The Evolution of Data Citation: From Principles to Implementation

Micah Altman and Mercè Crosas [1]

Keywords: Data Citation; Bibliographic Practices

Forthcoming: IASSIST QUARTERLY

Abstract

Data citation is rapidly emerging as a key practice in support of data access, sharing, reuse, and of sound and reproducible scholarship. In this article we review the evolution of data citation standards and practices – to which Sue Dodd was an early contributor – and the core principles of data citation that have emerged through a collaborative synthesis. We then discuss an example of the current state of the practice, and identify the remaining implementation challenges.

1. Background

Data is, as they say, the new black. Scientific data are increasingly being made available online, and access to large collections of data is increasingly sought for education, science, policy, and commerce. Lowering barriers to discovery and use of these data and increasing our ability to link data with publications have the potential to enable new forms of scholarly publishing, promote interdisciplinary research, strengthen the linkage between policy and science, and lower the costs of replicating and extending previous research.

Many problems arise when research findings become disconnected from the underlying data that forms the evidence for these findings. The most well-publicized of these problems is scientific fraud. Access to data and the documentation of clear connections between the research results and the data facilitate detection of structural fraud both before and after publication. Other problems arising from this disconnect include irreproducibility, lack of reuse and wasted effort collecting new data, a proliferation of unmanaged versions and subsets of the 'same' data, and weak incentives for data sharing.

This is why the submission requirements for *Science*, one of the most cited, read, and respected journals in the sciences, requires that "all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*" and that "*citations to unpublished data* and personal communications cannot be used to support claims in a published paper" (emphasis added). (Science 2014)

Too often, this proscription, and others like it, have been honored only in the breach. The history

of data sharing makes this clear – despite clear recognition of the benefits of data sharing (Fienberg, *et al*. 1985) many research findings are based on data that is not made available -- making this research surprisingly difficult to replicate and even more difficult to extend. Furthermore, most research articles fail to provide clear citations to data, or the code necessary to reproduce, reuse, or extend results (CODATA 2013).

Within the social sciences, the vast majority of datasets produced by sponsored research is never deposited or shared (Pienta 2006), and, as a result, reproducing published tables and figures, and directly extending prior results is often difficult or impossible (Dewald, *et al.,* 1986; Altman, *et al.,* 2003; Hamermesh 2007). Similar problems exist in other fields: A recent study by Vines *et al.* (2014) of a sample of zoology articles found that less than 30% of even the most recent publications made data available, and that research data availability declined rapidly with article age, while loss of data increased. Moreover, a study of articles published in high-impact journals during 2009 showed that only 41% minimally complied with the journal's own data-sharing policies, and of these only 9% deposited the full primary raw data corresponding to the paper online (Alsheikh-Ali, et al 2011).

The research community has begun to take wider notice of this. And in the past two years a number of efforts have been launched by publishers, funders, professional associations, and organized projects to improve reliability, reproducibility, and data availability across a variety of scientific fields. We are optimistic that these projects will succeed, and if they do a key part of their success is likely to be through better scholarly recognition of data authorship.

There is increasing recognition that researchers are more inclined to share their data when they get credit (Borgman, 2012, p. 1072). Conversely, recent studies also suggest that researchers receive more credit when they share their data (Piwowar & Vision 2013). Publications that shared data from earlier years yielded an increase in citations of up to 30%.

Data citation, which has existed for 40 years in principle, is finally emerging as a pivotal norm for promoting data accessibility and accountability. Robust data citation practices and infrastructure will play a critical role in the widespread adoption of data citation and in the promotion of data sharing and its benefits.

2. The Emergence of Data Citation Principles and Practices

Within traditional print publishing, scholarly citation was widely formalized over a century ago. The first edition of the *Chicago Manual of Style*, published in 1906 under the title *Manual of Style: Being a compilation of the typographical rules in force at the University of Chicago Press* exemplified (and helped catalyze) the extent of standardization in scholarly citation. (Pollack, 2006) Within this tradition, a "bibliographic citation" referred to a formal, structured reference to another scholarly work that appeared in the text of a work. Typically, citations were either marked off with parentheses or brackets, such as: "(Altman 1992)," although in some fields footnotes were used. A standard reference entry included author(s), a title, a date, and a publisher (publishing house for books, journal name for articles) (Van Leunen 1992, pg. 186-

208). In addition, citations could include “pinpointing” information that identified which part of the cited work was being referenced, typically in the form of a page range. Citations to a single work could be repeated throughout the text. The reference list, typically appearing at the end of the main text, provided more detailed bibliographic information for each work cited in the text. Many variations were used for references to archival sources, correspondence, government documents, and artworks. However, each of these reference formats provided as well as possible at least three elements: author/creator, dates of the work, and the publisher or distributor of the work.

When the first scientific digital data archives were established in the late 1960s, their design focused on issues of access, storage, formatting, costs, and information retrieval (Bisco 1965). Bibliographic standards for cataloging data were developed over the next decade. In 1970 the American Library Association (ALA) formed a subcommittee on Rules for Cataloging Machine-Readable Data Files (MRDF), and tasked it with, among other things, bringing bibliographic control to MRDF. It was a long time before academic citation practices started to catch up with archiving practices, as summarized in Figure 1.

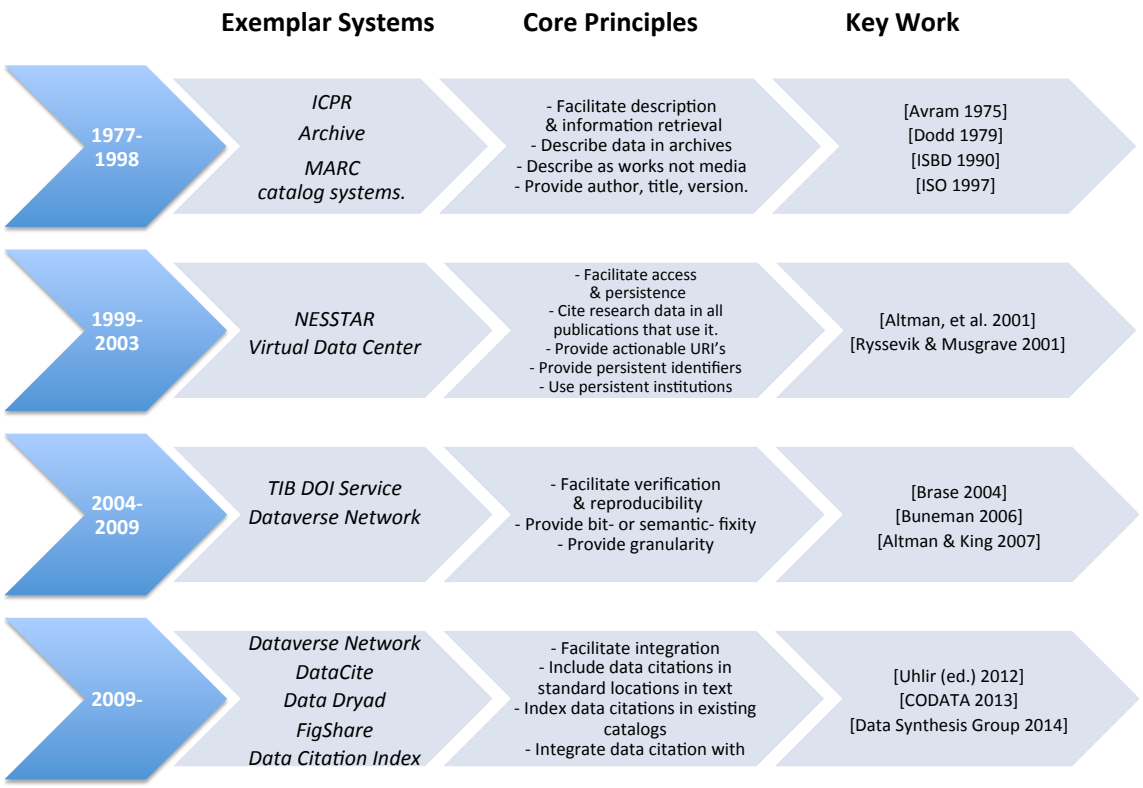

Figure 1: A chronology of data citation principles and related systems

(In the figure, “Exemplar Systems” indicate key software and or technical infrastructure supporting practices. “Core Principles” summarizes the principles identified for data citation –

as described in the text below. “Key Work” indicates the work related to principles of data citation and bibliographic practice – not responsibility for exemplar systems.)

The *American Standard for Bibliographic Reference* (ANSI Z29.29-1977, aka ASBR) provided a minimal “Data File” type to be used as part of the general material designator element in bibliographic metadata. Dodd (1979) quickly noted the shortcomings of ASBR in practice – notably the inconsistencies in describing the same dataset when presented in different physical formats, and the fact that the general approach conflated specific media with the intellectual works temporarily stored in those media. Dodd proposed using existing ASBR elements in a consistent and systematic way to bibliographically describe datasets *as intellectual works*. The key elements of Dodd’s approach emphasized the use of consistent *Title*, *Author*, and *Edition (*which included date). They were used along with a general media designator of “Machine Readable Data File(s)” (MRDF), which was format and media agnostic.

Dodd’s work was prescient – it contained many of the core elements of modern data citation and represented the state of the art for decades. The first major international standard that addressed data citation in more detail wasn’t published until 1990 (ISBD 1990, ISO 690-2 1997), and it incorporated Dodd’s treatment of databases as an intellectual work, requiring Title, Author, and Version (which corresponded to Dodd’s “Edition”). The ISO standard also took a step backward, however, in explicitly requiring reference to specific media types such as “CD-ROM” and in lumping together databases and software programs.

The recognition of data as a public good[2] was, however, insufficient by itself to support or incentivize data sharing. In general, public goods in the absence of effective norms, regulation, or subsidies will be under-supplied. The state of the art in data citation, as well as in data sharing, did not progress quickly until catalyzed through advances in information technology, open source software development practices, and legal infrastructure. The growing recognition among scholars that data is a fundamental product of research, a trend identified in the National Research Council’s foundational report on data sharing (Fienberg 1985), began to build slowly in momentum through the leadership of individual scholars such as Sieber (1991) and King (1995). Then, rapid advances in Internet and web infrastructure greatly decreased the technical barriers to data sharing. More recently the rapid growth of the Open Software movement generally, together with development of the legal “technology” of robust standardized open licenses, have sparked initiatives in academia to build open tools in support of scholarly access, discovery, collaboration, and research sharing.

Building on these trends, and supported by the NSF Digital Library Initiative (Griffin 1998), Altman, King & Verba developed one of the first open source (and open access) data publishing systems, the *Virtual Data Center* (Altman *et al*. 2001). This system successfully fielded the largest federated catalog of social science datasets in the world (Altman *et al*. 2009). The virtual data center was designed to support persistent access to research data through federated institutional curation. Data citation was deeply integrated into the *Virtual Data Center* – each dataset managed was assigned a persistent identifier, and a citation. Moreover, the *Virtual Data*

Center was based on the principle that all data supporting published research should be cited, and that these citations and identifiers should be machine-actionable through the web (e.g. through machine-actionable URI's). *Nesstar,* a system developed in parallel by Ryssevik & Musgrave (2001), and later used by many European archives, also incorporated the concepts of actionable web links, and persistent federated curation – although it did not initially support or emphasize citation.

Incorporating work by Altman, *et al.* (2003) and Altman and King (2007), the Virtual Data Center incorporated both support for "deep citations" (Buneman 2006) that identify precise subsets of a larger dataset; and for semantic fixity information that enables verification of a dataset using the citation itself. These capabilities were further extended in the Dataverse Network (King 2007), which succeeded the Virtual Data Center. The Dataverse Network has since been adopted by the Harvard University as its data publication infrastructure and is used by hundreds of researchers in dozens of institutions to curate and publish data. (Crosas 2011, 2013)

In parallel work, Brase (2004) lead an initiative to systematically archive datasets associated with research outputs, and to systematically associate these datasets with Digital Object Identifiers (DOI, 1997) – a robust form of persistent identifier used in the publication community. This was the first step toward integration of data citation and data publication into the larger publishing ecosystem.

To summarize, from 1977 through 2009 there were three phases of development in the area of data citation.

- The first phases of development focused on the role of citation to facilitate description and information retrieval. This phase introduced the principles that data in archives should be described as works rather than media, using author, title, and version.
- The second phase extended citations to support data access and persistence. Building upon the principle that research data used in publication should be cited, this phase introduced the principles that those citations should include persistent identifiers, and that the citations should be directly actionable on the web.
- The third phase of development focused on using citations for verification and reproducibility. Although verification and reproducibility had always been one of the motivations for data archiving – it had not been a focus of citation practice. This phase introduced the principles that citations should support verifiable linkage of data and published claims, and it started the trend towards wider integration with the publishing ecosystem.

The importance and urgency of scientific data management and access is now starting to be recognized broadly. Many publishers recognized this, in theory, in 2006, when the "Brussel's Declaration" put forth the principle that data associated with publications should be openly available. This same year, the U.S. National Science Foundation introduced a policy requiring every grant proposal to be accompanied by a data management plan. Also that same year, data management was the theme of the annual meeting of the Society of Scholarly Publishers, the premier conference in that field. This continues a trend of funders and publishers adopting data

publication and management policies. Universities have likewise become involved and have started to develop their own policies requiring data management, while journals, archives, and research libraries are increasingly grappling, largely independently, with the issues of data management.

Even the media has taken note. This is reflected by numerous articles drawing attention to particular high-profile cases of scientific fraud, such as the Stapel affair (e.g., Carey 2011), to increased rates of retractions (e.g. Ionaddis 2005, Steen 2010, Fang *et al* 2012), and to the practice of Open Science more generally (e.g, Lin 2011).

The culmination of this trend, thus far, is an increasingly widespread consensus by researchers and funders of research that data is a fundamental product of research and therefore a citable product. The fourth and current phase of data development work focuses on integration with the scholarly research and publishing ecosystem. This includes integration of data citation in standardized ways within publication, catalogs, tool chains, and larger systems of attribution. It is exemplified by systems such as Data Dryad (Vision 2010) and Figshare (Hahnel 2013) which integrate data deposition into publisher workflows, and DataCite and the Thomson Reuters Data Citation Index, which integrate data citations into index and discovery of other published work; and by community standardizations efforts, such as those coordinated by the National Academies (Uhlir 2012), CODATA (2013), and the Data Citation Synthesis Group (2014).

Across these various groups there has been a developing agreement over the years that an essential part of connecting research publications or claims to data is formal data citation that includes a persistent link to guarantee long-term data accessibility. Global persistent identifiers, such as DOIs and Handles, offer a mechanism to provide a permanent link that can be configured to always resolve to a web page from which the data can be accessed, independent of whether the location of that page changes over time. An increasing number of data repositories generate DOIs which can be directly used in a publication to reference the data. However, until now, there has not been a single set of principles or guidelines for data citations which represents and is in agreement with all these initiatives.[3]

What has emerged in the bibliographic and research community is a substantial core of agreement over the need for citation to support attribution and verification; the recognition that citations must support both human and machine clients; the existence of robust persistent identifiers and the understanding of the core role; and the publication of key reference documents such as the National Academies and CODATA reports.

3. Converging Data Citation Principles

Given the rise of these parallel, variously implemented initiatives on data citation, as well as the lack of unified guidance for publishers, journal editors, and funding agencies, there was a need for a synthesis set of general recommendations and good practices for data citation. In the summer of 2013, a synthesis group was formed to unify the various recommendations. It came to be known as the Data Citation Synthesis Group. It met weekly from July to November of 2013

to thoroughly deconstruct previous data citation principles defined by CODATA, the Amsterdam Manifesto, and DataCite, and to produce a synthesis set that included the input of more than 25 organizations. During that time, the group met as part of the RDA (Research Data Alliance) conference in Washington, DC in September, in two half days of public workshop. As a result, in November 2013, the proposed *Joint Declaration of Data Citation Principles* was released to the public for open comment, and finalized at the end of February 2014 (Data Citation Synthesis Group, 2014)

The scope of the synthesis principles is solely to provide data citation recommendations, and does not intend to include detailed specifications for implementation or to focus on technologies or tools or research data repositories. The principles should extend to all disciplines and all types of data. Some of the challenges for specific types of data will be discussed in the next sections. As will be seen below, the *Joint Declaration of Data Citation Principles* reflect the various efforts described in the last section and a broad convergence on core principles:

1. **Importance** Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. **Credit and Attribution.** Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. **Evidence.** In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
4. **Unique Identification.** A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. **Access.** Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. **Persistence.** Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.
7. **Specificity and Verifiability.** Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.
8. **Interoperability and flexibility.** Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities[8].

At the time this article was completed, less than a month after the principles had been finalized, they had been officially endorsed by thirty organizations, including many major publishers and data archives. The synthesis group has also committed to a dissemination plan that includes reaching out to a large number of stakeholders from multiple organizations and disciplines for

an endorsement of the principles.

We anticipate that the impact of the unified, widely broadcasted *Joint Declaration of Data Citation Principles* will be substantial and will: change current publication workflows, create new data citation technologies, define new metrics for scholarly impact and recognition, and, more importantly, provide persistent access to the data supporting scientific results to validate and extend previous scientific work. The *Principles* will facilitate interoperability across existing and new implementations, and will help guide enhancements and new versions of the current implementations. Several data repositories are already compliant, or close to compliant, with these principles (e.g., Dataverse, DataDryad). In section five, we describe, as an example, the Dataverse Network data citation implementation.

4. A Generic Example

A generic example for a data citation can be represented as:

> Author(s), Year, Dataset Title, Global Persistent Identifier, Data Repository or Archive, version or subset

The *authors* and the *data repository or archive* elements directly support principle two, providing credit and attributions to the creators of the data as well as to their publishers or distributors. As in citations of literature, in some cases the creators are not individual authors, but instead an entity or organization that produced the data. Also as in citations of literature, authorship can be challenging and ill-defined in a simplified citation format when there is a large number of individuals who have contributed to the scholarly product in a wide range of ways (e.g., from designing the instrument and software to cleaning and analyzing the data.) We already find these authorship challenges in publications in high-energy physics, such as articles related to the observation of the Higgs Boson having nearly 3,000 authors (e.g., CMS Collaboration, 2012). This wide array of authors is more common for data products than for articles. The *Principles* and this citation example do not address the authorship problem, but, as described below, the metadata associated with the dataset can allow annotation of various levels of contribution during the creation and processing of the data, and also allows reference to related datasets or other scholarly products.

The *year* in which the dataset is first published and the *title* are not directly related to a principle. However, these elements are common in traditional literature citations, and such consistent and informative formats contribute toward giving data citation the same importance as citations of other scholarly records, as stated in principle one.

The *Global Persistent Identifier* is an essential piece of the citation of a digital object and directly supports principle four. The persistent identifier or URL allows separation of the link given in the citation with the URL to which it resolves, thus guaranteeing that even if the hosting or location of the dataset's web page changes, the link in the citation will always go to the same dataset page. In a forthcoming article by Pepe, *et al* (2014), based on a study of 7,641 astronomy

publications from four main astronomy journals, we show that 44% of the links in publications from ten years ago are broken. These are regular links to web sites, and not global persistent identifiers. The persistent identifier or URL solves a technical problem, but it is not sufficient without a publisher that supports and guarantees the validity of its persistent identifiers. In the case of data, the publisher is usually the data repository or archive. The more commonly used global persistent identifiers are handles (Sun, *et al.* 2003) and DOIs (Paskin, 2002). The persistent identifier in the data citation example also supports principles five and six. In support of principle five, the handle or DOI should resolve to a dataset page, which contains sufficient information describing the data and facilitating their reuse. In the rare cases in which the data cannot be made accessible any longer or must be destroyed, the data citation should still be valid. That is, the persistent identifier should resolve to a page with information about the discontinuation of that dataset (principle six).

The last element in the generic citation example is the version, subset, or timestamp, which supports principle seven. This element is particularly relevant when citing data. Contrary to most literature publications, a dataset is often altered or expanded with time. The frequency with which a dataset might be changed can vary, from a static dataset that never changes once published, to a dataset that is updated once in a while with a new version, to datasets that are constantly changing, as is the case of dynamic data from meteorological sensors or streaming data Twitter feeds that grow constantly over time. Dynamic and streaming data offer a number of challenges for both citation and replication of published results contingent upon the reuse of a specific version of a dataset. Those challenges are described in section six.

The generic citation example might vary in style from community to community (principle eight), but across all cases it should be considered as important as other citations and should be part of either the standard reference section of a publication or a similar section for data citations, in accordance with principle one. The Data Citation Synthesis Group also recommends that when a published claim is made based on the data, enough information should be provided in the text to identify the data citation listed in the reference section, in the same fashion as other citations. When the published work makes a claim based on a subset of the data, specific information about the subset should be referenced by that claim.

Due to the possible complexity of such a citation, it is not always feasible to include in the reference section all the information needed to fulfill the core data citation principles. For this purpose, as stated in principle four, an important component of any data citation is machine-actionable metadata that is bound to the data citation and persists with it. For example, the DataCite metadata schema and ontology (DataCite 2013) describe a detailed set of fields that may be used to complete a data citation. Typically, additional fixity and provenance information is required to support the verification requirements – such that future users of the citations can ensure that the data they use is identical to that cited. Such information might include bit-level fixity information (such as a MD5, SHA-256 or other cryptographic hash), or preferably, where available, semantic fixity information (such as a UNF or perceptual fingerprint).

Additional information on contributors will be required to fulfill the attribution requirements

wherever the authors explicitly listed in the reference are ambiguous or incomplete. Unstructured metadata such as a contributors list may fulfill the bare legal requirements for attribution; however, structured name authority or identifiers such as ORCID's (Open Research Identifier) or ISNI's (International Standard Name Identifiers) are much preferred, because they facilitate scholarly attribution (credit). This information can be embedded in published documents in machine-accessible form, included in the metadata stored with the DOI or other persistent identifier by its resolver service, or stored in an associated community index, such as CrossRef or DataCite. Such metadata should also be presented through the landing page provided to humans when the persistent identifier for the data is resolved.

5. Implementing the State of the Practice

Data repositories, or data publishers, are often responsible for implementing and generating data citations for the datasets hosted within them. As noted above, there are a number of repositories that are already generating data citations upon deposit of a dataset, and those citations are often compliant with the principles above (e.g., Dryad, Dataverse, Figshare). The generic data citation example in section four is based on the citation format generated by the Dataverse Network software application. This application is a data repository platform that allows organizations to host dataverses, where each dataverse contains datasets, and where each dataset contains data files and metadata. A dataverse is, in essence, a virtual archive, which can be branded and administered individually, giving control to the data owner or distributor, while its data and metadata are stored by the repository in accordance to professional archival practices, metadata standards, and preservation formats (King 2007, Crosas, 2011). The software is open-source and developed at the Institute for Quantitative Social Science at Harvard University (King, 2014). The Harvard Dataverse is one of the Dataverse Network instances open to all researchers and to all data types. It supports a variety of types of dataverses, from journal dataverses, to dataverses for individual researchers, to dataverses for data associated with an institutional department (Crosas, 2013). In this section, we describe the implementation of data citation as it is built in Dataverse version 4.0.

When a new dataset is added to a Dataverse, the required metadata fields that must be entered by the depositor include the author(s) or producer organization and the dataset title. In addition, an extensive set of metadata fields are provided, some required and others optional. The citation metadata supported by Dataverse maps closely to the DataCite metadata, and can also be mapped to the format developed by the Data Documentation Initiative (DDI, http://www.ddialliance.org/Specification/) and Dublin Core Metadata Initiative Terms (DCTERMS, http://dublincore.org/documents/dces/). The dataset, when created, is in a draft form that is unpublished, and data files and additional metadata can be added at a later time. Upon dataset creation, however, even if the dataset is not yet published, a draft data citation is instantly generated following these steps:

1. Authors and title are obtained from the metadata fields entered by the data depositor. If instead of individual authors, a producer (organization or institution) is entered, the producer

is used in place of the authors.

2. In the draft citation, the year is automatically populated by the year of deposit. At the time when the dataset is released, the final citation is updated with the year of the released or published date, which is often, but not always, the same year the dataset was deposited.
3. The Dataverse Network software supports both handles and DOIs as persistent identifiers. If a Dataverse Network is configured to use handles, each handle is registered to the Handle System. The Harvard Dataverse is configured to use DOIs, which are registered to DataCite through the EZID API (http://ezid.cdlib.org/home/documentation). Upon deposit, the dataset is registered with status "reserved", an option provided by the EZID API. When the dataset is released, the status becomes "public". This means that the DOI at that point resolves to a public dataset page, which includes description information about the dataset, as well as information on how to access the data. Even when data cannot be completely open, and one or more data files in the dataset are restricted due to data user agreements or confidential information, the DOI resolves to a dataset page where access can be requested.
4. The publisher or data repository element in the citation is automatically populated as the repository name, in this case, the Harvard Dataverse. If additional distributors or archives are responsible for those data, they can be listed in the dataset page, as part of the additional metadata.
5. The Dataverse Network software supports versioning of datasets because, unlike traditional literature publications, data are often updated even after being published. The data citation generated by Dataverse includes the version of the dataset. When the dataset is released, the version in the citation is set to 1. If the dataset metadata or files are updated in the future, a new version is created, and a new citation, with the same DOI, but a new version number, is created. This allows reference to a specific previous version, and access to that version from the dataset page within a dataverse. It is important to note that a DOI or other persistent identifier is not equal to a data citation. The data citation is the composition of all the elements that form it, and the DOI is one of these elements. Therefore, one can cite two versions of a dataset with the same DOI, as long as the citation provides unambiguous information about the version. This is similar to citing a subset of the entire dataset, or in other type of citations, citing a set of pages in a book.

The data citation generated by the Dataverse Network software also supports Universal Numerical Fingerprints (UNF) for tabular datasets (Altman and King, 2007). The UNF guarantees fixity; it's a unique fingerprint on the semantics of a dataset. That is, even if a dataset changes format, if the data values remain the same, the UNF remains the same. When a UNF cannot be calculated, the Dataverse calculates bit-level fixity information (the MD5) of the data file(s) contained in the dataset.

The Dataverse Network implementation is fully compliant with the data citation principles discussed throughout this article. However, it does not support, in its current form, dynamic or streaming data. This is discussed in more detail in the next section.

6. Remaining Challenges

At the broadest conceptual level, the substantial remaining challenges for implementing robust data citation systems fall into three categories: [4]

- *Challenges 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.
- *Challenges of identity.* These theories involve defining 'data' themselves, 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.
- *Challenges 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.

Provenance is a particularly important concern because many data citations are used to document a direct evidentiary relationship between a published assertion and the underlying evidence that supports it. However, supporting this evidentiary relationship does not require recreating or establishing the entire provenance chain – and much of provenance can be considered as orthogonal to citation, as Groth (2012) argues. Notwithstanding, as Smith (2012) points out, enabling readers to establish authenticity of the cited object is an important use for citation and requires that citation be connected to provenance information. The maintenance of this connection and of the associated provenance information is a major challenge for developing reliable citable scientific workflows.

Identity is close to the heart of creating a citation. To cite something requires it to be identified – the citation should enable us to find the *same thing* that was used in the citing article. Identity is relatively straightforward for immutable data in the original formats and used as a whole. However, when data that changes over time is manifested in different formats, or is used only in part, a number of practical questions emerge:

- *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?

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 and Lee, 2002); creation of semantic fingerprints for data objects, such as UNF's (Altman *et al.*, 2003, 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, *et al*. 2004). Algorithms are being developed for generating persistent granular citations of specific forms of dynamic data objects, particularly of databases.[5] Moreover, open annotation frameworks and ontologies are being developed to allow interoperable annotation of digital objects that define spatial (logical) and temporal granularity which might be used generally to complement bibliographic data citations and support deep citation (Van de Sompel, 2012).

Natural corollaries to these questions involve considerations of scalability. For example, how does one track and recreate versions of 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?

A third challenge is that of attribution. Citation should support unambiguous attribution of credit for all contributors. As the scale of the data increases, and more people contribute to its creation and maintenance, practical challenges with attribution arise. These include supporting attribution for contributors that may number in the hundreds of thousands in crowd-based citizen science (e.g. Wiggins and Crowston 2011), distinguishing among different contributor roles (IWCSA 2012), and capturing the nature of the relationship between the cited and citing objects (e.g. Cronin 1984).[6]

Summary

Scientific data are increasingly being made available online. Lowering barriers to discovery and use of these data, and increasing our ability to link data with publications have the potential to enable new forms of scholarly publishing, promote interdisciplinary research, strengthen the linkage between policy and science, and lower the costs of replicating and extending previous research. Robust data citation practices and infrastructure will play a critical role in achieving these outcomes.

Bibliographic standards for cataloging data developed gradually from the early days of data archives but it was a long time before academic citation practices started to catch up with archiving practices. Over four decades ago, however, several core principles for data citation and bibliographic description were recognized – in part based on the pioneering work of Sue Dodd. For the next 25 years, data citations had little attention from or impact on either the scientific or library community – despite the fundamental soundness of many of the early principles and the implementation of citation practices by selected major data repositories. More recently data citation principles and practices have made a resurgence – fueled both by advances in web and network technologies and by a growing public and scientific recognition of the importance of scientific reproducibility, data sharing, and reuse. Recently, a wide

convergence on principles has emerged, and the deployment of production infrastructure to support data citation across the research lifecycle is rapidly advancing.

Key enablers of a successful synthesis process have included a substantial core of agreement concerning the need for citation to support attribution and verification; the recognition of the need for citation to support both human and machine clients; the existence of robust persistent identifiers and the understanding of their core role; and the publication of key reference documents such as the National Academies and CODATA reports.

A number of central challenges remain, particularly related to the frontiers of data – big data, complexly structured data, dynamic data, and data in changing formats. These are being addressed gradually through groups such as RDA and through state-of-the-practice development of systems such as the Dataverse Network.

Acknowledgments

We would like to thank the members of the Data Citation Synthesis Task Group and of the Co-Data Data Citation Working Group for commentary on this paper and on the ideas leading into it: Amy Brand, Amye Kenall, Andras Rauber, Anita deWaard, Bonnie Caroll, Christinge Borgman, Dan Cohen, David Shotton, Eefke Smit, Elizabeth Arnaud, Elizabeth Iorns, Fiona Murphy, Franciel Linares, Giri Palanisami, Hannelore Vanhaverbeke Heige Sagen, Hylke Koers, Ivan Herman, Jan Brase, Jianhui Li, Jo McEntyre, Joan Starr, Joe Hourcle, John Helly Maren Morgenroth, Kathleen Cass, Kerstin Lehnert, Koji Zettsu, Mark Hahnel, Mark Parsons, Martie van Deventer, Maryann Martone, Michael Diepenbroek, Michael Wit, Mustapha Mokrane, Natalia Moanola, Paul Groth, Paul Uhlir, Phil Archer, Puneet Kishor, Ruth Duerr, Sarah Callaghan, Simon Hodson, Stefan Proell, Stephanie Hagstom, Tim Clark, Tim Smith, Todd Carpenter, Vishwas Chavan, Yannis Ionnadis, Yasuhrio Muryama

References

Alsheikh-Ali, A. A., W. Qureshi, M.H. Al-Mallah, & J.P. Ioannidis. (2011). “Public availability of published research data in high-impact journals.” *PloS one*, 6(9), e24357. < http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0024357#pone-0024357-g001>

Altman, M. (2008) "A fingerprint method for scientific data verification."*Advances in Computer and Information Sciences and Engineering*. Springer Netherlands: 311-316.

Altman, M., L. Andreev, M. Diggory, G. King, A. Sone, S. Verba, and D.I L. Kiskis. (2001) "A Digital Library for the Dissemination and Replication of Quantitative Social Science Research The Virtual Data Center." *Social Science Computer Review* 19(4): 458-470.

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

Altman, M, and G. King. (2007) . "A proposed standard for the scholarly citation of quantitative data." *D-lib Magazine* 13.3/4. <http://www.dlib.org/dlib/march07/altman/03altman.html>

Altman, M., M.O. Adams, J. Crabtree, D. Donakowski, M. Maynard, A. Pienta and C.H. Young. (2009). "Digital preservation through archival collaboration: The data preservation alliance for the social sciences." *American Archivist* 72, no. 1: 170-184. <http://archivists.metapress.com/content/eu7252lhnrp7h188/fulltext.pdf>

Avram, H. D. (1975). *MARC, its history and implications*. Washington: Library of Congress.

Bisco, R. L. (1965). "Social Science Data Archives: Technical Considerations." *Social Science Information* 4:3, 129-150.

Borgman, C. (2012) “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.

Brase, J. (2004) "Using digital library techniques–registration of scientific primary data." In *Research and advanced technology for digital libraries*, pp. 488-494. Springer Berlin Heidelberg.

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.

Cano, E. Batle, T. Kalker, J. Haistma, (2002) “A Review of Algorithms for Audio Fingerprinting”, *IEEE Workshop on Multimedia Signal Processing,* IEEE Press:169- 173.

Carey, B. (2011), “Fraud case seen as red flag for psychology research.” *New York Times*, A3. November 3, 2011.

CMS Collaboration, (2012) “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC.” *Physics Letters B*, volume 716, Issue 1, pages 30-61.

CODATA/ITSCI Task Force on Data Citation, (2013). “Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation.” *Data Science Journal* 12: 1-75., <http://dx.doi.org/10.2481/dsj.OSOM13-043>

Cronin, Blaise. (1984)*The citation process. The role and significance of citations in scientific communication*. London: Taylor Graham.

Crosas, M. (2011). “The Dataverse Network: An Open-Source Application for Sharing,

Discovering and Preserving Data." *D-Lib Magazine* 17 (1–2). < http://www.dlib.org/dlib/january11/crosas/01crosas.html>

Crosas, M. (2013). "A Data Sharing Story." *Journal of eScience Librarianship* 1 (3):173–79. <http://escholarship.umassmed.edu/jeslib/vol1/iss3/7/>

Crosas, M., T. Carpenter, C. Borgman, D.M. Shotton. *(*2013). "The Amsterdam Manifesto on Data Citation Principles." *Force11*.

Data Citation Synthesis Group, (2014). *Joint Declaration of Data Citation Principles,* <http://www.force11.org/datacitation>

DataCite, (2013). "DataCite Metadata for the Publication and Citation of Research Data" doi:10.5438/0008

Dewald, W.G., J.G. Thursby, and R.G. Anderson. (1986). "Replication in empirical economics: The journal of money, credit and banking project." *American Economic Review*, 76(4):587-603.

Dodd, S. A. (1979) "Bibliographic Reference for Numeric Social Science Data Files: Suggested Guidelines." *American Society for Information Science Journal* 30:2, 77-82.

DOI, (1997). *DOI Handbook* <http://www.doi.org/hb.html>

Fang, F. C., R.G. Steen and A. Casadevall. (2012). "Misconduct accounts for the majority of retracted scientific publications." *Proceedings of the National Academy of Sciences*, 109(42), 17028-17033.

Fienberg, S. E., M.E. Martin and M.L. Straf. (1985). *Sharing Research Data.* Washington, D.C.: National Academies Press.

Griffin, S. (1998). "NSF/DARPA/NASA digital libraries initiative." *D-Lib Mag*, 4(7). <http://www.dlib.org/dlib/july98/07griffin.html>

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.

Hahnel, M. (2013) "Referencing: The reuse factor." *Nature* 502.7471: 298.

Hamermesh, D.S. (2007). "Viewpoint: Replication in Economics," *Canadian Journal of Economics.*

Hedstrom, M. and C. Lee (2002). "Significant properties of digital objects: definitions,

applications, implications." *Proceedings of the DLM-Forum*: Parallel session : 218-113.

ISBD (1990). International Standard Bibliographic Description for Computer Files. Recommended by the Working Group on the International Standard Bibliographic Description for Computer Files set up by the IFLA Committee on Cataloguing. ISBN 0-903043-56-4

Ioannidis, J. P.A. (2005). "Why most published research findings are false." *PLoS medicine* 2.8: e124.

ISO, (1997). Information and documentation -- Bibliographic references -- Part 2: Electronic documents or parts thereof. 690-2:1997. International Standards Organization.

IWCSA Report. (2012). *Report on the International Workshop on Contributorship and Scholarly Attribution, May 16, 2012.* Harvard University and the Wellcome Trust. Available at: <http://projects.iq.harvard.edu/attribution_workshop>.

King, G. (1995). "Replication, replication." *PS: Political Science and Politics* 28.3: 444-452.

King, G. (2007). "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." *Sociological Methods and Research* 36 (2): 173–99.

King, G. (2014). "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science." *PS: Political Science and Politics* 47, no. 1: 165-172.

Lin, T., (2011) "Cracking open the scientific process." *New York Times*, D1. January 17, 2011.

Paskin, N. (2002). "Digital object identifiers." *Information Services and Use* 22.2: 97-112.

Pepe, A., A. Goodman, G. Muench, M. Crosas, C. Erdmann (2014). "Sharing, Archiving and Citing Data in Astronomy" *PLOS ONE* (Forthcoming).

Pienta, A. (2006). "LEADS Database Identifies At-Risk Legacy Studies." *ICPSR Bulletin* 27(1).

Piwowar, H. and T. Vision (2013). "Data Reuse and the Open Data Citation Advantage" *PeerJ*

Pollak, O. B. (2006). "The Decline and Fall of Bottom Notes, op. cit., loc. cit., and a Century of the Chicago Manual of Style." *Journal of scholarly publishing* 38.1: 14-30.

Proll, S. and A. Rauber (2013). "Scalable data citation in dynamic, large databases: Model and reference implementation." *Big Data, 2013 IEEE International Conference on*. IEEE.

Ryssevik, J. & S. Musgrave (2001). "The Social Science Dream Machine: Resource Discovery, Analysis, and Delivery on the Web" *Social Science Computer Review* 19(2) 163-174.

Science (2014). General information for authors. Retrieved from http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml

Sieber, J. E. (1991). *Sharing social science data: Advantages and challenges*. Sage Publications, Inc.

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

Steen, R.G. (2010). "Retractions in the scientific literature: is the incidence of research fraud increasing?" *Journal of Medical Ethics* 37: 1-5.

Sun, S., L. Lannom, and B. Boesch (2003). "Handle system overview." *RFC 3650*, November, 2003.

Uhlir, P. F., (Ed.) (2012). *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop*. Washington, D.C.: National Academies Press.

Van de Sompel, H. (2012), "Data citation - technical issues - identification" in Uhlir, P. F., (Ed.) (2012). *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop*. Washington, D.C.: National Academies Press.

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

Vines, T. H.; A.Y.K. Albert, R.L. Andrew, F. D barre, D.G. Bock, M.T. Franklin, K.J. Gilbert, J-S Moore, S. Renaut, D.J. Rennison (2014). "The Availability of Research Data Declines Rapidly with Article Age" *Current Biology* 24 (1): 94 - 97.

Vision, T. J. (2010). "Open data and the social contract of scientific publishing."*BioScience* 60, (5): 330-331.

Wiggins, A., and K. Crowston (2011). "From conservation to crowdsourcing: A typology of citizen science." *System Sciences (HICSS)*, 2011 44th Hawaii International Conference on. IEEE.

Notes

[1] Authors listed alphabetically; the authors have made equal contributions to this work.
Micah Altman is Director of Research, MIT Libraries at the Massachusetts Institute of Technology.

He can be contacted at <escience@mit.edu>. Mercè Crosas is Director of Data Science, Institute for Quantitative Social Science at Harvard University. She can be contacted at <mcrosas@iq.harvard.edu>.

[2] Or more precisely, in some cases it is a "club good" – nonconsumptive and only partially excludable.

[3] Efforts in this area have been made by CODATA, as part of an extensive report on data citation (CODATA 2013), DataCite principles, DCC as part of the core guidelines on data curation, Harvard's Institute for Quantitative Social Science through a data citation workshop hosted in 2012, and Force11 in the form of the *Amsterdam Manifesto for Data Citations Principles*, born at the Beyond the PDF 2 conference in 2013 (Crosas *et al,* 2013), among others, and by multiple research data repositories that offer to generate data citation upon deposit of a dataset (such as Dataverse, DataDryad, Figshare, and the Inter-University Consortium for Political and Social Research (ICPSR)).

[4] This section in part summarizes and updates section 7.2 in the CODATA report (2013), which was originally written by one of the authors of this article.

[5] See Buneman (2006) for fundamental work in this area; also Proll and Rauber (2013) for a more recent approach.

[6] Cronin [1984] reviews over 10 different proposed taxonomies of citation types and roles, some of which identify dozens of individual relationships.