIB) and identified using a DOI. A sample citation is provided below.

> Giesecke, T et al. (2013): Validated sample ages from 823 EPD sites. doi:10.1594/PANGAEA.804597, *Supplement to:* Giesecke, Thomas; Davis, Basil AS; Brewer, Simon; Finsinger, Walter; Wolters, Steffen; Blaauw, Maarten; de Beaulieu, Jacques-Louis; Binney, Heather; Fyfe, Ralph M; Gaillard-Lehmdahl, Marie-José; Gil-Romera, Graciela; van der Knaap, Pim Willem O; Kunes, Petr; Kühl, Norbert; van Leeuwen, Jaqueline F N; Leydet, Michelle; Lotter, André F; Ortu, Elena; Semmler, Malte; Bradshaw, Richard H W(2013): Towards mapping the late Quaternary vegetation change of Europe. *Vegetation History and Archaeobotany*, 12 pp, doi:10.1007/s00334-012-0390-y

Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC)

The ORNL DAAC (http://daac.ornl.gov/) for biochemical dynamics is one part of the NASA's Earth Observing Systems Data and Information System (EOSDIS) managed by the Earth Science Data and Information System (ESDIS) project. The DAAC is operated by the Environmental Science Division and hosts data produced by NASA's Terrestrial Ecology Program. These data sets are produced by projects related to field campaigns, land validation, regional and global studies, as well as models.

The Center is assigning DOIs to their collection of published data sets because they are widely used by publishers and they contain key information needed to assist with reducing the barriers to adopting a data citation method (Wilson, 2012, p. 147).

Data principles at the ORNL DAAC require that data sets be cited and include metrics to track the times that data sets are downloaded. This information is used to help determine how the DAAC is providing added value to the scientific community of practice. An example of the data citation is as follows:

> Gu J.J., E.A. Smith, and H.J. Cooper. 2006. LBA-ECO CD-07 GOES-8 L4 Gridded Surface Ration and Rain Rate for Amazonia: 1999. Data Set. Available on-line [http://www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/831

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Federation of Earth Science Information Partners (ESIP)

The ESIP Federation (http://www.esipfed.org/) is an open networked community of more than 100 data centers in the United States. It includes organizations such as the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), the Environmental Protection Agency (EPA), the National Science Foundation (NSF), the United States Geological Survey (USGS), the Department of Energy (DOE), and others. Data stewardship and principles of the Federation include "reducing barriers between data providers and data users through IT, training, and standards education and promoting the use of technical standards and best practices for data management, stewardship, and application development" (Federation of Earth Science Information Partners, 2011, para. 2). Current recommendations for citing data encompass the guidelines followed by the International Polar Year, DataCite, DataVerse, and the Digital Curation Center. ESIP suggests core elements of author, date, title, version, archive, identifier, and access date although they allow for additional elements as appropriate. A sample citation is provided below.

> Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. Parsons and M. J. Brodzik. National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://dx.doi.org/10.5060/D4MW2F23z.

Natural Environment Research Council's (NERC) Environment Data Centers

In the United Kingdom, many of the data collected by NERC (http://www.nerc.ac.uk/) provide a unique and irreplaceable record of the environment. NERC's network of data centers provides support and guidance in data management to those funded by NERC, is responsible for the long-term curation of data, and provides access to NERC's data holdings in subject areas including atmospheric science, Earth sciences, Earth observation, marine science, polar science, science-based archaeology, terrestrial and freshwater science, hydrology, and bioinformatics.

NERC has the ability to issue DOIs to data sets held in its Environmental Data Centers, which is a result of collaboration between the NERC data centers, the British Library, and DataCite. NERC recommends the standard DataCite format for data citation and is currently working with academic publishers, such as Wiley and Elsevier, to improve the linking between data sets mentioned in research papers and the data center where the data set is archived. An example citation is provided below.

> Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan SA, Waight J, Walden CJ, Agnew J, Ventouras S]. 2009b. GBS 20.7 GHz slant path radio propagation measurements, Chilbolton site. NERC British Atmospheric Data Centre. doi:10.5285/639A3714-BC74-46A6-9026-64931F355E07.

International Coastal Atlas Network (ICAN)

ICAN (http://ican.science.oregonstate.edu/) is a group of organizations that work to implement data interoperability for coastal web atlases (CWAs) and is an International Oceanographic Data and Information Exchange (IODE) project under the International Oceanographic Commission (IOC). The strategic aim of ICAN is to develop common solutions to CWA development (e.g., user and developer guides; handbooks and articles on best practices; information on standards and web services; expertise and technical support directories; and other opportunities for education, outreach, and funding) while ensuring maximum relevance and benefit for end users. The long-term view is for global-level operational interoperability, which will evolve as the ICAN community strives to increase awareness of the opportunities that exist for increased coastal and marine data sharing among policy makers and resource managers as strategic users of a CWA. ICAN participants are leaders in forging international collaborations of value to the participating nations thereby optimizing regional governance in coastal zone management. A major goal is to help build a functioning digital atlas of the worldwide coast based on the principle of shared distributed information. This is done by organizing a cooperative interoperability network for the integration of locally maintained CWAs as the premier source of spatial information about coastal zones throughout the world.

4.4.3 Life sciences

The Global Biodiversity Information Facility (GBIF)

Established in 2001 with the goal of providing open access to biodiversity data, GBIF (http://www.gbif.org/) is an international government initiative with 52 countries as members and 47 international organizations (Chavan, 2012a). GBIF specializes in two types of data: species occurrence records and names and classification of organisms.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Because of the nature of biodiversity data, classification schemes are required and essential for searching occurrence data. GBIF utilizes authoritative classifications, such as the Integrated Taxonomic Information System (ITIS) Catalog of Life (http://www.itis.gov/), the International Plant Names Index (http://www.ipni.org/), and others. Two types of citation styles are recommended: publisher-based citation and query-based citations. The former is for immediate uptake by publishers and data owners, custodians, and aggregators while the latter requires the implementation of a data citation service (Chavan, 2012b). Examples of both are provided below.

Publisher-based complete formulation:[5]

> Chavan, V. S. (1996). Amphibians of the west coast of India. 1223records, published online, http://www.vishwaschavan.in/indfauna/amphibians_west_coast/,released on 12 June 1998, doi: 10.5284/1000164.

Publisher-based short formulation:

> Chavan, V. S. (1996), doi: 10.5284/1000164.

Query-based complete formulation (User-driven citation for subset from single dataset):

> http://www.ncbi.org.in/indfauna/ (2012), Hornbills and India, 989 records, accessed on 12January 2012: 22:10:10 hrs, user doi: 99.6672/100.324.2012, publisher doi: doi: 10.3897/ncbi.ncl.2001.

Query-based short formulation:

> http://www.ncbi.org.in/indfauna/ (2012), doi:10.3897/ncbi.ncl.2001.

Dryad

Dryad (http://datadryad.org/) is an international data repository of peer reviewed scholarly literature that specializes in biosciences data. Dryad cooperates with other specialized repositories, such as GenBank and TreeBASE, to provide a repository for data that does not have a discipline-specific repository. The repository is governed by several stakeholders from partner journals as well as other organizations from the publishing, funding, and scientific community.

Once an author's work is approved or under review for publication, the supporting data may be submitted in multiple formats to Dryad. After publication and verification by a Dyrad curator, the data files receive persistent, resolvable DOIs for use in citation, and the data are released to the public under the CC0 Creative Commons Public Domain Dedication. An example follows:

> Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC (2012) Data from: More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters* doi:10.5061/dryad.75nv22qj

4.4.4 Physical sciences

The Cambridge Structural Database (CSD) and the Protein Data Bank (PDB)

The discipline of crystallography has a number of long-established curated databases of crystal structure data sets, each containing three-dimensional atomic coordinates, data about inter- and intra-molecular geometry, atomic displacement parameters, and the like. In most cases, these structures are associated with literature publications, but there are significant numbers of unpublished structures. Among the most important of these databases are the Cambridge Structural Database (CSD) (http://www.ccdc.cam.ac.uk/products/csd/), which holds over 600,000 organic and metal-organic structures, and the Protein Data Bank (PDB) (http://www.wwpdb.org/), which holds over 80,000 proteins and nucleic acid structures. In both cases, each structure is identified by a unique reference code with a distinctive format. Each structural record carries a literature citation of the original publication where there is one.

The use of a PDB identifier (e.g., 4FDY) or CSD "refcode" (e.g., ABACOF) in a particular context (i.e., within a structure report article) enables location of all metadata associated with a structure by entering the reference code

[5] Note that there are a total of six styles for publisher-based citations and two styles for hypothetical based citations. For more description on all styles, see Chavan (2012b).

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

alone. In the editorial process for crystallography journals published by the International Union of Crystallography (IUCr) (http://www.iucr.org/), the appearance of such a reference code in a published article prompts a check for a corresponding reference to the literature describing that structure. Furthermore, every occurrence of a PDB code in the online edition is hyperlinked to its entry in the PDB. The link brings the user to a landing page, with a summary of relevant metadata described in or extracted from the structural data set. At present, such hyperlinking is done only for the PDB, which is an open-access resource. Associated with each structure in the PDB is a DOI, registered through CrossRef. In current practice, the DOI resolves to the data set itself and not to an intermediate landing page. The CSD also plans to assign DOIs to individual structures.

For small-molecule structures published in IUCr journals, the journal archives the full structural data set as supplementary material. The data set is a superset of information that may be deposited in the CSD as well as the processed experimental data sets (in the form of diffraction structure factors or Rietveld profiles) that were used to solve and refine the crystal structures. Each such data set has its own DOI, and, in each case, the DOI links directly to the data file. Current practice is to cite the parent article as a conventional literature citation. The article also has a DOI, and in this case, the DOI resolves to the traditional landing page associated with publications, which typically includes bibliographic information, an abstract, links to supplementary data sets, visualizations and, in some cases, validation reports.

4.4.5 Social sciences

Dataverse

Hosted by Harvard's Institute for Quantitative Social Science (IQSS), the Dataverse Network (http://thedata.org/) is an open source repository for digital collections management, dissemination, exchange, and citation. Its infrastructure allows for administration activities through a web-based client that interacts with the host network and for collection owners to serve their data via their websites. Dataverse networks can be used to host metadata, text, images, and data from numerous disciplines, and networks and collections can interact with each other via open protocols such as Z39.50 (Altman, 2012). Users simply fill out a form with the required metadata (based on the data citation principles at IQSS), and each dataverse includes an automatic handle and Universal Numeric Fingerprint (UNF). The repository provides backups and replications of data, reformatting for preservation, extraction of metadata, and inter-operability (King, 2007).[6] An example citation follows:

> Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]

The Inter-university Consortium for Political and Social Research (ICPSR)

Established in 1962, the Inter-university Consortium for Political and Social Research (ICPSR) (http://www.icpsr.umich.edu/icpsrweb/landing.jsp) is an international consortium of about 700 academic institutions and research organizations that maintains and provides access to social science data. According to its website, the ICPSR maintains a data archive of more than 500,000 files of research and hosts sixteen specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields. In addition to data access, the ICPSR also provides training in data management and curation. The Consortium recommends the data citation standards and best practices of its collaborative partner, the Data Preservation Alliance for the Social Science (Data-PASS) (http://www.data-pass.org).

Data-PASS currently includes ICPSR, the Institute of Quantitative Social Science, the National Archives and Records Administration, the Roper Center for Public Opinion Research, the Odum Institute, and the UCLA Social Science Data Archive. Data-PASS partners share a common cataloging and citation infrastructure, the Dataverse Network, for their union catalog and collaborate to replicate and safeguard the content partnership (Altman et al., 2009). Based on Data-PASS, the ICPSR recommends straightforward citations that include the elements of title, author, data, version, and persistent identifier. In addition to these elements, ICPSR also recommends the addition of fixity information, such as a checksum or Universal Numeric Fingerprint (UNF), which are definitive ways to establish provenance of the data. Examples are provided below.

[6] Note that the prior progenitor to the Dataverse Network, the Virtual Data Center system, supported an earlier data citation format using PURLs (Altman et al., 2001).

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Deschenes, Elizabeth Piper, Susan Turner, and Joan Petersilia. Intensive Community Supervision in Minnesota, 1990-1992: A Dual Experiment in Prison Diversion and Enhanced Supervised Release [Computer file]. ICPSR06849-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000. doi:10.3886/ICPSR06849

Esther Duflo; Rohini Pande, 2006, "Dams, Poverty, Public Goods and Malaria Incidence in India", http://hdl.handle.net/1902.1/IOJHHXOOLZ UNF:5:obNHHq1gtV400a4T+Xrp9g== Murray Research Archive [Distributor] V2 [Version]

Sidlauskas B (2007) Data from: Testing for unequal rates of morphological diversification in the absence of a detailed phylogeny: a case study from characiform fishes. Dryad Digital Repository. doi:10.5061/dryad.20

4.4.6 Other data citation practices

Journal Supplemental Materials Practices

Practices for dealing with supplementary materials are emerging. Because of the policy push to provide data management plans and access to this material, new models are being discussed, but the ecology of supplemental materials is in an early stage of development. Many scholarly journals routinely accept and distribute supplemental material, and these materials may include further details, or supporting information, even including data sets. Although this practice is common, it is far from standardized, and supplementary files suffer from being poorly peer reviewed, frequently lack unique identifiers, and are typically not discoverable except through reading the related article because they are rarely indexed by web search engines. Signs that the current practices are coming under strain because of increased submission of data are a significant challenge.

As an example of this strain, the *Journal of Neuroscience* announced in 2010 that it would no longer accept supplemental materials (Maunsell, 2010). This was the first unambiguous admission from a journal that it was no longer able to cope with all of the data submitted to them in supplementary article files. With the exception of multimedia, it no longer accepts any article supplements because they are too much of a burden on the refereeing system. "A supplemental material arms race" is noted by the Editor-in-Chief of this journal.

The announcement further highlights that much of what is submitted as supplementary material should be considered essential to the understanding of the article – yet this material practically disappears when published as supplements: "Supplemental material also undermines the concept of a self-contained research report by providing a place for critical material to get lost. Methods that are essential for replicating the experiments, analyses that are central to validating the results, and awkward observations are increasingly being relegated to supplemental material. Such material is not supplemental and belongs in the body of the article, but authors can be tempted (or, with some journals, encouraged) to place essential article components in the supplemental material" (Maunsell, 2010, para. 7).

There is hope expressed in this article that more reliable repositories will be established as journals become less and less a deposit place for data: "It is conceivable that removing supplemental material from articles might motivate more scientific communities to create repositories for specific types of structured data, which are vastly superior to supplemental material as a mechanism for disseminating data" (Maunsell, 2010, para. 11).

As a response to these strains, NISO initiated a process to create recommendations on best practices for supplemental materials (2013). NISO has since published multiple versions of these recommendations, and it is approaching final publication. In essence, the NISO recommendations make a distinction between Integral Content, Additional Content, and Related Content. The first category, integral content, is considered an essential part necessary for the understanding of the article, even if the contents are placed outside the article for technical, business, or logistical reasons. The second category, additional content, will help in gaining further understanding of the article. The third category, related content, will comprise data in external repositories.

With regard to this final category, the document states that some publishers refer to these data as similar to bibliographic citations while others do not. Publishers may have no formal curation responsibility over the data, but proper linking and citation should be considered important.

Thus, in contrast to the citation practices discussed in this chapter, the categorization adopted by the NISO group for supplementary materials prioritizes the locational/administrative properties of the additional content (where the related work is and who manages it) over its evidentiary relationship to the article. Data that are essential to the understanding of the article, but are not submitted to the journal as supplementary material, are not classified as

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

"integral", despite their strong evidentiary relationship to the article. Citation policies such as those described above are potentially valuable in improving the usability and usefulness of supplementary materials; they could function to maintain the connection between publications and related data that are integral to understanding of those publications but are administratively separated.

4.4.7 Other initiatives in the developing data infrastructure

The Data Hub

The Data Hub (http://datahub.io/) is an open source software data management system for publishing linked data, created by the Open Knowledge Foundation (http://okfn.org/). The aim of the Data Hub is to make institutional data open, available, and reusable on the web. The Data Hub has 4,586 data sets published thus far, and of those, 332 data sets are Linked Open Data (LOD).[7] Each entry of the LOD representing a fact has a URI and can be referenced and linked on the web. The high interconnectivity between entries potentially increases discoverability, reusability, and the utility of the information (Mendes, Jakob, Garcia-Silva, & Bizer, 2011).

Characteristics of data published on the Data Hub include:

- Government encouragement to publish their own data on the Data Hub.
- The data on the Data Hub can be assigned open data licenses, such as the Open Data Common Open Database License (ODbL).
- API providing data can also be published on the Data Hub.
- Data providers should provide metadata; a simple data citation includes a data set title, data set provider, and a retrieved time.

Examples of citations are provided below:

1. Taiwan Geographic Name:

http://thedatahub.org/en/dataset/edit/taiwan-geographic-names

2. Linked Open Data of Ecology, which is collected in a LOD cloud:

http://thedatahub.org/en/dataset/linked-open-data-of-ecology

Figshare

Based in London, Figshare (http://figshare.com/) is a commercial repository supported by Digital Science (http://www.digital-science.com/), which is owned by the same parent company as *Nature*. Figshare permits researchers to upload any file format and disseminate all of their outputs in a way that current scholarly publishing does not allow. It is unique in that it caters to all research domains and users can post null results. Allowing users to post negative results helps avoid the "file drawer effect" discussed in Chapter 7. Users may choose to keep their data private or public. Metrics are included for public uploads so users can see the impact of their research. All Figshare data made publicly available get allocated a DataCite DOI at the time of publication. In 2013, Figshare partnered with *PLOS* ("Figshare partners," 2013) and formed an agreement that will allow authors to host their data on Figshare. An example citation follows:

> SHILLELAgh: A data-driven solar wind model for studying solar energetic particle events. Paul Higgins, David Pérez-Suárez, Peter Gallagher.figshare. http://dx.doi.org/10.6084/m9.figshare.726152
>
> Retrieved 16:11, Jun 23, 2013 (GMT)

OpenAIRE

OpenAIREplus (https://www.openaire.eu/en/home), the 2nd Generation of Open Access Infrastructure for Research in Europe, is a 30-month project funded by the EC 7th Framework Program. It extends on the work done in OpenAIRE to facilitate access to the entire Open Access scientific production of the European Research Area,

[7] Resource Synchronization and Exchange is a related area of development that builds on LOD. See http://www.niso.org/workrooms/resourcesync/

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

providing cross-links from publications to data and funding schemes. This large-scale project brings together 41 pan-European partners, including three cross-disciplinary research communities. The project will capitalize on the successful efforts of the OpenAIRE project, which is rapidly moving from implementing the EU Open Access Pilot project into a service phase, enabling researchers to deposit their FP7 and ERA funded research publications into Open Access repositories. The current publication repository networks will be expanded to attract data providers from domain specific scientific areas. Innovative underlying technical structures will be deployed to support the management of and inter-linking between associated scientific data. OpenAIRE has adopted the DataCite metadata schema, with minor modifications, as the basis for harvesting and importing metadata about data sets from data archives. In particular, it is interested in data sets that are cited by articles already held in the OpenAIRE repository.

In May 2013, OpenAIRE and CERN released a data repository called ZENODO (http://www.zenodo.org/), which will accept publications and data from researchers. The main characteristics are provision of DOIs for data, long-term preservation techniques of the CERN digital library, and links of data to publications and projects that implement OpenAIRE guidelines.

The Scientific Data Center of the Computer Network Information Center, Chinese Academy of Sciences (SDCCAS)

The Scientific Data Center (SDC) (http://english.cnic.cas.cn/), a division of the Computer Network Information Center (CNIC) (http://english.cnic.cas.cn/) in the Chinese Academy of Sciences (CAS) (http://english.cas.cn/), is an information infrastructure supporting organization in charge of the construction and operation of the CAS digital application environment. SDC provides data storage and services, data application technical services, disaster recovery backup, and scientific data long-term preservation and sharing.

The SDC has taken the lead in implementing the CAS Scientific Databases Project (SDB) (http://www.csdb.cn/) for more than 26 years, with over 60 institutes joining the project to supply database services. The SDB holds a variety of scientific data, including geology, biology, energy, and others. As of 2013, more than 20 percent of the mass of data collected from different disciplines has been shared, with over 80 million online visits to the database and almost 680TB of downloaded data (http://msis.csdb.cn).

Moreover, a recent study from the National Knowledge Infrastructure (CNKI) (http://www.global.cnki.net/kns50/single_index.aspx), one of the largest sources of China-based information sources, analyzed cited frequency of the SDB from 2006-2011. Results show that the databases have been cited 1,082 times and suggest the SDB plays an important role in scientific and social research (Yanhua & Lianglin, 2012). In order to face future data handling capacity and improve data sharing and reuse efficiency, SDC data citation guidelines, funded by the Chinese Ministry of Technology, were published in 2012 (Wiehua, 2012). Specifications are based on comparisons of data citation set elements that are used by STD-DOI, DataCite, PANGAEA, the ICPS, the Roper Center, and the Dataverse network. The elements are listed as follows: author, title, version, publisher, publication year, distributor, distribution data, unique identifier, and parsed URL.

Citations without the version element are used as a substitute for concise expressions while the other eight elements are all required elements to compose a certain citation for data. An example is provided below:

> Institute of Geographic Sciences and Natural Resources Research. Metadata standardization for man-earth database (V2.0). Institute of Geographic Sciences and Natural Resources Research [Publisher], 2010. Computer Network Information Center [Distributor], 2011-01-19.csdb:cn. csdb. TR-REC-015-01; http://citation.csdb.cn/csdb:cn. csdb. datamirror.LE71230392011343EDC00

Compared with traditional literature citation methods, the CNIC data citation guideline utilizes three of the elements listed above: distributor, unique identifier, and parsed URL to cite data. The unique identifier is a unique character string to set up a permanent relationship between the citation and the real data. The parsed URL is included because most unique identifiers (except the DOI) are difficult for navigators to recognize and parse directly. The parsed URL makes cited data easier to access and use.

4.5 Conclusion

The emergence of the importance of data management in an increasingly data intensive world has led to significant activities that will help build the cyberinfrastructure to facilitate distributed data and information-intensive collaborative research. As we have seen though the many programs and initiatives in this chapter, the challenge of

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

building the infrastructure is being addressed in both a top-down approach through national and international policy and as a bottom-up community approach through many operating entities.

What is most interesting is that the level of attention to data management is increasing. Every day we hear of new initiatives. Data citation is increasingly recognized as a lynchpin in the evolving cyberinfrastructure. To date, although there has been growing policy attention to general data management and data planning, there has chiefly been only technical community activity regarding guidelines and standardization. As the ecology of data sets and data citation evolves, the research community will need to explore more systematically and in a more interdisciplinary manner the possibilities for more formal standards. That will be the task of the next few years for at present it is unfortunately the case that, with exceptions in a small number of disciplines, the majority of researchers remain unaware of many of the activities discussed in this chapter and do not have data management or data citation on their lists of priorities.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Chapter 5 THE TECHNICAL INFRASTRUCTURE

5.1 Common practices for data citation

The caveat that "one size does not fit all" is applicable to data citation, but there are elements underlying data citation for which disciplinary or Community of Practice (CoP) guidelines are emerging. Objectives include discovering, describing, sharing, and preserving the data and also making them interoperable. While some of the elements traditionally used in citations of literature can be transferred readily to the citations of data sets as part of a standard citation system, other elements still need to be formulated, implemented, and adapted to address the different needs of data citation.

Data citations derive from metadata elements, or components, that uniquely identify a data set and make it discoverable. These elements may also vary by CoP guidelines. Core elements of a traditional publication citation, such as author or contributor, title, source or publisher, and access date, are generally transferrable from long established norms. However, there are less obvious elements that are in need of review and consensus development.

In addition to standardizing a "citation" as the proxy for a data set, the content of some metadata elements that will comprise the citation should also be standardized. Elements to support the disambiguation of a data set, describe its granularity, and identify its version need to be considered in light of the unique properties of data sets. Initiatives to develop interdisciplinary standardized authority files for those elements are currently underway. Through such standardization, ambiguity regarding the data set can be reduced. Other metadata elements, such as those describing provenance, privacy controls, and reuse rights, are, to some degree, novel to data citation, and these should be standardized as well.

5.1.1 Elements of a data citation

In recent years, a growing number of repositories and publishers have developed good, consistent practices in data citation (Ball & Duke, 2012). The elements that would make up a complete citation are still a matter of some debate. The following list is taken from the Digital Curation Centre (http://www.dcc.ac.uk/) and combines tenets of several research papers (Altman & King, 2007; Green, 2009; Lawrence et al., 2007; Star & Gastl, 2011). See also the next section, which focuses on the citation structure rather than the metadata elements selected as core for a citation.

Author

- The creator of the data set.

Title

- As well as the name of the cited resource itself, this may also include the name of a facility and the titles of the top collection and main parent subcollection (if any) of which the data set is a part.

Publisher

- The organization (or repository) either hosting the data or performing quality assurance.

Publication date

- Whichever is later: the date the data set was made available, the date all quality assurance procedures were completed, or the date the embargo period (if applicable) expired. In other standards an "Access Date" field is used to document the date the data set was successfully accessed.

Resource type

- Examples: "database" or "data set."

Edition

- The level or stage of processing of the data, indicating how raw or refined the data set is.

Version

- A number increased when the data changes, such as the result of adding more data points or rerunning a derivation process.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Feature name and URI

- The name of an ISO 19101:2002 "feature" (e.g., GridSeries, ProfileSeries) and the URI identifying its standard definition, used to pick out a subset of the data.

Verifier

- Information to verify the identity of the content.

Identifier

- A resolvable web identifier for the data, according to a persistent scheme. There are several types of persistent identifiers, but the scheme that is gaining the most traction is the Digital Object Identifier (DOI).

Location

- A persistent URL or UNF from which the data set is available. Some identifier schemes provide these via an identifier resolver service.

The consensus seen in previously cited research indicates that among the required components that should be present in any citation are the creator or author, the title and date, and the location. These components, respectively, give due credit, allow the reader to judge the relevance of the data, and permit access to the data, respectively. In theory, these elements should uniquely identify the data set; in practice, a formal identifier is often needed. The most efficient solution is to give a location that consists of a resolver service and an identifier.

Finally, in addition to the above quantifiers, the way in which these elements would be styled and combined together in the finished citation depends on the style in use for citations within individual textual publications. The list below provides examples of the same data citations drawn from commonly used style manuals.

APA

- Cool, H. E. M., & Bell, M. (2011). Excavations at St Peter's Church, Barton-upon-Humber [Data set]. doi:10.5284/1000389.

Chicago (notes)

- 2. H. E. M. Cool and Mark Bell, Excavations at St Peter's Church, Barton-upon-Humber (accessed May 1, 2011), doi:10.5284/1000389.
- Cool, H. E. M., and Mark Bell. Excavations at St Peter's Church, Barton-upon-Humber (accessed May 1, 2011). doi:10.5284/1000389.

MLA

- Cool, H. E. M., and Mark Bell. "Excavations at St Peter's Church, Barton-upon-Humber." Archaeology Data Service, 2001. Web. 1 May 2011. <http://dx.doi.org/10.5284/1000389>.

Oxford

- Cool, H. E. M. & Bell, M. (2011) Excavations at St Peter's Church, Barton-upon-Humber [data-set]. York: Archaeology Data Service [distributor] <DOI 10.5284/1000389>.

5.1.2 Persistent resource identifiers in citations

The persistence of the links to information provided in a scholarly citation is a critical aspect of a good citation. In the digital sphere, information is much more dynamic and fragile than in the sphere of print publication. Digital data can be more easily altered or corrupted than information previously provided by paper and ink, and such changes may be less obvious. Additionally, the availability of instantaneous and near zero-cost distribution of electronic information means that it is more common to have subsequent "editions" or versions that correct or enhance the original edition. Fixity of location also poses a challenge as the demise of a single repository or the degradation of the bits stored there could mean complete loss of the information. Even the transfer of electronic data to a different single repository may render them impossible to find unless there is some persistent identifier and resolver system in place to provide a persistent location service. Furthermore, instability from the server side often causes hyperlinks and URLs to be unstable, and this can also interfere with persistence. The scholarly community has developed a number of systems to incorporate more permanence into the data distribution system.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Organizations are increasingly recognizing the importance of technology in properly citing data. Persistent identifiers (e.g., resolvable DOIs, URIs, and Handles) provide both human-and machine-readable means that can direct one to the data set of interest. Their benefit is the fact that they identify the data set, regardless of its physical location, thereby avoiding the common issue of changing or disappearing URLs. More importantly, for the purpose of citing data, persistent identifiers provide digital data citations with additional findability characteristics, making them easier to access and reuse the data. However, it is important to note that failure to maintain registries of persistent identifiers will cause the same instability problem that URLs present.

Although there is no current model for persistent identifiers that is universally used for digital citation, the one system that has received the most use is the DOI. Recently published as ISO Standard 26324:2012, the DOI is a unique alphanumeric string assigned by an authorized DOI registration agency to an object in a digital system. Objects identified with a DOI can be digital or analog, but in the case of data citations, we expect that the objects in question will be digital. Several registrants (repositories, consortia, and publishers), such as Dryad, PANGAEA, and BioMed Central,[8] issue DOIs for data sets. DataCite[9] and CrossRef are membership registration agencies within the International DOI Foundation that authorize their member organizations (repositories, publishers, etc.) to act as registrants, issuing DOIs for data sets (and in the case of CrossRef's registrants, also for textual publications such as journal articles) and keeping centralized records of the metadata associated with each DOI collected by each registrant. Benefits of using the DOI include the commitment to maintaining the links to content for which the DOI is known, long established awareness of the DOI, and its application in citation structures elsewhere in scholarly communications. As such, people are highly aware of the system and its use. A potential disadvantage, however, can be the small cost incurred for each DOI issued. For example, small granularity could increase costs because of the need for a greater number of DOIs.

Persistent identifiers (e.g., DOIs and URIs), therefore, can assist with consistency, credit, and findability of data when it comes to digital data citation. Their inherent nature permanently identifies any given data set independent of its location. However, design and development of the infrastructures to facilitate digital data citation are a work in progress, and the existing technologies are not entirely intuitive nor well documented. Additionally, there is still a lack of consensus as to the necessary and sufficient characteristics of an identifier. At a minimum, an identifier must be unique and cannot be reusable or transferable. An underlying issue is the level of information being identified. Some identifiers are sufficient for assigning credit but do not provide enough information to identify the source of the data.

In addition to DOIs, other systems for persistent linking of content exist, notably the Persistent Uniform Resource Locator (PURL), the Archival Resource Key (ARK), and—albeit less reliably—permalinks and redirects. There are also some who advocate for the use of location-based URIs (HTTP-URIs), which might include the HTTP version of DOIs. The unique DOI number that is assigned by a registration agency begins with the number ten and contains a suffix and a prefix separated by a slash. The prefix is a unique number of at least four digits assigned to identify an organization, and the suffix is assigned by the publisher to meet publication standards. The distinction between using DOIs and the DOI-based HTTP-URI may seem a very modest one, but the latter imparts added functionality in terms of direct actionability, machine interoperability, and resolution to other associated metadata or services.

5.2 Incomplete practices and gaps within metadata elements and data citation

While the basics of data citation can be derived by analogy to the citation of textual publications, especially electronic ones, finer points of concern exist, such as issues of granularity, version control, microattribution (fine-grained and unambiguous credit), contributor identifiers, and facilitation of reuse, that merit special attention.

---

[8] In conjunction with the British Library and the Digital Curation Center, BioMed Central maintains a comprehensive "List of Research Data Repositories," which has been updated through October 2012. https://docs.google.com/spreadsheet/ccc?authkey=COmDvOUB&key=0Aok0Od_Hhd1XdEdiRXVCbDlFWk8wNW5FYlBBTndyaVE&hl=en_US&authkey=COmDvOUB#gid=0

See also the German initiative of data repository registry RE3DATA: http://www.re3data.org/

[9] For a list of data centers that utilize DOIs at DataCite see: http://stats.datacite.org/.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

5.2.1 Granularity

It is important to represent the level of granularity in a citation in order to enable reproducibility and findability. A citation may refer to the database, to a set of records within the database, or to a specific record. The citation system should allow for more than one level of granularity if needed (Buneman, 2006).

An open question remains, however, about the extent to which granularity should be applied to data, data sets, and even individual data elements in a set. Obviously, the answer to this question is specific to each instance and varies on contextual information, such as the scale, type of data, and coherence of individual datum. For example, a data set comprised of images might be assigned datum level identifiers because an image could be used outside of the context of the entire data set whereas an atmospheric measurement datum at a specific time point might be meaningless out of context.

Within identification there is the concept of functional granularity, described by Rust and Bide (2008), which should be applied here. Functional granularity describes a decision-making process for the assignment of identifiers at the most detailed operational level where it makes business sense to assign an identifier. Because there is a real cost to the assignment, metadata creation, and management of identifiers at any level, a balance must be struck between the costs of creating and maintaining those identifiers and the value that is brought by their existence. For most data sets, it is likely not of sufficient value to assign identifiers at levels below the data set or subsection levels for discrete subsets. For citation purposes, a similar principle should be applied; the granularity of citation and identification should be sufficient to locate the data being referenced at the level at which it makes the most sense to assess or replicate the finding. Obviously, some situations arise where such access or subsectioning is impossible. At this point, the citation, identification, and description should at least focus on the discrete data set level.

5.2.2 Version control

Because multiple versions of a data set may be cited throughout the lifecycle of the data, it is important to have mechanisms in place that differentiate between fluctuating versions and revisions. Good version control practices can address problematic issues, such as updates done by multiple people on multiple systems and citing different granularities of the same data set, such as a subset of an original data set. Just as granularity refers to the level of detail of data sets, version control tracks revisions to those data sets (regardless of their granularity level). Defining the citable unit can be challenging. Version control software or services are useful tools to track data and metadata revisions. Such tools might be adapted to facilitate good data citation practices.

5.2.3 Microattribution

Microattribution is a current workaround for the scholarly recognition of smaller contributions of a particular author (Fenner, 2011; Partrinos et al., 2012). In many domains, such as the life sciences, data sets may be compiled from several contributors. Listing every creator may be perceived as cumbersome, but technology will relieve this perception and enable equitable attribution practices. The Scholarly Contributions and Roles Ontology (SCoRO) (http://purl.org/spar/scoro/), for example, provides a vocabulary for such credit assignment for data contributions, and the Scholarly Contributions Report Form (SCORF) (http://purl.org/spar/scoro/scorf/) provides an easy-to-use spreadsheet for entering such ontology-compliant information. Where previously these contributors were simply left out of the scholarly contribution, a table listing the data and agent now provides the contribution and is included as supplementary data. Additionally, microattributions are key to providing credit to nanopublications (the smallest unit of publishable information that can be uniquely identified). These types of publications can be attributed by implementing the elements of assertion and provenance. An instance of microattribution occurs in the March 2011 issue of *Nature Genetics*, and the authors conclude that microattribution: "demonstrably increased the reporting of human variants, leading to a comprehensive online resource for systematically describing human genetic variation in the globin genes and other genes contributing to hemoglobinopathies and thalassemias" (Giardine et al., 2011).

5.2.4 Contributor identifiers

Contributions made to the creation and maintenance of data collections are often unacknowledged. This is in part because of a lack of formal citation practices around data but also in part because of the fine granularity of many contributions and the problem of disambiguating contributor names and identities ("Credit Where Credit is Due," 2009).

A number of emerging approaches to this issue include library name-authority databases (e.g., Virtual International Authority File [VIAF]), bottom-up and top-down researcher identifier registries (e.g., ORCID, ISNI), web-scale

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

identification frameworks (e.g., OpenID), and researcher and professional profile networks (e.g., ResearchGate, LinkedIn). SCoRO and SCoRF are also relevant.

The International Standard Name Identifier (ISNI) is an ISO standard (ISO 27729:2012) that identifies public identities of parties—the identities used publicly by individuals or organizations involved in creating, producing, managing, and distributing content. The ISNI system uniquely and authoritatively identifies public identities across multiple fields of creative activity (http://www.isni.org/).

ORCID aims to provide a researcher-centric solution to the name-ambiguity problem in scholarly communication by creating a registry for researchers. ORCID has established an author self-registration service and an author claim service for publications and is working with stakeholders, such as publishers, research institutions, and funders, in order to link digital research to this registry (Haak et al., 2012). Early adopters include most of the major scientific publishers, CrossRef, and Wellcome Trust.

Furthermore, ORCID is also actively working to integrate ORCIDs, (i.e., ORCID IDs) into data citation practices. ODIN, a 2-year project with the DataCite Consortium, CERN, the British Library, Dryad, arXiv, the Australian National Data Service, is funded by the European Commission and will integrate ORCID identifiers into the DataCite citation registry (see http://odin-project.eu/).

Identifier systems such as ORCID would make possible microattribution that enables researchers to have a constantly updated summary of all of their contributions to science "going far beyond the simple publication list" (Butler, 2012).

5.2.5 Facilitation of reuse

Access to research data, as facilitated by data citations, requires technological infrastructure that is appropriately designed and based on interoperability best practices that include data quality control, security, and authorizations. Currently, interoperability at both the semantic and the infrastructure levels is important to ensure that data citations facilitate access to research data. However, organizations working to develop improved infrastructures that foster interoperability should widely communicate the standards, guidelines, and best practices that are being implemented; adopt standards for data documentation (such as metadata) and dissemination (data citations, including bidirectional links from data to publications and vice versa); and maintain an up-to-date knowledge of the evolution of not only the technologies implemented but also the best practices efforts being executed by the community of practice.

5.3 Current and emerging tools, technology, and infrastructure

5.3.1 Introduction

The creation of policy and standards for the communication, management, and creation of research data is becoming increasingly important to the development of the scientific community as a whole. The research community needs ontologies, tools, technologies, and infrastructure that allow researchers to describe and manage data while facilitating the creation of digital data citations that enable discovery, use, reuse, and knowledge extraction. Having standards in place will foster an environment that incorporates data citation into the information lifecycle and adds value to research materials. The goal is to make these citations available in a format that is understood by both humans and machines.

Although some organizations have generally accepted processes for data creation, use, and citation, there is no standardized implementation of citing data. Only a few journals require formatted data citations and attributions. At the same time, most funding agencies and research institutions do not recognize data citations as part of the evaluation and recognition process. Simultaneously, more granting agencies expect researchers to collect and share their data sets as they are considered research products.

Data centers create their own infrastructure, which is designed to encourage the participation of data producers. The data center infrastructures follow a basic common model: a database to store data and a web interface for searching and browsing. The interface may implement protocols for interoperability and information exchange, but it has elements that change according to the scientific fields and frequently even within the same fields. While many repositories assist with the data archival, preservation, and sharing processes, there is no standardized or widely used workflow system or repository tool. The research community should encourage interaction among data centers while promoting common principles for best practices, including the implementation of open architecture concepts.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

5.3.2 Current tools for data citation discovery, tracking, and reuse

Citing data sets and resulting documents as well as other peripheral products of research, such as software programs, protocols, and workflows, poses significant challenges for facilitating comprehensive discovery and retrieval for reuse. One challenge is how data citations deal with identification and location of data. Another concern is how to catalog these data effectively and at the appropriate level of granularity.

As of January 2013, there were no tools for automatically indexing citations to data sets made in the reference sections of scholarly papers although such tools are being developed. Several products have emerged in recent years, however, which offer some methods to track data use. Only one of these products currently provides data citation indexing. The products are summarized briefly below.

A. Thomson Reuter's Data Citation Index

The Data Citation Index (DCI) (http://wokinfo.com/products_tools/multidisciplinary/dci/) is a new subscription-based tool from Thomson Reuters that aggregates data set metadata and citation information. At this point, all of the "cited by" data that DCI reports for data sets are based on data provided to it from the data repositories themselves. Repositories such as the Inter-university Consortium for Political and Social Research scour the literature for citations to their data sets and report this information to the DCI. The DCI plans to automatically index data citations in reference lists in the future. Citation information in the DCI is only available by subscription. The data may not be openly built upon or redistributed, and, so far, no application programming interface to support integration is advertised.

B. Other tools

Other useful tools that work in context with efforts on digital data citation include the following: Thomson Reuter's Web of Science provides access to citation databases although data citations are not indexed in the Web of Science. Elsevier's Scopus is a large abstract and citation database, but data citations are not currently indexed in Scopus (Piwowar, 2010). It plans to automatically index citations to data sets at some point in the future. Microsoft Academic Search also is a public search engine for academic papers and literature.

Google Scholar does not consider data sets to be first-class scholarly products so data sets are not included in search results or researcher's profiles unless they are formatted to look like articles (Boettinger, 2012). Google Scholar does not have an API to support reuse and has stated that due to agreements with publishers, they will not offer an API for the foreseeable future. However, Google Scholar can be used to track data reuse another way: searching for attribution to data sets in the full text of papers rather than just their reference lists. In these cases, it is not necessary that the data sets themselves be indexed: A search in Google Scholar for the term 10.5061/dryad.* returns many hits to papers that include this DOI somewhere in their text. The search term, when used with a period (.) and wildcard (*), directs Google to match any number of characters on a page that start with the same query word, in this case Dyrad DOIs (10.5061/dryad), but has any number of characters afterword. Querying full text to detect data reuse has several disadvantages, however. In addition to the Google Scholar export, API, and scope limitation mentioned above, searching in full text rather than just the references section returns data sharing statements in addition to attribution in the context of data reuse. Full text queries would be needed to exclude these statements of data availability from estimates of third-party use. Such queries do not have high accuracy.

As an alternative to the commercial sources of citation information described above, the Open Citations Corpus (http://opencitations.net) is a JISC-funded database of scholarly citations in which the citation data are published under an open CC0 license in both BibJSON (http://www.bibjson.org/) and RDF formats. At present, it contains citations harvested from the reference lists of the Open Access papers in PubMed Central and in the arXiv preprint server, totaling about 40 million citations. However, in 2013, it will start to harvest bibliographic citations via CrossRef in an ongoing manner from both open-access and subscription-access journals and will ingest data citations from both CrossRef and DataCite.

5.3.3 Emerging tools

The current system of available tools is not sufficient for effective management of the data collected because the technology tools available do not provide all the flexibility needed to manage data from different communities, generated in different ways, and in different formats.

Several emerging tools return alternative metrics on data sets. These new metrics are based on the Social Web and assist in the analysis of scholarship (Priem, Taraborelli, Groth, & Neylon, 2010). Alternative metrics include

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

mentions in blog posts, Twitter, Wikipedia, bookmark managers, and other online sites and services, some of which are summarized below.

A. Altmetric.com

Altmetric.com (http://www.altmetric.com/), a startup affiliated with Macmillan Publishing through Digital Science, reveals alternative metrics of anything with a DOI or other standard identifier. For example, it can find mention of a data set in blog posts, tweets, and mainstream media. Such data can be queried for free on an item-by-item basis through their bookmarklet or by embedding the Altmetric.com Score on a webpage. Metrics are rolled up into a single proprietary Altmetric Score. Viewing full provenance data requires a subscription. Altmetric.com also has an API that allows free use of data for public research projects and small projects of nonprofit organizations in a noncommercial setting. A fee is required for commercial use, however.

B. ImpactStory

ImpactStory (http://impactstory.org/) is a nonprofit startup initially funded by the Alfred P. Sloan Foundation. The organization reports alternative metrics for a wide variety of research objects: articles, data sets, software, blog posts, posters, and online lab webpages. Metrics are pulled from many sources, including blog posts, Twitter, Wikipedia, and attributions in full text of articles in PubMed Central. ImpactStory has integrated with several data repositories (Dryad and FigShare as of January 2013) to also report download statistics.

ImpactStory's metrics are reported with open links to full provenance data and with percentiles to provide a reference to other data sets published the same year within the same data repository. Researchers can build and display an impact CV, including their data sets alongside their research articles. ImpactStory data can also be embedded on a webpage through a javascript widget or accessed through the API. Data are free to everyone once ImpactStory has collected them, but there is a fee to register more than 1,000 items for ongoing collection.

5.3.4 Technology design to leverage tools for data citation

The world of data citation is as expansive as the World Wide Web, where the online documents are associated through hyperlinks in an ever-growing manner. Conventional web tools for data citation should be enhanced and new technology should be created that better allows users to traverse links from web document to data sets. We suggest the following system design leveraging tools for data citation:

A. Browser tools

Data browsers should be an integral part of the web browser through the browser plug-in function. In addition, bi-directional access from data to its linking documents should also be integrated to data citation infrastructure because it will allow for findability of related documents that share the same or similar data as well as different interpretations or "value" of the data in varying documents.

B. Dynamic citation tools

Data citation searches should be made available not only as standalone applications but also as services embedded in conventional online document editors. Embedded search services are especially useful because they also collect information on reference sources (documents) and calculate statistics on data citations. The use of data reference management tools could greatly assist with the data management process. Numerous reference management software tools are available, both open source (such as Zotero from the Center for History and New Media at GMU) and proprietary (such as EndNote from Thomson Reuters and Mendeley from Elsevier), which assist with the management of bibliographies and references when doing research. However, because they were all designed for bibliographic references, no known reference management software packages to manage digital data references currently exist.

C. Search tools

At least one search paradigm of data citation will be analogous with semantic web searches that employ the Resource Description Framework (RDF) or SPARQL Protocol and RDF Query Language (SPARQL) (http://www.w3.org/TR/rdf-sparql-query/). The data model of data citation thus becomes a semantic triple: subject (i.e., source document attributes), object (i.e., target data attributes), and properties (i.e., citation parameters, if any). For instance, unlike a conventional web search, a simple query keyword such as "climate change" will be interpreted primarily as the subject keyword so that it will find data sets cited by "climate change" related documents (Zettsu, Gonzales, Ong, & Murayama, 2012). Additionally, the same keyword acts as an object keyword to find any

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

documents citing "climate change" related data. In this way, the data citation search is necessary to coordinate the triple semantics. Dynamic citation links can be realized by embedding search queries in the data citation format in the same way as in a dynamic hyperlink of the conventional web (RFC3986, Section 3.4) (http://tools.ietf.org/html/rfc3986#section-3.4) and would provide additional functionality for navigation of data citation based on user context or personalization.

D. Archiving tools

Data citation archiving is also important to preserve complete snapshots of data citation over time, which contains versions of referring documents, referred data, and the links among them. Database management technology for handling extremely large citation graph data with version control is required, as is a standard protocol for data citation harvesting.

E. Data citation mining tools

Data mining technologies will be enhanced, and new technologies will be developed for discovering specific knowledge from the data citation archive:

- Association rule discovery: finds associations of typical source documents attributes and target data attributes. They can also be applicable for automatic generation of data citations (Gonzales, Zhang, Akahoshi, Murayama, & Zettsu, 2012).
- Community discovery: finds tightly connected citation clusters or "data leveraging communities."
- Hub/authority analysis: discovers authoritative data cited by many documents as well as hub documents citing many data sources.
- Co-citation analysis: finds data frequently cited together from a specific subject.

F. Platform Support Tools

Citations should be web service ready and in compliance with web service standards and technologies. Platform support for security and privacy of data citations should include provenance tracking (Moreau et al., 2011) and security protocols for data access control through data citation.

5.4 Conclusion

As technology evolves and becomes more sophisticated, it facilitates the generation, management, analysis, and dissemination of data. Standards, such as protocols for data exchange, best practices for publishing, and tools for cataloging, need to be developed by the communities of the Internet in parallel with technological progress. Achieving attribution standards and best practices with common data citation will make the management and dissemination of data sets easier and promote collaboration, reuse, and discovery. Optimally, data products should be managed on an infrastructure that supports data collection, curation, citation, and attribution while also fostering usability, definition of identity, persistence, and discoverability. A deeper understanding of how technologies associated with data citation affect the use and reuse of data needs to be developed.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Chapter 6 THE SOCIO-CULTURAL DIMENSION: EXAMINING THE BENEFITS AND CHALLENGES TO ADOPTING GOOD DATA CITATION PRACTICES

6.1 Introduction

Previous chapters have surveyed which data citation practices already have been adopted in different research communities and subdisciplines and what needs to be considered in developing such practices. This chapter looks at the benefits and challenges from the application of good data citation practices by major stakeholder groups in the research enterprise.

6.2 The benefits of good data citation practices

While all of the aforementioned strategies support good data citation practices for data management and sharing, implementing them is a crucial challenge. To help move adoption forward, it is important to first understand the identifiable benefits of data citation that can appeal to a wide range of players, including data creators, administrators, data centers, funding organizations, publishers, and the research community as a whole.

6.2.1 Benefits for data creators

For those who produce data sets, sharing those data effectively can open up opportunities for new collaboration, publication, increased impact, and reputational benefits. Researchers who build, maintain, validate, and collect the data for large databanks have to ensure that the data are of high quality and that the associated metadata and documentation are complete and understandable. Completing these tasks can be a major undertaking, which leaves little time for the analysis of the data required to produce a paper suitable for journal publication. Citation of data therefore can provide an alternative route to intellectual credit, additional publications, and wider recognition to data scientists so they can continue their primary work of ensuring data quality and curation while still getting professional recognition.

6.2.2 Benefits for universities and research institute administrators

Managers at universities and other research institutes perhaps have the smallest stake in the implementation of data citation practices because these entire stakeholder groups generally are neither the direct producers nor users of research data. They nonetheless should be concerned with the overall health of the research enterprise, and the broad implementation of data citation practices should improve accountability, recognition, and effectiveness of research, among other benefits discussed above.

They also need to be accountable for complying with the research funders' policies and the prevention of fraud. For example, research administrators are now increasingly compelled by funder mandates to manage the data produced by their researchers for prolonged periods of time after the end of the relevant research projects when the researchers themselves may have changed projects, fields or institutions or indeed may have retired or died. They may require information about the data published by members of their institutions so as to be able to better report on all the research outputs of their institutions, not least for the purposes of quality evaluation and research assessment.

6.2.3 Benefits for data centers

Data centers or repositories are the intermediaries for curating and disseminating the data made available by data producers to the users of those data in the research community and, in many cases, the general public. Libraries also increasingly provide this function, but for the sake of simplicity, we subsume libraries under the "data center" label.

To the extent that data centers operate openly and make their data available, either freely or with some user restrictions, such institutions can benefit substantially from a well-implemented data citation practice. Because the mission of such institutions is to disseminate the data in their repositories and to provide to the users of their data a well-documented method of attributing credit to the data providers and to themselves, it will clearly be of value to them and their depositors to have systematized data citation practices. They also are fully engaged in data-related work and have the personnel and infrastructure to implement data citation systematically and thus have a strong reason to adopt data citation best practices wholeheartedly. These should include not only the facilitation of citations of data held in their repositories by publications but also the citation by and alongside the data sets curated in their

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

centers in all related publications. The uptake of such practices should therefore be most rapid and perceived as least onerous by data centers.

More specifically, providing data citation links allows data centers to promote further discovery, analysis, and collaboration and to track and record reuse of data sets (Piwowar, forthcoming). The value of data centers in their role in the advancement of science and in addressing global scientific problems can be objectively understood, which can have a bearing on their funding. Data centers may also choose to utilize citation measures and impact factors similar to those currently used in journals thereby creating further incentives for data creators and scientists to engage in good data management and citation practices. Data centers promoting and employing such practices, among many others, include the ICPSR and ORNL DAAC, both discussed in Chapter 4.

6.2.4 Benefits for funding organizations

A driver for funding organizations is obtaining the best possible science for their money. Modern research can be expensive. Data citation may translate to better financial return from research investments as the data can be reused and sometimes repurposed.

A key principle of citation noted in Chapter 3 is that the citation and the object being cited should be persistent. Citation therefore encourages the data set creator to upload the data set to a trusted repository where it will be backed up, properly archived, and curated. As a result, the problems of data stored on obsolete or degrading media will be avoided thereby minimizing the need to repeat costly research. Similarly, citation allows for the data set to be linked to other publications, enabling it to be discussed in the literature and providing a mechanism for formal or informal peer review. This review process can reassure the funder that the published data set is of good quality and that the research was carried out appropriately.

A related benefit for funding organizations can include better utilization of data, for example, through the superior discoverability of cited data. Finally, other potential benefits for this stakeholder could include better trust and relationships with members of the public, who either directly or indirectly support the funders and their research clients.

6.2.5 Benefits for publishers

Data citation provides publishers with a means of tracking publications that are using or citing researchers' data sets. Early indications exist that publications that make underlying data available are associated with increased citation rates (Piwowar & Vision, 2013).

As a result, publishers that make data citations available will benefit from increased exposure because the citation of data publications will increase citations of related research papers and thereby attract more authors and readers. In general, better integration of data and publications that moves beyond simple links between articles and data sets creates new ways of making research publications actionable for reuse (Kotarski, Reilly, Schrimpf, Smit, & Walshe, 2012). It also should create a positive reflection on those publications that link to databases as sources of supplemental information.

In addition, formal provenance and attribution practices for data strengthen accountability, increase incentives to disseminate data for reuse, decrease the ambiguity of replication studies, and deter scientific fraud. This is increasingly salient, as reflected by numerous articles drawing attention to scientific fraud and increase in retraction rates (Steen, 2011). Furthermore, journals are being approached to accept data as supplementary materials–especially as funder data sharing mandates have become implemented–but have limited willingness and capacity to host data sets. Nonetheless, as discussed in Chapter 4, standards and practices for managing data associated with publications, such as the NISO *Recommended Practices for Online Supplemental Journal Article Materials* (2013), are beginning to emerge.

6.2.6 Benefits for researchers and research communities

Citation of data provides the wider research community with a mechanism to locate and discover data sets and to be confident that the data sets will be persistent. It also provides information on the appropriate responsible parties to contact about reuse of the data set. Data citation can be useful to researchers outside the immediate field as centralized metadata catalogues for cited data are being created (for example, by DataCite), which will provide a convenient starting point for browsing. These catalogues will encourage inter-disciplinary collaboration and open up the user base for the data sets and the supporting data repositories. Moreover, the ability to cite data sets makes it

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

easier to validate conclusions through the reanalysis of those data sets. As a result of these and other benefits, the wider research community, as a whole, will likely benefit from better science.

6.2.7 Benefits for the broader society

Of course, if all the benefits from data citation practices described above come to pass for the stakeholder groups in the research process, the whole society will benefit from improved efficiency, return on the public investment in research, and the general progress of science and its applications. In addition, many members of the public will wish to use some of the research data for other, even unintended, purposes, to create new applications, to educate themselves or others, and to pursue other serendipitous results. Data citation will also foster public trust in research results as all citizens will be better able to find the facts underlying the research results on which public decisions or policies may be based (Boulton et al., 2012).

6.3 Socio-cultural and institutional challenges

The socio-cultural and institutional obstacles to the broad implementation of new data citation practices have both a general, overarching dimension and more specific considerations from the perspective of different stakeholder communities. As a general matter, there are substantial barriers to the uptake of any new processes, attributable to a "first-mover" problem and questions of legitimacy and trust.

The rise of data-intensive research or "data science" began less than two decades ago with the advent of global digital networks and became more mainstream only in the past few years (Commission in Physical Sciences, Mathematics, and Applications, 1997; Hey et al., 2009; The Interagency Working Group on Digital Data, 2009). Pervasive recognition of "big data" research and applications has been even more recent.

The purpose of this report is not to describe in detail the processes of data-intensive research and applications. The point is that that these activities are quite new. Although researchers and businesses have always generated and managed their own data, the technologies have grown in capacity and decreased in cost, thus making the changes quantitatively and qualitatively different. As with any new technological developments, however, the social systems are much slower to adapt and respond than the technical systems in ways that are more efficient and effective (Uhlir, 2006). It takes time for potential users to develop a sense of legitimacy and trust in new technologies and processes and to reach a coordinating equilibrium among actors around policies and standards. This area is no different.

Data citation practices would form an underlying social and scholarly infrastructure that could go a long way in helping to build such legitimacy and trust, providing a means for enhancing recognition and reward for individuals involved in data related work. However, the situation is further complicated by specific social and cultural barriers that operate in different stakeholder communities.

6.3.1 Data producers

Data set creation takes time and effort. Even if the process of data collection is automated, there often are further organizing, processing, documentation, and quality control steps to be taken before the data can be properly published for third-party analysis and reuse. This pre-processing work is largely disregarded when it comes to obtaining academic credit because the result of this work is a data set rather than a formal publication. It should be noted, however, that in some fields (e.g., climate science, astronomy), data sets are represented by and cited as proxy papers. Nevertheless, there can be potential risks associated with sharing, including reputational risks of the data contents, risks of the data being misinterpreted, or risks of the data being used to obtain research funding without collaboration with the data producer. One way to combat these risks is to ensure data citation standards and best data management practices are followed so proper use is made, proper citations can be given, and proper credit accrues to the producers of data.

6.3.2 University and research institute administrators

Improved citation and attribution of the data produced by university employees may or may not raise the visibility and rankings of certain departments or the university itself while incorporating those counts in the performance reviews of personnel can add some burden to the process. In addition, university administrators can suffer from the same general uptake barriers as others in the research system. On balance, this stakeholder group may have the most additional work and the least actual or perceived benefits.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

6.3.3 Data centers

Despite some good existing examples and practices, only a small percent of data centers have implemented a preferred data citation model thus far and made that a suggested practice to all their customers. Reasons for this might be insufficient awareness or knowledge, an absence of incorporation in their existing workflow, and a lack of demand by their contributors, customers, and funders. In any event, data centers should see a high value and be a primary audience for data citation practices.

6.3.4 Publishers and editors

Publishers of journals and other publications serving a research area that tends to use many data sets in publishing research results have much to gain from a well-established data citation system. Some, however, may view it as a greater burden than benefit. Any active support of data citation protocols adopted by the publishers will become an additional responsibility of their editors so there will be some cost associated with such an implementation.

6.3.5 Research funders

In principle, research funders should have a positive outlook as well on the broad implementation of data citation practices by the research community. As noted above, it can help to show additional value from some of their public investments in research, it can assist in the evaluation and planning of their programs, and it can be used as a further justification for their annual appropriations (if they are government funders) or donations (if they are private foundations). They are concerned as well about the overall health of the research system.

At the same time, research funders are demonstrably conservative organizations. Although they have begun to recognize the value of data work and data-intensive science, only a few in the world (e.g., the United States' National Institutes of Health, the National Science Foundation, the Wellcome Trust, and some of the Research Councils of the United Kingdom) have implemented even a weak data management plan mandate and a few positive requirements for the deposit of data. For the rest, their uptake of data citation protocols, much less requirements, may similarly be assumed to lag behind the activities of other stakeholders in the research system. New research practices tend to arise in the research community from bottom-up, rather than top-down, and data citation is likely to be no different.

6.3.6 Individual researchers and the research community

Both the need for data citation practices and the socio-cultural barriers to their uptake are most important for researchers themselves, as both the producers and users of data. The benefits are quite clear and have been identified above and in other places in this report. We focus here on the barriers.

As with any new practice or method, there is a counter-cultural element to adopting it in a comprehensive way. Researchers need to understand how it can benefit them, and they need to overcome the inertia or resistance associated with any new system. Concerns about additional work and time should be addressed. An acceptable workflow needs to be created. However, most researchers, while experts in their own fields, have little awareness of metadata standards for data publication and information science in general, leading to cognitive and skill barriers that prevent them from undertaking routine best-practice data management. Put crudely, the large amount of effort involved in preparing data for publication release, coupled with the negligible current incentives and rewards, prevents many researchers from doing so. Subject-specific librarians, with appropriate new skill training, could help researchers in this regard.

Additional barriers are the increasing pressures faced by academics to read the ever-increasing number of research papers of relevance to their own work, to attain more research grants, and to publish more papers in high impact journals. These activities are crucial both for their own career advancement and their departments' survival in the current economic climate. All other activities not on the critical path to success in these three areas, such as data management, are further down the priority list.

Moreover, many researchers also have the unfortunate desire to ensure that the only people who can reuse their data are themselves. Thus, the need for a sea-change in attitudes towards open data cannot be overstressed, and this has only partially been achieved thus far (Peroni & Shotton, 2012; Shotton, 2011).

Finally, it should be noted that significant differences among disciplines or types of research should be taken into account. Although the concerns or barriers specific to each subdiscipline are beyond the scope of this report, whether data are more likely or less likely to be made openly available for use by others is clearly important as a

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

threshold condition. Therefore, not all kinds of research will benefit equally from data citation practices, and actual or perceived barriers will vary as well.

6.4 Economic and financial challenges

Even a cursory cost-benefit analysis of data citation should be positive for any data set that is broadly available and expected to be used by others. The costs associated with data citation are low, but not zero. If the data set is likely to be reused, the producer of those data will almost certainly benefit from a proper digital citation of that work. Implementation of DOIs and other persistent digital tags costs some money and a little time by the researcher. These can be perceived as strong barriers to implementation as the analysis of voluntary open access author deposits has shown in another context.

Additionally, the persistence of the connection between data citation and the actual data ultimately must also depend on some form of institutional commitment and allocation of economic resources (Altman & King, 2007). Given the huge increases in the amount and types of data being generated or used by the scientific community, this is likely to involve a spectrum of approaches–from top-down centralized archiving, to collaborative institutional arrangements, to bottom-up web-crawling and capture, and will, in turn, place new financial demands that need to be met from the available sources of research funding. This pressure may be felt particularly acutely in the Arts and Humanities, many of whose researchers operate without external grant funding.

From a broad perspective, there will be shifts over time and by discipline, as data citation practices become more routine, more standardized, and more embedded in the normative processes. Nevertheless, there will be some winners and losers. Without mandates or a default rule of openness for the public disclosure of research data, the winners will be self-selecting because they will come from those data-intensive and open data disciplines where the benefits will be obvious. If data disclosure or "publishing" mandates are imposed, those producing data most used by others will be those with the most to gain. In an era that is likely to be characterized by smaller research budgets, however, care must be taken not to implement unfunded mandates that do not have a strong cost-benefit ratio and do not avoid unnecessary negative externalities. Currently, there is a lack of empirical data about the costs and benefits of data citation, and this will need to be studied more closely as the practice develops.

6.5 Conclusion

Significant normative barriers to acceptance of data citation practices by different stakeholder groups in the research system exist, and overcoming them will be a challenge. While researchers, research institutions, data centers, scholarly publishers, and research funders generally have common goals in the research enterprise, broadly defined, they may have different perspectives and interests that they will defend and promote. In order to implement good data citation practices comprehensively in certain disciplines, there will need to be buy-in by most if not all these stakeholders. Both a specialized focus and a substantial and sustained effort will need to be devoted to achieving this.

Good data citation practices can be expected to have a high benefit-to-cost ratio, both from a general, systemic standpoint and from the perspective of major stakeholder groups and discipline areas. Because the latter will vary according to discipline, time, and institution, studies should collect empirical data on the costs and benefits for implementation approaches. Examples of successful data citation implementation approaches can help make the case to the different stakeholder communities.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Chapter 7 OPEN RESEARCH QUESTIONS

7.1 Introduction

Data citation is a rapidly developing area of research and implementation—yet much remains to be done. Previous chapters have implicitly and explicitly identified a set of challenges.

Chapter 5 described the tools and infrastructure supporting data citation and, in so doing, identified a set of technical implementation challenges. In particular, that chapter identified the need to integrate data citation support into cataloging and indexing infrastructures, citation management tools, workflow systems, manuscript management systems, and digital repository tools. Chapter 5 also identified more general technical challenges, such as versioning, standards, licenses, interoperability mechanisms including protocols for human and machine readable landing pages, processes and services for indexing data citations, and support for granular references in "deep citation."

Computing pioneer and long-time senior researcher at Microsoft Research, Jim Gray, purportedly coined a benediction uniquely appropriate to, and common among, scientists and engineers: "May all your problems be technical." In that spirit, Chapter 6 identified a number of socio-cultural, institutional, and economic challenges to be met in the implementation of a data citation infrastructure. Resolving these challenges requires coordination as well as appropriate resources and research.

The ultimate goal of improved data citation practices is to advance the practice of science and science policy. We describe how data citation can contribute to domain research and science policy.

7.2 Enabling research

As a complement to the inventory of implementation and socio-technical challenges in Chapter 5, this section describes the core research challenges related to implementing data citation and evaluating the results. These research challenges cluster in several groups, each of which lies along a spectrum of abstraction, from basic theory to applied research. In addition, as described in Chapter 6, socio-cultural problems require investigation and solutions, and, as discussed below, evaluation research will be required to follow implementation of data citation practices and standards in order to gauge their effectiveness. The required research activities are shown in Figure 7.1 below. Each column represents a stage of research activity (note that socio-cultural investigations may be conducted in parallel) while the categories within columns reflect research topics. Arrows are used to indicate dependencies among related topics across research stages.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

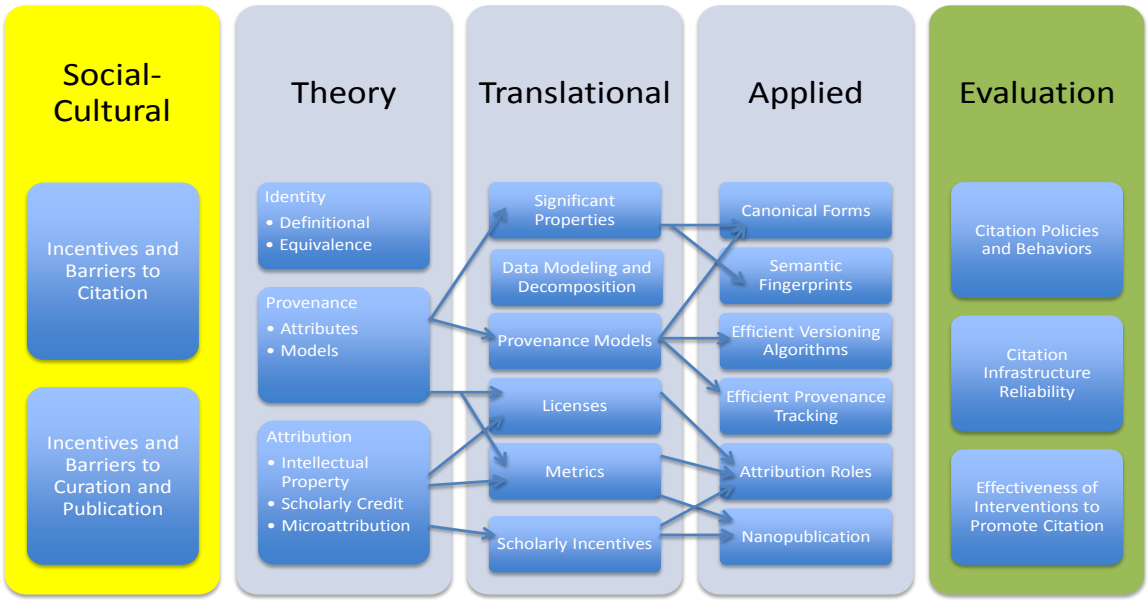

Figure 1. Research activities required to enable ubiquitous, seamless, and interoperable data citation

The most abstract of challenges involves theory*;* once a sufficiently robust theory has been established, *translational research* is necessary to embody core theoretical concepts and basic methods and approaches that can be applied to specific domains.

Applied research is then needed to adapt existing informatics and computer science research in order to incorporate these methods and approaches, yielding innovative solutions to the technical challenges of data citation. Much of the technical infrastructure can, and should, be implemented using existing tools and methods or marginal extensions thereof. However, before seamless, ubiquitous, interoperable data citation is possible, some existing methods and approaches developed in computer and information sciences must be adapted and incorporated into production-quality infrastructure. Finally, *evaluation research* is needed to assess the reliability and effectiveness of the research application in practical settings.

The research needed to enable data citation falls into several categories at the broadest theoretical level. As illustrated in Figure 1 under "Theory," there are three areas of theory which require elucidation:

- *Models of identity.* These theories involve defining 'data' itself, the identity of data and how to define equivalence and derivation relationships, and the granularity and structure of data. Theories of data have strong implications for determining what should be cited.
- *Models of provenance.* Provenance includes the chain of ownership of an object and the history of transformations applied to it. Models of provenance have strong implications for how data citation is integrated into the data curation workflow.
- *Models of attribution.* Attribution plays a key role in the incentives for citation. Models of attribution have strong implications for determining the presentation of data citations.

7.2.1 Identity

Identity is near to the heart of citation. To cite something requires that it be identified and that the identification mechanism be persistent. The current approaches to data citation (described in Chapter 4) all incorporate some form of persistent identifier.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

In Chapter 3, we recognized that a citation should provide an unambiguous identifier to the data cited, its location, and means of access. Although this does not imply that the unambiguous identifier is one-to-one, in many cases it will be useful to maintain a one-to-one relationship between the citation and the data cited. As Borgman (2012b) notes, however, the dimensions of data identity are unsettled; there is not yet a clearly agreed upon set of dimensions for data identity, nor for the levels needed to read, interpret, combine, compute upon and trust data. Similarly, Wynholds (2011) states that data are currently "unruly and poorly bounded objects" within scholarly works, and Renear, Sacchi, and Wickett (2010) assert that "Although definitions of data sets do appear to fit a common pattern, with recurring phrases and semantically similar terms, it is clear that there is no single well-defined concept of data set" (p. 3).

Chapter 3 also discusses the importance of granularity for citation. At present, effective granularity is determined by the practices at the hosting repository. In theory, the granularity of a citation should relate to the portion of evidence cited to support a particular assertion, finding, or analysis.

The challenges of modeling data identity are exemplified by three practical questions, which were raised, in different forms, in the findings of the 2011 Harvard "Data Citation Principles Workshop" (http://projects.iq.harvard.edu/datacitation_workshop/) and in Uhlir (2012).

- *The equivalence question.* How does one determine whether two data objects, not bitwise identical, are semantically equivalent (interchangeable for scientific computation and analysis)?
- *The versioning question.* How does one unambiguously assign at the time of citation a 'version' to a data object, such that someone referencing the citation later can retrieve or recreate the data object in the same state that it was at the time of citation?
- *The granularity question*. How does one unambiguously describe components and/or subsets of a data object for purposes of computations, provenance, and attribution? How does one incorporate this granularity with a bibliographic data citation to create a "deep" citation?

Natural corollaries to these questions involve considerations of scalability. For example, how does one track and recreate versions on large and dynamic databases? What data structures enable fine-grained access to data? How does one compute equivalence over the members of large collections for the purposes of de-duplication? Although there are no complete solutions to these problems, a number of promising approaches are emerging. These approaches include:

- Systematic identification of the "significant properties" of digital objects – those attributes that are used in later substantive/semantic interpretation of the object (Hedstrom & Lee, 2002; Hockx-Yu & Knight, 2008; Matthews, McIlwrath, Giaretta, & Conway, 2008).
- Semantic fingerprints for data objects, such as UNFs (Altman, Gill, & McDonald, 2003; Altman, 2008), which compute cryptographic hashes over canonicalized representations of an object, and perceptual fingerprints, which characterize uniquely the way that a data object is perceived (Cano, Batlle, Kalker, & Haitsma, 2005).
- Open annotation frameworks and ontologies (Van de Sompel, 2012) to allow interoperable annotation of digital objects. These frameworks include defining spatial (logical) and temporal granularity and could be used as a complement to bibliographic data citations to support deep citation.

7.2.2 Provenance

Citations are one of the tools that scholars use as part of documenting provenance. For data citations, provenance is a particularly important concern because many data citations are used to document a direct relationship between a published assertion and the underlying evidence that supports it.

Supporting this evidentiary relationship does not necessarily require recreating or establishing the entire provenance chain, and much of provenance can be considered as orthogonal to citation, as Groth (2012) argues. Nevertheless, as Smith (2012) points out, enabling readers to establish authenticity of the cited object is an important use for citation, and this often requires use of provenance information. Furthermore, as Greenberg (2009) shows, citation chains may distort findings, even converting hypotheses into 'facts'.

There has been significant work in developing models of provenance for web databases (see Buneman, Khanna & Wang Chiew, 2001; Moreau & Missier, 2013) and for scientific workflow (Davidson & Freier, 2008). However, these models, and the tools supporting them–which are almost all experimental–have yet to be integrated at an operational or even theoretical level.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Chapter 3 established the principles that citations should enable resolution to the version of the data cited and that citations should contain sufficient fixity and other provenance information to verify that a data object later accessed is equivalent to the data object originally cited. This illustrates that the elements of provenance that are most important to citation are those overlapping with identity. Further, citing the data (evidence) associated with a particular claim may act to limit the distortion resulting from citation chains.

7.2.3 Attribution

As discussed in Chapter 3, a citation should facilitate giving scholarly credit and legal attribution to all parties responsible for those data.

Systems of attribution must navigate several different concerns:

- Legal attribution is founded on intellectual property rights and licenses as well as on strong normative values in the research community, and the data citations concern individual rights and norms of credit and publicity.
- Ontological attribution is based upon classifications of the intellectual relationships between citing and cited objects.
- Scholarly attribution is concerned with the incentives and systems of scholarly credit and evaluation.

There is active research in each of these areas of concern. In the legal area, while issues related to intellectual property have received considerable attention (see Chapter 2 and Pearson, 2012), management of the privacy rights attached to data is poorly understood. Where these privacy rights conflict with the right of publicity, there are implications for legal attribution and therefore consequences for citation.

From an ontological point of view, citations remain limited indicators of relationship. A citation to another article may serve the purpose of providing background information, paying homage, or offering supporting evidence, or it may serve to refute or critique a previous work (Case & Higgins, 2000). More sophisticated ontologies are required to describe the subtleties of these relations: Cronin (1984) reviews over 10 different proposed taxonomies of citation types and roles, some of which identify dozens of individual relationships. A more recent example is CiTO, which is an ontology that provides the means to express a subset of these complex relationships and to record them in machine-readable (RDF) form (Shotton, 2010).

Finally, from the viewpoint of scholarly attribution and credit, citation currently is connected directly with only a small part of the scholarly research lifecycle and value chain, as Bollen (2012) discusses. Related to this, scientific impact is a multi-dimensional construct; a single measure of impact does not necessarily describe the true impact of traditional publications (Bollen, Van de Sompel, Hagberg, & Chute, 2009; Bollen, Van de Sompel, Smith, & Luce, 2005), nor will existing measures likely capture the true impact of data. As Harley (2012) explains, a citation is embedded in a complex system of scholarly evaluation and incentives, and related issues of time, credit, and peer review still need to be adequately understood.

7.3 Social-cultural research

Borgman (2012b) suggests that data citation is best framed in terms of the general infrastructure for digital objects, which include combinations of technical, social, local, and global concerns. Key issues concerning the social infrastructure are the relationship of the reward system to data citation and the integration of related social practices for data reuse. As Borgman notes:

To reuse data, it is necessary to determine what useful data exist, where, in what form, and how to get them. In turn, data must be described in some way if they are to be discoverable. For people to invest effort in making data discoverable, they should receive credit for creating, cleaning, analyzing, sharing, and otherwise making data available and useful. To get credit, some means must exist to associate names of individuals and organizations with specific units of data. (p. 4)

There is a clear gap between policy and citation practices for data sets. One small-scale study found that 61 percent of the articles reviewed failed to provide even informal citation of the underlying data (Mooney, 2011). Similarly, a study of recently published articles in high-impact journals found that in the majority of cases, data to replicate articles was not clearly identified nor made available in a manner consistent with the policies of the publishing journal (Alsheikh-Ali et al., 2011). An analysis of 5,600 peer-reviewed articles that cited the NASA EOS instrument data over nine years found the citation record was substantially incomplete, many more articles used the data than cited it, and that citation of data sets in general is lacking (Major, 2011). Furthermore, because of the lack of

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

standards for data citation, constructing an index of citations to data from publications is a laborious manual process (Vardigan, 2012).

Data citation has been identified as one of a set of broader incentives that would promote effective data publication and sharing (Griffiths, 2009). However, the incentives and barriers to data curation and publication are complex and sometimes include competing factors, such as effort, skill, community norms, incentives, and risks (Borgman, 2007, 2012a; Edwards, Jackson, Bowker, & Knobel, 2007; Harley, 2012; Hedstrom, Niu, & Marz, 2008). Compounding this underlying level of complexity is the lack of uniformity in practices, requirements, and cultural norms across disciplines, which suggests that discipline-specific targeted research and implementation efforts will be needed to fully implement data publication and reuse (Griffiths, 2009).

Although the general categories of benefits and barriers are becoming clear (see Chapter 6), systematic measurement of these requires applied research. Where the net benefits are positive, applied research is required to determine the effectiveness of incentives and other policies in promoting good practices, both globally and within specific scientific communities.

7.4 Evaluation research

The last set of implementation challenges involves *evaluation.* This research is aimed at determining what implementation strategies work and how well. This generally requires establishing a baseline, testing the reliability and precision of the intervention, and then assessing the effectiveness of interventions in promoting proximate outcomes. In the area of data citation evaluation, research should include analysis of baseline rates of citation and existing policies, testing the citation infrastructure itself, and an assessment of the effectiveness of interventions, such as the introduction of new tools and policies, in influencing researcher's actions related to citation.

For example, existing work has provided some preliminary estimates of baseline data citation rates. White (1982) examined the indexing of three central data sets in social sciences and found that citation of data sets is highly inconsistent and incomplete—and that "hard digging" was required to determine usage of these data files. Sieber and Trumbo (1995) investigated the citation of the General Social Survey, one of the most important social science surveys in the United States, and found that less than twenty percent of authors included survey in-structure citation, and nine omitted any mention at all of the data source. Mooney (2011) analyzed a nonrandom sample of 49 journal articles representing 39 journals and 33 data set series and found that 61 percent failed to provide *any* type of citation. ICPSR's (2011) description of the methodology required to generate its "Bibliography of Data Related Literature" illustrates both the high degree of manual effort and the many specific challenges of identifying data use within current publications.

Some relevant evaluation research questions include:

- What are the range of and best practice examples of data citation practices – especially with respect to those issues discussed in Chapter 4 of citation granularity, citation syntax and presentation, licenses, versioning, dynamic works, provenance, and fixity?
- How does the prevalence of data citation vary across disciplines?
- Which journals, funders, repositories, and research organizations have data citation policies, and what are the main similarities and divisions among policy approaches?
- What proportion of data citations appearing in research articles can be resolved to data objects?
- What proportion of tables, figures, statistical coefficients, and other data-derived results provide sufficient granularity (e.g., through "deep citation") to identify the subset of data supporting the specific result?
- What proportion of citations provide sufficient provenance to verify that the version of the data referenced corresponds to the version cited?
- What are the effects of funder and journal policies regarding citation, replication, and access on data citation rates and on the replicability of published articles and the accessibility of associated data?

These and related evaluation research questions can be answered using a variety of approaches. These might include: a census and analysis of policies promoted by journals, funders, societies, and repositories; bibliometric analysis of citations among articles and data; automated tests of the resolvability of data citations; and manual evaluation of the provenance of cited data. The census will collect and code existing published journal policies on citation, data management, and data archiving.

Although we expect that analysis using complete replications will not be feasible, evaluation research could include: the development of coding rules, the coding of sample articles based on the availability of the data, the clarity of

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

data provenance and whether or not the necessary information is provided to replicate the article, and the independent coding of subsamples to establish measures of inter-coder reliability.

7.5 Domain research

Widespread adoption of better data citation promises to facilitate domain research as well as science metrics and science policy (which we discuss in the following section). For example, Borgman (2012a), in her analysis of the related topic of data sharing identifies benefits including reproducing/verification of research, both simple and complex reuse, making publicly funded research available to the public, and advancing the state of research and innovation. To a degree, data citation would be expected to yield these benefits as well. Other potential benefits of systematic data citation are summarized in Figure 2.

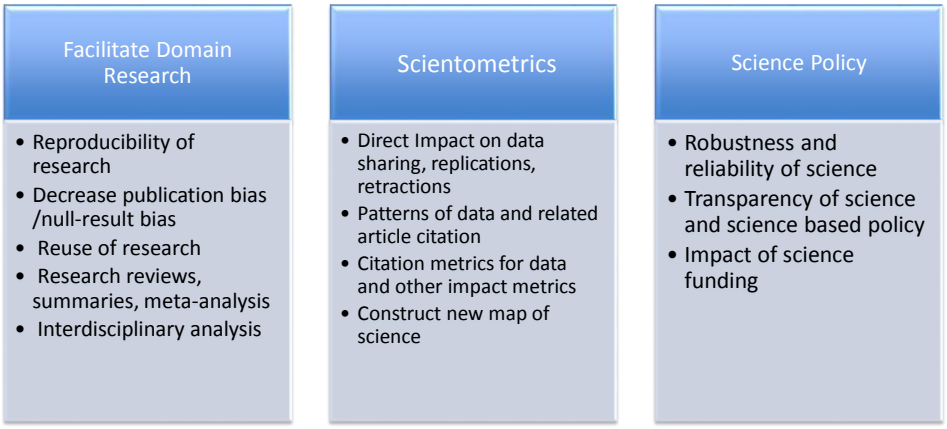

Figure 2. Research building on data citation infrastructure

In the remainder of this section, we discuss in more depth ways in which data citation may facilitate domain research through reproducibility, reuse, and interdisciplinary analysis. In the following section, we discuss the potential use of data citation to conduct scientometric analysis and to inform science policy.

7.5.1 Reproducibility

Reproducibility, or replication of research, is viewed as "the gold standard" for science. However, reproducibility is a value often honored in the breach. Although this problem has been noted for decades (Buckheit & Donoho, 1995; King, 1995; Rosenthal, 1979; Simon & Lesage, 1989), recently there has been increasing attention to reproducibility as some researchers have become increasingly concerned with the failure of articles to replicate (Ioannidis, 2005; Vul, Harris, Winkielman, & Pashler, 2009). A special issue of *Science* examines the approaches, benefits, and challenges to reproducibility across fields (Jasny, Chin, Chong, & Vignieri, 2011; Ioannidis & Khoury, 2011; Peng, 2011; Ryan, 2011; Santer, Wigley, & Taylor, 2011; Tomasello & Call, 2011).

Research reproducibility in social science has been particularly well studied, in part, because of all the sciences, it has the longest history of sharing and managing key databases, such as government statistics and major surveys (Alonso & Starr, 1989; Glasser & Bisco, 1966; Desrosières & Naish, 2002). Nevertheless, individual social-science research findings are surprisingly difficult to replicate. Most research articles fail to provide clear citations to data or to provide the computer code necessary to reproduce results. As a result, reproducing published tables, figures, and

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

results from the raw data are often difficult or impossible (Dewald, Thursby, & Anderson, 1986; Altman et al., 2001; Hamermesh, 2007; McCullough, McGeary, & Harrison, 2008; McCullough, 2007, 2009).

Scientists and other observers have proposed three types of approaches to solving the problem of replicability and reuse. The first approach focuses on tools for reproducible computation ranging from "statistical documents" (many of which incorporated Knuth's (1992) concept of literate programming to workflow systems to reproducible computing environments (Buckheit & Donoho, 1995; Deelman & Gil, 2006; Freire, Silvia, Callahan, Santos, Scheidegger, & Vo, 2006; Gentleman & Lang, 2007; Leisch & Rossini, 2003; Schwab, Karrenbach, & Claerbout, 2000). The second approach focuses on data management methods and tools (Altman et al., 2001; King, 2007; Anderson, Greene, McCullough, & Vinod, 2008). Increasingly, work in this area centers on issues of enabling long-term and interdisciplinary access to data (Altman & King, 2007; Gutmann et al., 2009). This requires that the researchers' tacit knowledge about data formats, measurement, structure, and provenance, and of appropriate data processing software, be more explicitly documented.[10]

The third approach focuses on the norms, practices, and licensing associated with data sharing, archiving, replication and the related incentives embedded in scholarly communication (Altman & King, 2007; Hammermesch, 2007; Hedstrom et al., 2008; McCullough, 2009; Pienta, 2006; Stodden, 2009). This approach seeks to create the necessary conditions to enable data sharing and reuse and to examine and align citations, data sharing, and peer review to encourage replicability and reusability.

Incentives for replication and reuse are currently weak in many disciplines; journals and citation practices are contributing factors. The reluctance of journal editors to publish articles either confirming or non-confirming replications taxes authors' incentives to create replicable work. Lack of formal provenance and attribution practices such as data citation also weaken accountability, raise barriers to conducting replication and reuse, reduce incentives to disseminate data for reuse, and increase the ambiguity of replication studies, making them difficult to conduct and publish. Funders are also reluctant to fund replication work in preference to novel research.

7.5.2 Systematic reviews

Verification essentially reviews the information and conclusions related to a single published article. A related method is the systematic review, in which the evidence for a particular finding is summarized, evaluated, and quantified across an entire set of studies. Arguably, the most successful example of the practice of systematic review is in medicine, in which the *Cochrane Collaboration* (http://www.cochrane.org/) has constructed, over the last 20 years, a database of over 5,300 systematic reviews. Cochrane Reviews are highly cited and widely regarded as being of exceptional quality. They have been found to be less biased than industry reviews (Jørgensen, Hilden, & Gøtzsche, 2006).

Review of the original individual patient data is considered the gold standard for producing these systematic reviews, and access to the data allows for the application of evidentiary analyses that is otherwise impossible (Stewart, Tierney, & Clarke, 2008). Notwithstanding, the statistical core of the Cochrane review protocol is meta-analysis, which accounts for the vast majority of reviews published and involves only a review of published results (Deeks, Higgins, & Altman, 2008).

The primary advantage of meta-analysis as compared to the gold standard is that meta-analysis does not require access to the original data. Thus, standardization of data citation and availability would reduce the barriers to gold-standard reviews in medicine. Further, it could enable systematic reviews to take place in fields where published results are less standardized than in medicine and thus less amenable to second-best methodologies such as meta-analysis.

7.5.3 Publication bias

Even a gold-standard review of previous research studies can analyze only those that are actually discoverable, which generally means that these reviews rely on published articles and data. New findings garner more attention

[10] Also see for example the "CRAN Reproducible Research Task View" (2013): http://cran.r-project.org/web/views/ReproducibleResearch.html and the "Reproducible Research Tools Page" (2013): http://reproducibleresearch.net/index.php/RR_links#Tools

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

than confirmations, and statistically significant findings garner more than null results. As a result, there is substantial evidence that journals are less likely to publish articles presenting null findings (Neuliep & Crandall, 1993; Sterling, Rosenbaum, & Weinkam, 1995) when submitted, and there is even evidence that researchers engage in data dredging, model fishing, or other methodological peccadilloes in search of results that are "significant" (Humphreys, Sanchez de la Sierra, & van der Windt, 2013). Furthermore, research studies with null or negative findings are more likely to end up in "the file drawer" (Rosenthal, 1979). (A related bias, noted in Greenberg (2009), is that positive findings may be cited more often than negative findings.)

The file drawer problem, or "publication bias" (as it is now known), was first raised as a concern in psychology in 1979 (Rosenthal). Following this, researchers in medicine found evidence of publication bias in medical clinical trials and observational studies (Dickersin, Chan, Chalmers, Sacks, & Smith, 1987; Mayes, Horwitz, & Feinstein, 1988), in biological research (Csada, James, & Espie, 1996), and in sociological research (Gerber & Malhrota, 2008). Because the incentives for publishing significant results exist in practically all fields, none is immune, and some have even argued that publication bias accounts for *most* published results in many fields (Greenberg, 2009; Ioannidis, 2005).

Despite attempts to address it, publication bias remains a difficult problem. A recent review of 91 separate meta-analyses in the field of psychology found that 41 percent of these detected some publication bias (and 25 percent of the meta-analyses provided evidence of serious bias) (Ferguson & Brannick, 2012). Moreover, a recent review of 383 meta-analyses using participant data in medicine found that 71 percent of these did not take appropriate measures to detect publication bias (Ahmed, Sutton, & Riley, 2012).

A widespread practice of data citation would provide a counter-incentive to the file-drawer problem for data sets. By encouraging the publication of data independent from a published article, data citation has the potential to contribute to studies of publication bias and citation bias and perhaps even reduce the prevalence of such bias.

7.5.4 Data reuse

In Chapter 6, we discussed some potential benefits of data sharing for research. Data sharing has been studied extensively and numerous field benefits of research data sharing have been identified (Berns et al., 1996; Committee on the Preservation of Geoscience Data and Collections & Committee on Earth Resources, 2002; Committee on Responsibilities of Authorship in the Biological Sciences, 2003; Feinberg et al., 1985; Sieber, 1991; Uhlir & Schröder, 2007).

Borgman (2012a) further places the rationales for data reuse into two broad categories. The first type of rationale involves simple reuse by combination of data from multiple sources, times, and places to ask new questions. The second type of rationale involves broader reuse that leads to the general advancement of research. This reuse is more likely to be interdisciplinary. A leading example of arguments for the broadest reuse are those related to the claimed "fourth paradigm" in which data and computational science combine to form a new branch of methodology (Bell, Hey, & Szalay, 2009; Gray, Liu, Nieto-Santisteban, Szalay, DeWitt, & Heber, 2005; Hey et al., 2009). Borgman notes that reuse is easier and more reliable when data are collected and processed systematically using common metadata, data structures, and ontologies.

To this observation we add that common data citation standards, which allow for uniform discovery, access, and verification of supporting evidence, also contribute to the easy and reliable integration of data from multiple sources. Moreover, as noted in Section 7.2, data use is costly to track and is underreported overall; thus a systematic set of data citation practices and infrastructure would improve the reliability, cost, and scale of research to evaluate the patterns, consequences, and value of data reuse.

7.6 Science metrics and science policy

Data citation, sharing, and curation also may help to reduce or mitigate the effects of biases, errors, and misconduct. Data citation offers the promise of facilitating research to better understand the "science of science" and, as a consequence, to improve science policy.

There has been increasing attention to reproducibility and integrity in scientific research. Controversies such as "Climate Gate" ("Clouds," 2010), the crumbling of high-profile personalized medical research at Duke University (Ince, 2011), and the finding of a long history of fraud by a prominent psychologist (Carey, 2011) have garnered media attention and resurrected public inquiry about the capacity of science to self-correct.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Within the research and scholarly publishing community, a steep rise in retraction rates has been a cause for concern. In the largest study of retractions to date, a detailed review of 2,047 retracted articles indexed in PubMed found a 10-fold increase in retractions since 1975, which far exceeds the increase in rate of indexed publications (Fang, Steen, & Casadevell, 2012). As importantly, roughly two-thirds of these retractions were attributable to misconduct while more than 43 percent of retractions were due to suspected data manipulation or falsification. Moreover, although retraction of a paper substantially decreases future citations (Furman, Jensen, & Murray, 2012), the influence of retracted papers lives on. Retracted papers continue to be cited with alarming frequency and without acknowledgement of their retracted status. Furman et al. (2012) estimate a rate of bad citation at less than 50 percent while another study has estimated preliminary rates of over 90 percent (Van Noorden, 2011).

Advocates of reproducible research argue that data fraud can be reduced or eliminated through the use of open data and open research reporting. For example, Stodden (2009) asserts that "Knowing work will be fully open to inspection in the future creates an incentive for researchers to do better, more careful, science now. Openness prevents any desire, even unconscious, to modify results in such a way that departs from the paper's underlying methodology" (p. 11).

Although the simple citation of data does not guarantee reproducibility, data citation has the potential to reduce misconduct related to data and its impact. First, as discussed in Chapter 4, data citations can and should contain information that enables later verification of the citing work against the original data. This increases accountability and allows for further confirmation, for example, that the data being made available in archives correspond to data used in the article. Second, citations to data could enhance the analysis and handling of retractions by enabling problems with the underlying evidence base to be addressed directly. Instead of simply retracting the publication based on suspect data, the suspect data and any article that relies on them could be traced as well and in turn be retracted.

7.6.1 Mapping the dark matter of science

Tracing the use of data through data citation has the potential not only to mitigate data related misconduct but also has the potential to illuminate the flow of knowledge in research. Building on progress in computation-intensive visualization and network analysis, the last ten years have seen rapid advances in the mapping of scientific knowledge and the relationships among science funding, scientists, and science output.[11] These "maps of science" can be used to support decision making for resource allocation (Boyack, Klavins, & Börner, 2005), to study co-authorship and collaboration patterns (Newman, 2004), and to chart emerging research frontiers and national and organization research profiles (Börner et al., 2002).

More generally, as Cronin's (2005) review of a wide range of information science, semiotic, and economics literature reveals, citation data (at least in certain subject fields) are sufficiently robust to serve as a reasonable proxy for properties such as research quality, the relatedness of research and scholars, and the impact of attention given to scholarly works.[12]

These maps of science and much of quantitative science metrics studies rely upon bibliographic citation databases. However, such citation and publication comes only at the end of a long chain of scholarly activity and yields an incomplete picture of science and science practice (Bollen, 2012). The result is that our understanding of science is incomplete.

Moreover, by relying solely on citations among articles, we miss a great deal of the relationships among scientific articles and lines of research – because articles may rely on the same evidence but not cite each other. In this way, data in its current state represents a type of the "dark matter" in our scientific universe that underlies the current visible evidentiary relationships among publications. Creation and management of data require substantial allocation of resources; it underlies the evidentiary relationships among publications—some of which may rely on the same data as evidence but not cite each other. Thus, establishing systematic indices of data citation promises to give us a richer understanding of science and science practice.

[11] See for a review, Börner (2010).

[12] See Cronin (2005), particularly p.125, Ch. 6, and Ch. 8, respectively.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

7.7 Conclusions

As Clay Shirky (2010, p.138) observes: "The problem with alchemy wasn't that the alchemists had failed to turn lead into gold—no one could do that. The problem, rather, was that the alchemists had not failed informatively." A robust infrastructure for data citation promises to substantially increase the frequency with which previously collected scientific evidence informs current research and future policy.

Data citation offers broad potential benefits for domain research and interdisciplinary research, in addition to generally improving our understanding of science. Yet achieving a robust and effective data citation infrastructure requires enabling research, coordination, and resources. Effort must be invested to build broad and coherent theoretical models and to translate and apply these to enable evaluation of the semantic equivalence of data, to strengthen theoretical and practical applications of data provenance, and to enable standardized approaches to managing the legal rights and responsibilities that adhere to research data. Evaluation research will be needed to assess the effectiveness and impact of the implementation of the data citation and to guide development of the infrastructure.

A robust data citation infrastructure, complemented by wider and more uniform data access, including free global access to bibliographic and data citations by means of open citation corpora, would catalyze domain research. Analysis that requires evidence from multiple measurement domains, scales, and populations or that requires extensive longitudinal data collection is challenging to fund and implement within the scope of a single research project. Uniform citation and access to research data would lower the barriers to this type of research. Similarly, data citation and sharing would make research that combines information across disciplines substantially easier to conduct. Furthermore, reviews of published work, such as Cochrane reviews and other systematic evidential reviews that synthesize the findings of previous studies, would be easier to conduct and would yield more informative results.

Moreover, while work within each specific discipline is vital, creating a robust data infrastructure cannot be accomplished within a single discipline or by relying solely on the skills and perspectives of a single discipline. Data citation and access infrastructure require research and implementation at the intersections of several disciplines: law, computer science and informatics, and policy and systems research. Figure 3 summarizes this.

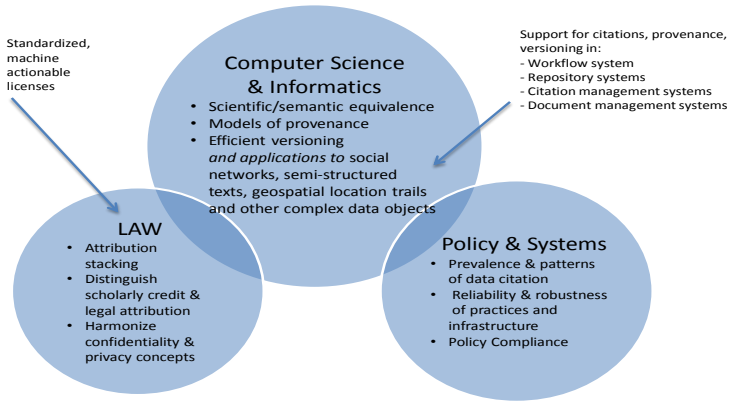

Figure 3. Primary research fields enabling data citation

In the last decade, substantial effort has been made and insights achieved in the "science of science" and in the "science of science policy." Much of this work is based upon analysis of scientific outputs and collaborations and uses bibliographic data describing research publications as its primary evidence base. It is primarily from bibliographic data (and secondarily from analysis of the texts of the publications themselves) that we have been able map the universe of scientific practice and results. If publications are the stars and planets of the scientific universe, data are the 'dark matter'—influential but largely unobserved in our mapping process. A robust and open data

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

citation infrastructure would render visible this dark matter, improving our understanding of the practice of science and improving science policy.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

ACKNOWLEDGEMENTS

We are grateful to the following for support of this project: the Alfred P. Sloan Foundation, grant number 2011-3-19 for the preponderance of funding for this project and the Committee on Data for Science and Technology (CODATA) and Microsoft Research for some additional funding.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with traditional procedures of peer review. The purpose of this independent review is to provide candid and critical comments that will assist the Task Group in making its published report as sound as possible. The review comments on the draft manuscript remained confidential to protect the integrity of the process.

We wish to thank the following individuals for their review of this report:

Wolfram Horstmann, Bodleian Libraries, University of Oxford; Paul Groth, VU University of Amsterdam; Alex Ball, University of Bath; Wim Hugo, South African Earth Observation System; Curt Tilmes, NASA Goddard Space Flight Center; David Shotton, University of Oxford; Natalia Manola, University of Athens; Tyng-Ruey Chuang, Academia Sinica (Taipei); and Monica Duke, University of Bath.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the content of the report, nor did they see the final draft before its release. Responsibility for the final content of this report rests entirely with the Task Group.

Numerous people devoted many months of effort to the writing of this report. The co-chairs of the CODATA-ICSTI Task Group on Data Citation Standards and Practices are Bonnie Carroll (until January 2013), Christine Borgman (from January 2013), Jan Brase, and Sarah Callaghan.

The lead authors of each chapter are: Sarah Callaghan (Chapters 1 and 2); Micah Altman, Daniel Cohen, and Puneet Kishor (Chapter 3); Bonnie Carroll (Chapter 4); Franciel Linares, (Chapter 5); Paul Uhlir (Chapter 6); Micah Altman (Chapter 7); and Yvonne Socha (Appendices and general editing).

Task Group members (in alphabetical order) are: Micah Altman, Elizabeth Arnaud, Todd Carpenter, Vishwas Chavan, Mark Hahnel, John Helly, Puneet Kishor, Jianhui LI, Franciel Azpurua Linares, Brian McMahon, Karen Morgenroth, Yasuhiro Murayama, Fiona Murphy, Giri Palanisami, Mark Parsons, Soren Roug, Helge Sagen, Eefke Smit, Martie van Deventer, Michael Witt, Koji Zettsu.

Consultants to the Task Group are: William Anderson, Daniel Cohen, Yvonne Socha, Melissa Turcios, and Lili Zhang. The Task Group Project Director is Paul Uhlir, and the CODATA Executive Committee Liaisons are Niv Ahituv (until April 2013) and Bonnie Carroll (since April 2013).

The short professional biographies of all these individuals are presented in Appendix A.

Christine Borgman, Jan Brase, Sarah Callaghan, and Bonnie Carroll

The Co-Chairs of the CODATA-ICSU Task Group on Data Citation Standards and Practice.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

REFERENCES

Ahmed, I., Sutton, A. J., & Riley, R. D. (2012) Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ*, 344. doi:10.1136/bmj.d7762

Alonso, W., & Starr, P. (Eds.) (1989) *The politics of numbers.* New York, NY: Russell Sage Foundation

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011) Public availability of published research data in high-impact journals. *PLoS ONE 6*(9), e24357. doi:10.1371/journal.pone.0024357

Altman, M. (2008) A fingerprint method for scientific data verification. In T. Sobh, (Ed.), *Proceedings of the International Conference on Systems Computing Sciences and Software Engineering, 2007* (pp. 311–316). New York, NY: Springer Netherlands. doi:10.1007/978-1-4020-8741-7_57

Altman, M. (2012) Data citation in the Dataverse Network. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 99-106). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (2009) Digital preservation through archival collaboration: The data preservation alliance for the social sciences. *American Archivist, 72*(*1*), 170–184. Retrieved July 30, 2013 from the WWW: http://archivists.metapress.com/content/EU7252LHNRP7H188

Altman, M., Andreev, L., Diggory, M., King, G., Kiskis, D. L., Kolster, E., . . ., & Verba, S. (2001) A digital library for the dissemination and replication of quantitative social science research. *Social Science Computer Review, 19*(4), 458–470. Retrieved July 30, 2013 from the WWW: http://www.box.net/shared/d3cf8u0gtyml2nqq3u2f

Altman, M., Gill, J., & McDonald, M. P. (2003) *Numerical issues in statistical computing for the social scientist.* Hoboken, NJ: John Wiley and Sons.

Altman, M. & King, G. (2007) A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine 13*(3/4). Retrieved July 30, 2013 from the WWW: http://www.dlib.org/dlib/march07/altman/03altman.html

Anderson, R. G., Greene, W. H., McCullough, B. D., & Vinod, H. D. (2008) The role of data/code archives in the future of economic research. *Journal of Economic Methodology, 15*(1), 99-119. doi:10.1080/13501780801915574

Atkins, D., Droegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., Messerschmitt, D., . . ., & Wright, M. (2003) Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure (Report No. cise051203). Retrieved July 30, 2013 from the WWW: http://www.nsf.gov/cise/sci/reports/atkins.pdf

Australian National Data Service (2011) Data citation awareness guide. Retrieved July 30, 2013 from the WWW: http://ands.org.au/guides/data-citation-awareness.html

Availability of supporting data (2013) BioMed Central. Retrieved July 30, 2013 from the WWW: http://www.biomedcentral.com/about/supportingdata

Ball, A. & Duke, M. (2012) How to cite data sets and link to publications. *DCC how-to guides.* Edinburgh: Digital Curation Centre. Retrieved July 30, 2013 from the WWW: http://www.dcc.ac.uk/resources/how-guides

Bell, G., Hey, T., & Szalay, A. (2009) Beyond the data deluge. *Science, 323*(5919), 1297-1298. doi:10.1126/science.1170411

Berns, K. I., Bond, E. C., & Manning, F. J. (1996) *Resource sharing in biomedical research*. Washington, D.C.: National Academy Press.

Boettiger, C. (2012) Citing lab notebook entries. [Web log post]. Retrieved July 30, 2013 from the WWW: http://www.carlboettiger.info/2012/11/23/citing-lab-notebook-entries.html

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Bollen, J. (2012) Attribution and credit: Beyond print and citations. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 15-22). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009) A principal component analysis of 39 scientific impact measures. *PLoS ONE, 4*(6), 6022. doi:10.1371/journal.pone.0006022

Bollen, J., Van de Sompel, H., Smith, J. A., & Luce, R. (2005) Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management, 41*(6), 1419-1440. Retrieved July 30, 2013 from the WWW: http://arxiv.org/pdf/cs/0503007v1.pdf

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

Borgman, C. L. (2012a) The conundrum of research sharing. *Journal of the American Society for Information Science and Technology*, 63(6), 1059-1078. doi:10.1002/asi.22634

Borgman, C. (2012b) Why are the attribution and citation of scientific data important? In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 1-10). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Börner, K. (2010) *An atlas of science*. Cambridge, MA: MIT Press

Börner, K, Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., Larivière, V., & Boyack, K. W. (2002) Design and update of a classification system: The USCD map of science. *PLoSONE, 7*(7). doi:10.1371/journal.pone.0039464

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., O'Neill, O., . . ., & Walport, M. (2012) *Science as an open enterprise: The Royal Society Science Policy Centre report*. Retrieved July 30, 2013 from the WWW: http://royalsociety.org/policy/projects/science-public-enterprise/report/

Bourne, P. (2005) Will a biological database be different from a biological journal? *PLoS Computational Biology, 1*(3), e34. doi:10.1371/journal.pcbi.0010034

Boyack, K. W., Klavans, R., & Börner, K. (2005) Mapping the backbone of science. *Scientometrics, 64*(3), 351-374. Retrieved July 30, 2013 from the WWW: http://scimaps.org/exhibit/docs/05-boyack.pdf

Brase, J. (2009, October) DataCite: A global registration agency for research data. Paper presented at ILDS Conference, Hannover, Germany. Retrieved July 30, 2013 from the WWW: http://www.ilds2009.eu/fileadmin/user_upload/Full_text/DataCite_Brase_COINFO.pdf

Buckheit, J. & Donoho, D. L. (1995) WaveLab and reproducible research. In A. Antoniadis & G. Oppenheim, (Eds.), *Wavelets and Statistics* (pp. 55-81). Berlin: Springer-Verlag.

Buneman, P., Khanna, S. & Wang-Chiew, T. (2001)Why and where: A characterization of data provenance. In *Database Theory – Proceedings of the ICDT 2001*, 316-330. Berlin: Springer-Verlag.

Buneman, P. (2006) How to cite curated databases and how to make them citable. *Proceedings of the 18th International Conference on Scientific and Statistical Database Management* (pp. 195-203). Los Alamitos, CA: IEEE Computer Society. Retrieved July 30, 2013 from the WWW: http://homepages.inf.ed.ac.uk/opb/homepagefiles/harmarnew.pdf

Butler, D. (2012) Scientists: Your number is up: ORCID scheme will give researchers unique identifiers to improve tracking of publications. *Nature*, *485*, 564. doi:10.1038/485564a

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., . . ., & Wright, D. (2012) Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. *International Journal of Digital Curation, 7*(1), 107-113. doi:10.2218/ijdc.v7i1.218

Cano, P., Batlle, E., Kalker, T., & Haitsma, J. (2005) A review of audio fingerprinting. *The Journal of VLSI Signal Processing, 41*(3), 271-284. doi:10.1007/s11265-005-4151-3

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Carey, B. (2011, November 2) Fraud case seen as a red flag for psychology research. *The New York Times*, p. A3. Retrieved July 30, 2013 from the WWW: http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html?_r=0

Case, D. O. & Higgins, G. M. (2000) How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51, 635–645. doi:10.1002/(SICI)1097-4571(2000)51:7<635::AID-ASI6>3.0.CO;2-H

Chavan, V. (2012a) Data citation mechanism and service for scientific data: Defining a framework for biodiversity data publishers. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 113-116). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Chavan, V. (2012b) *Recommended practices for citation of data published through the GBIF network*. Copenhagen: GBIF Secretariat. Retrieved July 30, 2013 from the WWW: http://links.gbif.org/gbif_best_practice_data_citation_en_v1

The clouds of unknowing (2010, March 18) [Briefing] *The Economist*. Retrieved July 30, 2013 from the WWW: http://www.economist.com/node/15719298

Commission in Physical Sciences, Mathematics, and Applications (1997) *Bits of power: Issues in global access to scientific data*. Washington, D.C.: National Academy Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/openbook.php?record_id=5504&page=R1

Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest (1999) *A question of balance: Private rights and the public interest in scientific and technical databases*. Washington, D.C.: National Academy Press. Retrieved July 30, 2013 from the WWW: http://books.nap.edu/catalog.php?record_id=9692

Committee on the Preservation of Geoscience Data and Collections & Committee on Earth Resources (2002) *Geoscience data and collections: National resources in peril.* Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=10348

Committee on Responsibilities of Authorship in the Biological Sciences. (2003). *Sharing publication-related data and materials: Responsibilities of authorship in the life sciences*. Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=10613

Consultative Committee for Space Data Systems (2002) *Reference model for an open archival information system (OAIS)*. Retrieved July 30, 2013 from the WWW: http://ddp.nist.gov/refs/oais.pdf

Credit Where Credit Is Due [Editorial]. (2009, Dec 17) *Nature* 462: 825. doi:10.1038/462825

Cronin, B. (1984) *The citation process: The role and significance of citations in scientific publication*. London, United Kingdom: Taylor Graham.

Cronin, B. (2005) *The hand of science: Academic writing and its rewards*. Lanham, MD: Scarecrow Press.

Csada, R. D., James, P. C., & Espie, R. H. M. (1996) The “file drawer problem” of non-significant results: Does it apply to biological research? *Oikos*, *76*(3), 591–593.

Davidson, S. B. & Freire, J. (2008). Provenance and scientific workflows: challenges and opportunities. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. Retrieved July 30, 2013 from the WWW: http://bigdata.poly.edu/~juliana/pub/freire-tutorial-sigmod2008.pdf

Deeks, J. J., Higgins, J., & Altman, D. G. (2008) Analyzing data and undertaking meta-analyses. In J. P. T. Higgins & S. Green, (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*, (pp. 243-296). Hoboken, NJ: Wiley.

Deelman, E. & Gil, Y. (2006, May) *Final report on workshop on the challenges of scientific workflows*. Workshop presented for the National Science Foundation, Arlington, VA. Retrieved July 30, 2013 from the WWW: https://confluence.pegasus.isi.edu/download/attachments/2031787/NSFWorkflow-Final.pdf?version=1&modificationDate=1254437518000&api=v2

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Desrosières, A. & Naish, C. (2002) *Politique des grands nombres: Histoire de la raison statistique*. Cambridge, MA: Harvard University Press.

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986) Replication in empirical economics: The journal of money, credit and banking project. *American Economic Review, 76*, 587-603.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith H. Jr. (1987) Publication bias and clinical trials. *Controlled Clinical Trials,* 8(4), 343-353.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007) Understanding infrastructure: Dynamics, tensions, and design. Retrieved July 30, 2013 from the WWW: http://hdl.handle.net/2027.42/49353

Engineering and Physical Sciences Research Council (2011) *ESPRC policy framework on research data*. Retrieved July 30, 2013 from the WWW: http://www.epsrc.ac.uk/about/standards/researchdata/Pages/policyframework.aspx

European Commission (2011a) Digital agenda: Turning government data into gold. [Press release]. Retrieved July 30, 2013 from the WWW: http://europa.eu/rapid/press-release_IP-11-1524_en.htm

European Commission (2011b) Open data portals. Retrieved July 30, 2013 from the WWW: http://ec.europa.eu/information_society/policy/psi/open_data_portal/index_en.htm

Fang, F. C., Steen, R. G., & Casadevall, A. (2012) Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences in the United States of America, 109*(42), 17028-17033. doi:10.1073/pnas.1212247109

Federation of Earth Science Information Partners (2011) Interagency data stewardship/principles. Retrieved July 20, 2013 from the ESIP wiki: http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Principles

Feinberg, S.E., Martin, M.E., & Straf, M.L. (1985) *Sharing research data*. Washington, D.C.: National Academy Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=2033

Fenner, M. (2011, April) On microattribution. [Web log] Retrieved July 30, 2013 from the WWW: http://blogs.plos.org/mfenner/2011/08/28/on-microattribution/

Ferguson, C. J. & Brannick, M. T. (2012) Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120-128. doi:10.1037/a0024445

Figshare partners with Open Access mega journal publisher PLOS (2013, Jan 30) [Web log post]. Retrieved July 30, 2013 from the WWW: http://figshare.com/blog/figshare_partners_with_Open_Access_mega_journal_publisher_PLOS/68

Finch, J. (2012) *Accessibility, sustainability, excellence: How to expand access to research publications: Report of the working group on expanding access to published research findings*. Retrieved July 30, 2013 from the WWW: http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf

Freire, J., Silvia, C. T., Callahan, S. P., Santos, S., Scheidegger, C. E., & Vo, H. T. (2006) Managing rapidly-evolving scientific workflows. *Proceedings from IPAW '06: International Provenance and Annotation Workshop, 4145,* 10-18. Berlin, Germany: Springer-Verlag.

Furman, J. L., Jensen, K., & Murray, F. (2012) Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy 41*(2), 276-290. doi:http://dx.doi.org/10.1016/j.respol.2011.11.001

Gentleman, R. & Temple Lang, D. (2007) Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics 16*(1), 1-23. Retrieved July 30, 2013 from the WWW: http://biostats.bepress.com/bioconductor/paper2

Gerber, A. S. & Malhotra, N. (2008) Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research, 37*(1), 3-30.

Glaser, W.A. & Bisco, R.L. (1966) Plans of the council of social science data Archives. *Social Science Information*, *5*(4), 71-96.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Giardine, B., Borg, J., Higgs, D. R., Peterson, K. R., Philipsen, S., Maglott, D., . . ., & Patrinos, G. P. (2011) Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nature Genetics*, *43*, 295-301. doi:10.1038/ng.785

Gold, A. (2007) Cyberinfrastructure, data, and libraries, part 1. *D-Lib Magazine*, *13*(9/10). Retrieved July 30, 2013 from the WWW: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html

Gonzales, E., Zhang X., Akahoshi Y., Murayama Y., & Zettsu K. (2012, October) Data Citation Wiki for Harnessing Collective Intelligence on Document-to-Data Associations to Interdisciplinary Data Access. Paper presented at the 23rd International CODATA Conference, Taipei, China.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2005) Scientific data management in the coming decade. *ACM SIGMOD Record*, *34*(4), 34-41. Retrieved July 30, 2013 from the WWW: http://arxiv.org/ftp/cs/papers/0502/0502008.pdf

Green, T. (2009) We need publishing standards for data sets and data tables. *OECD Publishing White Paper*, *22*. doi:10.1787/787355886123

Greenberg, J. (2005) Understanding metadata and metadata schemes. *Cataloging & Classification Quarterly*, *40*(3/4), 20. doi:10.1300/J104v40n03_02

Greenberg , S. A. (2009) How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 339, b2680. doi:10.1136/bmj.b2680

Griffiths, A. (2009) The publication of research data: Researcher attitudes and behaviors. *The International Journal of Digital Curation*, *4*(1), 46-56. Retrieved July 30, 2013 from the WWW: http://www.ijdc.net/index.php/ijdc/article/view/101/76

Groth, P. (2012) Maintaining the scholarly value chain: Authenticity, provenance, and trust. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop*, (pp. 31-42). Washington, D.C.: National Academies Press, 2012. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Guide to publication policies of the Nature Journals (2013) Retrieved July 30, 2013 from the WWW: http://www.nature.com/authors/gta.pdf

Gutmann, M., Abrahamson, M., Adams, M., Altman, M., Arms, C., Bollen, K., . . ., & Young, C. (2009) From preserving the past to preserving the future: The data-PASS project and the challenges of preserving digital social science data. *Library Trends*, *57*(3), 315-337. Retrieved July 30, 2013 from the WWW: http://gking.harvard.edu/files/gking/files/GutAbrAda09.pdf

Haak, L. L.,Fenner, M.,Paglione, L.,Pentz, E., & Ratner, H. (2012) ORCID: A system to uniquely identify researchers. *Learned Publishing*, *25*, 259-264. Retrieved July 30, 2013 from WWW: http://www.ingentaconnect.com/content/alpsp/lp/2012/00000025/00000004/art00004

Hamermesh, D. (August 2007) Viewpoint: Replication in economics. *Canadian Journal of Economics 40*(3), 715-733. Retrieved July 30, 2013 from WWW: https://webspace.utexas.edu/hamermes/www/CJE82007.pdf

Harley, D. (2012) Issues of time, credit, and peer review. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 81-94). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Hedstrom, M. & Lee, C. (2002) Significant properties of digital objects: definitions, applications, implications. *Proceedings of the DLM-Forum: Parallel session 3* (pp. 218-113). Retrieved July 30, 2013 from the WWW: http://www.ils.unc.edu/callee/sigprops_dlm2002.pdf

Hedstrom, M., Niu, J., & Marz, K. (2008) Incentives for data producers to create "archive/ready" data: Implications for archives and records management. *Proceedings from the Society of American Archivists Research Forum.* San Francisco, CA: SAA. Retrieved July 30, 2013 from the WWW: http://files.archivists.org/conference/2008/researchforum/M-HedstromJ-Niu-SAA-ResearchPaper-2008.pdf

Helly, J. (1998) New concepts of publication. *Nature*, *393*, 107. doi:10.1038/30086

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Helly, J., Elvins, T., Sutton, D., & Martinez, D. (1999) A method for interoperable digital libraries and data repositories. *Future Generation Computer Systems, 16*(1), 21–28. Retrieved July 30, 2013 from the WWW: http://www.sdsc.edu/~hellyj/papers/FGCS_jjh01.pdf

Helly, J., Elvins, T. T., Sutton, D., Martinez, D., Miller, S., Pickett, S., & Ellison, A. M. (2002) Controlled publication of digital scientific data. *Communications of the ACM, 45*(5), 97–101. Retrieved July 30, 2013 from the WWW: http://www.sdsc.edu/~hellyj/papers/CACM2002.pdf

Hey, T., Tansley, S., & Tolle, K., (Eds.) (2009) *The fourth paradigm: Data intensive scientific discovery*. Redmond, WA: Microsoft Research. Retrieved July 30, 2013 from the WWW: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Hockx-Yu, H. & Knight, G. (2008) What to preserve?:Significant properties of digital objects. *International Journal of Digital Curation, 3*(1), 141-153. doi:10.2218/ijdc.v2i1.49

Humphreys, M., Sanchez de la Sierra, R., & van der Windt, P. (2013) Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis, 21*(1), 1-20.

Ince, D. (2011) The Duke University scandal—what can be done? *Significance, 8*(3), 113-115.

Interagency Working Group on Digital Data (2009) Harnessing the power of digital data for science and society. Retrieved July 20, 2013 from the IT Law wiki http://itlaw.wikia.com/wiki/Interagency_Working_Group_on_Digital_Data

International Council for Science (2004) *Scientific data and information: A report of the CSPR assessment panel*. Retrieved July 30, 2013 from the WWW: http://www.icsu.org/publications/reports-and-reviews/priority-area-assessment-on-scientific-data-and-information-2004/PAA_Data_and_Information_report.pdf

Inter-University Consortium for Political and Social Research (2011) *About the bibliography of data-related literature*. Retrieved July 30, 2013 from the WWW: http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/citations/methodology.html

Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLoS Medicine, 2*(8), 124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. & Khoury, M. J. (2011) Improving validation in "Omics" research. *Science, 334*(6060), 1230-1232. doi:10.1126/science.1211811

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011) Again, and again, and again… *Science, 334*(6060), 1225. doi:10.1126/science.334.6060.1225

Journal Research Data Policy Bank (JoRD) (2013, February 1) A rather long post, but quite a brief summary. [Web log post]. Retrieved July 30, 2013 from the WWW: http://jordproject.wordpress.com/category/project-information/

Jørgensen, A. W., Hilden, J., & Gøtzsche, P. C. (2006) Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. *BMJ, 333*(7572), 782. doi:http://dx.doi.org/10.1136/bmj.38973.444699.0B

King, G. (1995) Replication, replication. *PS: Political Science and Politics, 28*, 443-499. Retrieved July 30, 2013 from the WWW: http://gking.harvard.edu/files/gking/files/replication.pdf

King, G. (2007) An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods and Research, 36*(2), 173-179. doi:10.1177/0049124107306660

Kirlew, P. (2011) Life science data repositories in the publications of scientists and librarians. *Issues in Science and Technology Librarianship, 65*. Retrieved July 30, 2013 from the WWW: http://www.istl.org/11-spring/refereed1.html

Knuth, D. E. (1992) *Literate programming*. (CLSI Lecture Notes 27). Stanford, California: Center for the Study of Language and Information.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Kotarski, R., Reilly, S., Schrimpf, S., Smit, E., & Walshe, K. (2012, June21) Report on best practices for citability of data and on evolving roles in scholarly communication. Retrieved July 30, 2013 from the WWW: http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-ReportBestPracticesCitabilityDataEvolvingRolesScholarlyCommunication.pdf

Kunze, J., Cruse, T., Hu, R., Abrams, S., Hastings, K., Mitchell, C., & Schiff, L. (2011) Practice, trends, and recommendations in technical appendix usage for selected data-intensive disciplines. Retrieved July 30, 2013 from the WWW: http://www.cdlib.org/services/uc3/docs/dax.pdf

Lawrence, B., Pepler, S., Jones, C., Matthews, B., McGarva, G., & Coles, S. (2007, May) Linking data publications in the environmental sciences: Claddier Project Workshop, Chilworth, Southampton, United Kingdom. Retrieved July 30, 2013 from the WWW: http://eprints.soton.ac.uk/46207/

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . ., & Alsyne, M. V. (2009) Computational social science. *Science*, *323*(5915), 721-723. doi:10.1126/science.1167742

Leisch, F. & Rossini, A. J. (2003) Reproducible statistical research. *Chance*, *16*(2), 46-50.

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*, *10*(5062). Retrieved July 30, 2013 from the WWW: http://www.istl.org/11-fall/article1.html

Matthews, B., McIlwrath, B., Giaretta, D., & Conway, E. (2008) *The significant properties of software: A study*. Retrieved July 30, 2013 from the WWW: http://www.jisc.ac.uk/media/documents/programmes/preservation/spsoftware_report_redacted.pdf

Maunsell, J. (2010, August 11) Announcement regarding supplemental material. *The Journal of Neuroscience*, *30*(32), 10599-10600. Retrieved July 30, 2013 from the WWW: http://www.jneurosci.org/content/30/32/10599.full

Mayes, L. C., Horwitz, R. I., & Feinstein, A. R. (1988) A collection of 56 topics with contradictory results in case-control research. *International Journal of Epidemiology*, *17*, 680–685.

McCullough, B. D. (2007) Got replicability? The journal of money, credit and banking archive. *Econ Journal Watch*, *4*(3), 326-337. Retrieved July 30, 2013 from the WWW: http://econjwatch.org/articles/got-replicability-the-journal-of-money-credit-and-banking-archive

McCollough, B. D. (2009) Open access economics journals and the market for reproducible economic research. *Economic Analysis and Policy*, *39*(1), 117-126.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2008) Do economics journal archives promote replicable research? *Canadian Journal of Economics*, *41*(4), 1406-1420. Retrieved July 30, 2013 from the WWW: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=931231

Mendes, P. N., Jakob, M., Garcia-Silva, A., & Bizer, C. (2011) DBedia spotlight: Shedding light on the web of documents. *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics), 2011*. doi:10.1145/2063518.2063519

Mooney, H. (2011) Citing data sources in the social sciences: Do authors do it? *Learned Publishing*, *24*(2), 99-108. Retrieved July 30, 2013 from the WWW: http://www.ingentaconnect.com/content/alpsp/lp/2011/00000024/00000002/art00004

Moreau, L. & Missier, P. (2013) PROV-DM: The PROV data model. Retrieved July 30, 2013 from the WWW: http://www.w3.org/TR/2013/REC-prov-dm-20130430/

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., . . ., & Bussche J. V. (2011) The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, *27*(6), 743-756. Retrieved July 30, 2013 from the WWW: http://eprints.soton.ac.uk/268332/1/opm.pdf

National Information Standards Organization (2004) *Understanding Metadata*. Bethesda, MD: NISO Press. Retrieved July 30, 2013 from the WWW: http://www.niso.org/publications/press/UnderstandingMetadata.pdf

National Information Standards Organization (2013) *Recommended practices for online supplemental journal article materials*. (Report No. NISO RP-15-201x). Retrieved July 30, 2013 from the WWW: http://www.niso.org/apps/group_public/download.php/10055/RP-15-2013_Supplemental_Materials.pdf

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

National Science Board (2005) *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Retrieved July 30, 2013 from the WWW: http://www.nsf.gov/pubs/2005/nsb0540/

National Science Foundation Cyberinfrastructure Council (2007) *Cyberinfrastructure Vision for 21st Century Discovery*. Retrieved July 30, 2013 from the WWW: http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Neuliep, J. W. & Crandall, R. (1993) Reviewer bias against replication research. *Journal of Social Behavior & Personality, 8*, 21-29.

Newman, M. E. J. (2004) Coauthorship networks and patterns of scientific collaboration. *Protocols of the National Academy of Science, 101*, 5200-5205. doi:10.1073/pnas.0307545100

Organization for Economic Cooperation and Development (2007) *OECD principles and guidelines for access to research data and public funding.* Retrieved July 30, 2013 from the WWW: http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

Patrinos, G. P., Cooper, D. N., van Mulligan, E., Gkantouna, V., Tzimas, G., Tatum, Z., . . ., & Mons, B. (2012) Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Human Mutation, 33*(11), 1503-1512. doi:10.1002/humu.22144

Paving the way to an open scientific information space: OpenAIREplus – linking peer-reviewed literature to associated data [press release] (2011) Retrieved July 30, 2013 from the WWW: https://www.openaire.eu/en/component/content/article/76-highlights/326-openaireplus-press-release

Pearson, S. (2012) Three legal mechanisms for sharing data. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 71-76). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Peng, R. D. (2011) Reproducible research in computational science. *Science, 334* (6060), 1226-1227. doi:10.1126/science.1213847

Peng, Z., Cheng, Y., Tan, B., Lin Kang, Tian, Z., Zhu, Y., . . ., & Wang, J. (2012) Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature Biotechnology, 30*, 253–260. doi:10.1038/nbt.2122

Peroni, S. & Shotton, D. (2012) FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web, 17*, 33-34. doi:10.1016/j.websem.2012.08.001

Pienta, A. (2006) LEADS database identifies at-risk legacy studies. *ICPSR Bulletin, 27*(1), 3-8. Retrieved July 30, 2013 from the WWW: http://www.icpsr.umich.edu/files/membership/publications/bulletin/2006-Q3.pdf

Piwowar, H. (2010) Tracking dataset citations using common citation tracking tools doesn't work. [Web log post]. Retrieved July 30, 2013 from the WWW: http://researchremix.wordpress.com/2010/11/09/tracking-dataset-citations-using-common-citation-tracking-tools-doesnt-work/

Piwowar, H. (2013) Altmetrics: Value all research products. *Nature, 493*(159). doi:10.1038/493159a

Piwowar, H. (forthcoming article) Making data count: Tracking impact through the scholarly literature and beyond.

Piwowar, H. & Vision, T.J. (2013) Data reuse and the open data citation advantage. *PeerJ PrePrints. 1*:e1v1. doi:10.7287/peerj.preprints.1

Policy on the management of research data and record (2012) University of Oxford. Retrieved July 30, 2013 from the WWW: http://www.admin.ox.ac.uk/rdm/managedata/policy/

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved July 30, 2013 from the WWW: http://altmetrics.org/manifesto/

Renear, A. H., Sacchi, S., & Wickett. K. M. (2010) Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology, 47*(1), 1-4. Retrieved July 30, 2013 from the WWW: http://mail.asist.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/240_Final_Submission.pdf

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Research Councils United Kingdom (2011) Common principles on data policy. Retrieved July 30, 2013 from the WWW: http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

Research data management policy (2011) University of Edinburgh. Retrieved July 30, 2013 from the WWW: http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy

Rosenthal, R. (1979) The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rust, G. & Bide, M. (2008) *The <indecs> metadata framework, principles, model and data dictionary*. (Report No. WPA1a.066.2.0). Retrieved July 30, 2013 from the WWW: http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

Ryan, M. J. (2011) Replication in field biology: The case of the frog-eating bat. *Science, 334*(6060), 1229-1230. doi:10.1126/science.1214532

Santer, B. D., Wigley, T. M. L., & Taylor, K. E. (2011) The reproducibility of observational estimates of surface and atmospheric temperature change. *Science, 334*(6060), 1232-1233. doi:10.1126/science.1216273

Schwab, M., Karrenbach, M., & Claerbout, J. (2000) Making scientific computations reproducible. *Computing in Science and Engineering, 2*, 61-67.

Shirky, C. (2010) *Cognitive surplus: How technology makes consumers into collaborators*. New York: Penguin.

Shotton, D. (2010) CiTO, the citation typing ontology. *Journal of Biomedical Semantics* 1 (Suppl. 1): S6. doi:10.1186/2041-1480-1-S1-S6

Shotton, D. (2011 August 4) The plate tectonics of research data publication [Web log post]. Retrieved July 30, 2013 from the WWW: http://semanticpublishing.wordpress.com/2011/08/04/the-plate-tectonics-of-research-data-publication/

Sieber, J. (1991) *Sharing social science data*. Newbury Park, CA: Sage Publications.

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

Simon, S. D. & Lesage, J. P. (1989) Assessing the accuracy of ANOVA calculations in statistical software. *Computational Statistics & Data Analysis, 8*(3), 325–332. Retrieved July 30, 2013 from the WWW: http://www.sciencedirect.com/science/article/pii/0167947389900480

Smith, M. (2012) Institutional perspectives on credit systems for research data. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 77-80). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Spengler, S. (2012) Data citation and attribution: A funder's perspective. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 177-188). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Starr, J. & Gastl, A. (2011). IsCitedBy: A metadata scheme for datacite. *D-Lib Magazine, 17*(½). doi:10.1045/january2011-starr

Staudigel, H., Helly, J., Koppers, A., Shaw, H. F., McDonough, W. F., Hofmann, A. W., . . ., & Zindler, A. (2003) Electronic data publication in geochemistry. *Geochemistry, Geophysics, Geoscience, 4*(3). doi:10.1029/2002GC000314

Steen, R. G. (2011) Retractions in the scientific literature: Is the incidence of research fraud increasing? *J Med Ethics, 37*(2), 113-17. doi:10.1136/jme.2010.038125

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995) Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 439*, 108–112. doi:10.2307/2684823

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Stewart, L. A., Tierney, J. F., & Clarke, M. (2008) Reviews of individual patient data. In J. P. Higgins & S. Green, (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*, (pp. 547-588). Hoboken, NJ: John Wiley and Sons.

Stodden, V. (2009) The legal framework for reproducible scientific research: Licensing and copyright. *Computing in Science and Engineering, 11*(1), 35-40. doi:http://dx.doi.org/10.1109/MCSE.2009.19

Tilmes, C., Yesha, Y., & Halem, M. (2011).Distinguishing provenance equivalence of earth science data. *Procedia Computer Science, 4*, 548-557. Retrieved July 30, 2013 from the WWW: http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20110015226_2011016012.pdf

Tomasello, M. & Call, J. (2011) Methodological challenges in the study of primate cognition. *Science, 334*(6060), 1227-1228. doi:10.1126/science.1213443

Towne, L., Wise, L. L., & Winters, T. M. (2004) *Advancing science in education*. Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=11112

Uhlir, P.F. (Ed.) (2012) *For attribution: Developing data attribution and citation practices and standards: Summary of an international workshop.* Washington, D.C: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Uhlir, P.F. (2006). The emerging role of open repositories as a fundamental component of the public research infrastructure. In G. Sica, (Ed.), *Open access: Open problems*. Monza, Italy: Polimetrica

Uhlir, P.F. & Schröder, P. (2007) Open data for global research. *Data Science Journal, 6*(36-53). Retrieved July 30, 2013 from the WWW: https://www.jstage.jst.go.jp/article/dsj/6/0/6_0_OD36/_pdf

Van de Sompel, H. (2012) Data citation - technical issues - identification. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop*, (pp. 23-30). Washington, D.C.: National Academies Press, 2012. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Van Leunen, M. (1992) *A handbook for scholars*. New York, NY: Oxford University Press.

Van Noorden, R. (2011) Science publishing: The trouble with retractions. *Nature, 478*, 26-28. doi:10.1038/478026a

Vardigan, M. (2012) Data citation for the social sciences. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop*, (pp. 55-58). Washington, D.C.: National Academies Press, 2012. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009) Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274-290. Retrieved July 30, 2013 from the WWW: http://www.pashler.com/Articles/Vul_etal_2008inpress.pdf

Wenger, E., White, N., & Smith, J. (2009) *Digital habitats: Stewarding technology for communities*. Portland, OR: CPsquare

White, H. D. (1982) Citation analysis of data file use. *Library Trends, 31*(3), 467-477. Retrieved July 30, 2013 from the WWW: https://www.ideals.illinois.edu/bitstream/handle/2142/7222/librarytrendsv30i3l_opt.pdf?sequence=1

Wiehua, W. (2012) TR-REC-069 specification referenced scientific data. Retrieved July 30, 2013 from the WWW: http://www.nsdc.cn/pronsdchtml/1.compservice.standards/pages/3440.html

Wilson, B. (2012) Data citation and data attribution: A view from the data center perspective. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 147-149). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

Witt, M. (2012) Roles for libraries in data citation. In P. F. Uhlir, (Ed.), *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop* (pp. 151-156). Washington, D.C.: National Academies Press. Retrieved July 30, 2013 from the WWW: http://www.nap.edu/catalog.php?record_id=13564

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

World Economic Forum (2012) *Big data, big impact: New possibilities for economic development.* Retrieved July 30, 2013 from the WWW: http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf

Wynholds, L. (2011) Linking to scientific data: Identity problems of unruly and poorly bounded objects. *Interantional Journal of Digital Curation, 6*, 215. doi:10.2218/ijdc.v6i1.183

Yanhua, Z. & Lianglin, H. (2012) Analyzing the Influence of Scientific Database Based on the Third-party Cited Marks. *China Science & Technology Resources Review*, *44*, 6, 17-22.

Zettsu, K., Gonzales, E., Ong, B. T., & Murayama, Y. (2012, October) Cross-Database Search for Interdisciplinary Use of Large-Scale, Multi-Domain and Heterogeneous Databases. Paper presented at the 23rd International CODATA Conference, Taipei, China.

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

APPENDIX A

Task Group members and their professional biographies

Co-Chairs

Co-Chair, Christine Borgman (US CODATA), Professor and Presidential Chair, University of California, Los Angeles; www.linkedin.com/pub/christine-borgman/a/6a3/36

Co-Chair, Jan Brase (Director, DataCite, and ICSTI representative), Technische Informations Bibliothek (TIB)/German National Library of Science and Technology, Germany; http://de.linkedin.com/pub/jan-brase/8/a4a/771

Co-Chair, Sarah Callaghan (UK CODATA), The NCAS British Atmospheric Data Centre, STFC Rutherford Appleton Laboratory, UK; http://uk.linkedin.com/pub/sarah-callaghan/1b/9a4/726

Membership List

Micah Altman, Director of Research, Head/Scientist, Program on Information Research MIT Libraries Massachusetts Institute of Technology, USA; http://micahaltman.com

Elizabeth Arnaud, Project Coordinator, Understanding and Managing Biodiversity Programme, Bioversity International, ITALY; http://fr.linkedin.com/pub/elizabeth-arnaud/18/618/42

Todd Carpenter, Executive Director, National Information Standards Organization, USA; www.linkedin.com/in/toddacarpenter

Vishwas Chavan, Senior Program Officer for DIGIT, Global Biodiversity Information Facility, Denmark; http://dk.linkedin.com/in/vishwaschavan

Mark Hahnel, Figshare, Digital Science, United Kingdom; http://www.linkedin.com/pub/mark-hahnel/1a/a36/355

John Helly, Scripps Institution of Oceanography and San Diego Supercomputer Center, University of California, USA; www.linkedin.com/pub/john-helly/0/688/4a4

Puneet Kishor, Creative Commons, USA; http://creativecommons.org/staff#puneetkishor

Jianhui LI, Director, Scientific Data Center, Computer Network Information Center, Chinese Academy of Sciences, China; http://sourcedb.cnic.cas.cn/yw/people/200908/t20090817_2404471.html

Franciel Azpurua Linares, Technical Director, Scientific Information Management & Technology Program, Information International Associates, USA; www.linkedin.com/in/francielazpurualinares

Brian McMahon, International Union of Crystallography, UK; http://bit.ly/ZuBIfD

Karen Morgenroth, National Research Council Canada, Canada Institute for Scientific and Technical Information, Canada; http://ca.linkedin.com/pub/karen-morgenroth/12/170/a68

Yasuhiro Murayama, Director, Integrated Science Data System Research Laboratory, National Institute of Information and Communications Technology, Japan; http://www.researchgate.net/profile/Yasuhiro_Murayama/

Fiona Murphy, Executive Journals Editor Wiley Europe Ltd, UK; http://uk.linkedin.com/pub/fiona-murphy/23/357/745

Giri Palanisamy ARM Archive Group Environmental Sciences Division Oak Ridge National Laboratory Oak Ridge, USA; www.linkedin.com/pub/giri-palanisamy/a/578/69b

Mark Parsons, Research Data Alliance/U.S. Center for a Digital Society Rensselaer Polytechnic Institute, USA; www.linkedin.com/pub/mark-parsons/8/a06/719

Soren Roug, EEA Coordinator GMES Bureau European Environmental Agency, Belgium; http://dk.linkedin.com/in/sorenroug

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

Helge Sagen, Head of Norwegian Marine Datacentre, Institute of Marine Research, Norway; http://no.linkedin.com/pub/helge-sagen/6/b58/403

Eefke Smit, International Association of STM Publishers, Director, Standards and Technology, The Netherlands; http://www.stm-assoc.org/whos-who-at-stm/

Martie J. van Deventer, Portfolio Manager for Information Services, Council on Scientific and Industrial Research, Soputh Africa; http://za.linkedin.com/in/martievandeventer

Koji Zettsu, Director, Information Services Platform Laboratory National Institute of Information and Communications Technology, Japan; http://www.nict.go.jp/en/univ-com/isp/members/zettsu/index.html

Consultants

Daniel Cohen, Program Officer, U.S. National Committee for CODATA and Board on Research Data and Information, National Academy of Sciences, USA (on detail from the U.S. Library of Congress); www.linkedin.com/pub/daniel-cohen-jd/1/3b9/24a/

Yvonne Socha, University of Tennessee Health Science Center; Minerva Consulting, USA; http://www.linkedin.com/in/yvonnesocha

Melissa L. Turcios, Research Assistant, U.S. National Committee for CODATA and Board on Research Data and Information, National Academy of Sciences, USA; http://www.linkedin.com/in/melissaturcios

Paul F. Uhlir, Director, U.S. National Committee for CODATA and Board on Research Data and Information, National Academy of Sciences, USA; www.linkedin.com/pub/paul-uhlir/7/88b/908

Lili Zhang, Computer Network Information Center, Chinese Academy of Sciences, China

CODATA EC Liaison

Bonnie Carroll (U.S. CODATA and CENDI), President, Information International Associates, USA; http://www.iiaweb.com/about/leadership/bonnieccarroll

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

APPENDIX B

Task Group bibliography on data citation publications

As part of the 2013 activities of the CODATA Digital Data Citation Task Group, we conducted an inventory of existing literature on data citation best practices and attribution activities. This document is the result of the collection of bibliographic sources, subsequent research, and corresponding analysis.

Prior versions of this bibliography have been published on the Task Group's website. [http://www.codata.org/taskgroups/TGdatacitation/Bilbliography_Links.html]. This version integrates all additional references gathered after the publication of the Summary Report of our August 2011 Workshop (sponsored jointly with the Board on Research Data and Information of the National Academy of Sciences), National Research Council. *For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: The National Academies Press, 2012.

The corpus of this bibliography was created by Task Group members and consultants who performed online database searches and captured information sources that are directly or peripherally focused on data citation practices and attribution. These contributions were then shared via email or via Zotero, a reference management software tool. Many new entries were contributed over time by authors and reviewers of the Task Group's reports, interviewees from various stakeholder communities, and other interested persons.

As of this version, we have found 441 resources and organized them into 16 different formats that cover the many facets of citation, such as policies, infrastructure, research practices, and best practices development. We concentrated our efforts on sources that were published during the past five years, with the occasional older, seminal item included because of additional context and background. Each source contains links, notes, or abstracts where applicable or possible. The document is accessible at: http://sites.nationalacademies.org/xpedio/groups/pgasite/documents/webpage/pga_084266.pdf

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

APPENDIX C

Organizations interviewed by Task Group members concerning data citation practices

In order to assess the progress of various stakeholder communities toward adoption of data citation and attribution practices, the Task Group identified stakeholder communities likely to have the greatest potential impact upon the development of citation and attribution practices (managers at data repositories and academic libraries, scholarly journal publishers, research institutions, and research funding organizations), recognizing that different stakeholder communalities might have different interests or concerns regarding data citation and attribution practices. While individual researchers were also identified as an important stakeholder community, in the interest of efficiency, the Task Group chose to focus its primary attention upon the institutional stakeholders with whom individual researchers would necessarily interact. Members of the Task Group then conducted telephone interviews with representatives of those stakeholder communities, in which the selected representatives were asked questions tailored specifically to each community. The interviews made no effort to achieve statistical validity but rather were designed to support our effort to assess the progress of those communities in their efforts to recognize and address issues regarding data citation as well as their perceptions of its importance.

The co-chairs of the Interview Subgroup were Sarah Callaghan, NCAS British Atmospheric Data Centre, STFC Rutherford Appleton Laboratory, UK and Todd Carpenter, National Information Standards Organization, USA.

The Interview Subgroup members who conducted the interviews were: Sarah Callaghan, NCAS British Atmospheric Data Centre, STFC Rutherford Appleton Laboratory, UK; Todd Carpenter, National Information Standards Organization, USA; Daniel Cohen, The National Academy of Sciences Board on Research Data and Information (on detail from Library of Congress), USA; Jianhui Li, Chinese Academy of Sciences, China; Helge Sagen, Norwegian Marine Datacentre, Institute of Marine Research, Norway; Eefke Smit, International Association of STM Publishers, The Nethrlands; Paul Uhlir, The National Academy of Sciences Board on Research Data and Information, USA; and Martie van Deventer, Council on Scientific and Industrial Research, South Africa.

Table 1. Summary of organizations interviewed by Task Group members concerning data citation practices.

Response #	Respondent	Institution	Country	Interviewer	Type
56	Jon Sears, Carter Glass	American Geophysical Union (AGU)	US	Todd Carpenter	Publisher
2	Fred Dylla, Evan Owens	American Institute of Physics (AIP)	US	Todd Carpenter	Publisher
57	Chris Biemesderfer	American Astronomical Society (AAS)	US	Todd Carpenter	Publisher
38	Jack Ochs, David Martinson	American Chemical Society (ACS)	US	Todd Carpenter	Publisher
3	Dr. Stuart Jeffrey	Archaeology Data Service (ADS)	UK	Sarah Callaghan	Repository
26	Susan Veldsman	Academy of Science for South Africa (ASSAf)	South Africa	Martie van Deventer	Repository

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

55	Ruth Lagring	Belgian Marine Data Center, Royal Belgian Institute of Natural Sciences	Belgium	Helge Sagen	Repository
39		British Oceanographic Data Centre (BODC)	UK	Sarah Callaghan	Repository
4	Project manager and scientific researcher	British Atmospheric Data Centre (BADC)	UK	Sarah Callaghan	Repository
35	Süenje Dallmeier-Tiessen	Conseil Européen pour la Recherche Nucléair / European Council for Nuclear Research (CERN)	CZ	Eefke Smit	Researcher/ Repository
63	Bob Downs	Center for International Earth Science Information Network (CIESIN)	US	Daniel Cohen	Repository
13	Dag Rosland and Olle Morten Grini	Climate and Pollution Agency	Norway	Helge Sagen	Funder
37	Clifford Lynch	Coalition for Networked Information (CNI)	US	Todd Carpenter	MISC - Libraries
22		Computer Network Information Center, Chinese Academy of Science	China	LI Jianhui	Repository
27	Adèle van der Merwe	Council for Scientific and Industrial Research (CSIR)	South Africa	Martie van Deventer	Repository
32	Lynn Woolfrey	DATA FIRST	South Africa	Martie van Deventer	Repository
60	Mark Martin	Department of Energy - Office of Scientific and Technical Information (DOE - OSTI)	US	Daniel Cohen	Repository
42	Carlson & Pfeiffenberger, Chief editors	Earth System Science Data journal	UK	Sarah Callaghan	Publisher
41		Elsevier	NL	Sarah Callaghan	Publisher
52	Thomas Hammond	Global Environment Fund	US	Paul F. Uhlir	Funder
28	Lucia Lötter	Human Sciences Research Council, South Africa (HSRC)	South Africa	Martie van Deventer	Repository

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

6	Head of ICES Data Centre, ICES Data manager	International Council for Exploration of the Sea (ICES)	UK	Helge Sagen	Repository
20	Geir Odd Johansen	Institute of Marine Research		Helge Sagen	Researcher
54	Graham McCann	Institute of Physics	UK	Eefke Smit	Publisher
36	Susan Reilly	Ligue Des Bibliothèques Européennes De Recherce / Association of European Research Libraries (LIBER)	EU	Eefke Smit	MISC - Libraries
50	Lisa Raymond, Associate Library Director	Library, Woods Hole Oceanographic Institution	US	Helge Sagen	Repository
61	Jeanne Behnke	National Aeronautics and Space Administration (NASA)	US	Daniel Cohen	Repository
15	George Strawn, Director	National Coordination Office for Networking and Information Technology Research and Development (NITRD)	US	Paul F. Uhlir	Funder
14	Regina Avila, Research Library	National Institute for Standards and Technology (NIST)	US	Paul F. Uhlir	Funder
53	Jerry Sheehan	National Institute of Health (NIH)	US	Paul F. Uhlir	Funder
17	Sylvia Spengler	National Science Foundation (NSF)	US	Paul F. Uhlir	Funder
45	Ruth Wilson	Nature	UK	Sarah Callaghan	Publisher
16	Jeffrey de la Beaujardiere, NOAA Data Mgmt Architect	National Oceanic and Atmospheric Administration (NOAA)	US	Paul F. Uhlir	Funder
58	Thomas Peterson and Nancy Ritchie	National Oceanic and Atmospheric Administration (NOAA)	US	Daniel Cohen	Repository
62	Dan Kowal	National Oceanic and Atmospheric Administration - National Geophysical Data Center (NOAA)	US	Daniel Cohen	Repository

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

8	Ken Casey	National Oceanic and Atmospheric Administration - National Oceanographic Data Center (NOAA-NODC)	US	Paul F. Uhlir	Repository
25	Daisy Selematsela	National Research Foundation (NRF)	South Africa	Martie van Deventer	Funder
64	Mark Parsons	National Snow and Ice Data Center (NSIDC)	US	Daniel Cohen	Repository
7	George Slesser	Oceanographic Group	UK	Helge Sagen	Repository
46		Open AIRE	EU	Sarah Callaghan	Repository
59	Phil Bourne	Protein Data Bank	US	Daniel Cohen	Repository
12	Ingunn Stangeby	Research Council of Norway	Norway	Helge Sagen	Funder
49	Paul Hardaker	Royal Meteorological Society	UK	Sarah Callaghan	Publisher
34	Richard Kidd	Royal Society of Chemistry	UK	Eefke Smit	Publisher
31	Henda van der Berg	South African Data Archive (SADA)	South Africa	Martie van Deventer	Repository
33	Wim Hugo	South African Environmental Observation Network (SAEON) & World Data Centre (WDC)	South Africa	Martie van Deventer	Repository
1	John Maunsell	Society for Neuroscience	US	Todd Carpenter	Publisher
5	Bruce Becker	South African National Grid	South Africa	Martie van Deventer	Repository
23	Bettina Goerner	Springer	Germany	Eefke Smit	Publisher
24	Eefke Smit	International Association of Scientific, Technical & Medical Publishers (STM)	NL	Eefke Smit	Publisher
51		Ubiquity Press	UK	Sarah Callaghan	Publisher
30		University of KwaZulu-Natal (UKZN)	South Africa	Martie van Deventer	Repository

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic

29		University of the Witwaters Rand (WITS)	South Africa	Martie van Deventer	Repository
9	Stein Tronstad			Helge Sagen	Repository
10	Lili Zhang		China	LI Jianhui	Repository
62	Dan Kowal	National Oceanic and Atmospheric Administration (NOAA)	US	Daniel Cohen	Repository
19	Heidi van Deventer	Council for Scientific and Industrial Research (CSIR)	South Africa	Martie van Deventer	Researcher
21	Hong Zhang	Council for Scientific and Industrial Research (CSIR)	China	Paul F. Uhlir	Researcher
40	Declan Vogt	Council for Scientific and Industrial Research (CSIR)	South Africa	Martie van Deventer	Researcher
43	Juanette John	Council for Scientific and Industrial Research (CSIR)	South Africa	Martie van Deventer	Researcher
44	J. Maritz	Council for Scientific and Industrial Research (CSIR)	South Africa	Martie van Deventer	Researcher
47	R. de Wind	Council for Scientific and Industrial Research (CSIR)	South Africa	Martie van Deventer	Researcher
48	R. Focke	Council for Scientific and Industrial Research (CSIR)	South Africa	Martie van Deventer	Researcher
Totals					
62				**9**	**Funder**
				14	**Publisher**
				28	**Repository**
				8	**Researcher**
				1	**Researcher / Repository**
				2	**MISC - Libraries**
				62	

(Article history: Available online 8 September 2013)

A correction article relating to this paper has been published and can be found at http://datascience.codata.org/artic