Title
Data citation as a bibliometric oxymoron

ISBN
978-3-11-029803-1

Author
Borgman, Christine L.

Publication Date
2016-03-21

Data Citation as a Bibliometric Oxymoron

Christine L. Borgman
Professor & Presidential Chair in Information Studies
University of California
Los Angeles, California
Christine.Borgman@ucla.edu
http://christineborgman.info

Please cite as:
Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication* (pp. 93–115). Berlin, Boston: Walter de Gruyter GmbH & Co KG.

Introduction

"Data citation" is a broad construct that incorporates credit, attribution, and discovery of data. It has taken on a life of its own, quite apart from the theory and method of bibliometrics, informetics, scientometrics, and other means to assess the flow of scholarly information via citations between published documents. International task groups have written hundreds of pages of reports. Manifestos abound. Principles and standards for data citation are being set and implemented in local practice (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013; Crosas, Carpenter, Shotton, & Borgman, 2013; Datacitation Synthesis Group, 2014; Uhlir, 2012).

The argument for data citation is made most succinctly in the first of ten principles promulgated by a joint task group of CODATA, the Committee on Data of the International Council of Scientific Unions and ICSTI, the International Council for Scientific and Technical Information (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013; "International Council for Scientific and Technical Information," 2015; Lide & Wood, 2012):

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

The status principle puts data on equal footing with other objects that are cited in scholarly communication – but without defining what those are or establishing the basis for equal treatment. This equivalence raises a host of theoretical, methodological, and practical problems for bibliometrics. Historically, bibliometrics involves "written communication" (Pritchard, 1969, p. 348), and specifically journals, periodicals, and books (Pritchard, 1969; Raisig, 1962). Bibliographic citation styles differ widely in the choice of data elements and citable units, as discussed further below. While the lack of agreement on bibliographic units for the purpose of citation remains problematic, at least these units usually can be aggregated into discrete documents such as journal articles or books. Of the many differences between data and written communication, the difficulty of defining citable units for data is the most problematic for bibliometrics.

Treating data citations as equivalent to bibliographic citations implies that data are publications, which in turn gives rise to the popular "data publication" metaphor. While "publication," strictly speaking, means only "to make public," publication in the sense of scholarly communication has a much higher bar. Scholarly publication normally requires peer review and dissemination in a venue with recognized status for credit and attribution (Borgman, 2007). Journals and books usually meet this standard of publication, whereas talks, blog posts, and objects posted on web pages generally do not. Reciprocal citation is a feature of bibliometrics and of related methods such as webometrics, scientometrics, and informetrics. Data are far more complex objects – if they are objects at all – than the entities to which bibliometrics applies. Units of data might receive citations, for example, but it is not clear that they can make references to other objects. The status principle for data citation and the data publication metaphor combine to muddy the waters of scholarly communication at time when far more clarity about the characteristics of data is needed (Borgman, 2007, 2012, 2015; Parsons & Fox, 2013).

The leap from citing publications to citing data is a vast one, but if data are to be discovered, exchanged, reused, and repurposed, robust mechanisms for citation are necessary. Transferring bibliographic citation principles to data must be done carefully and selectively, lest the problems associated with citation practice be exacerbated and new ones introduced. Determining how to cite data is a non-trivial matter. Giving credit for data, which is among the arguments for data citation, also raises the complex ethical and policy issues associated with the use of bibliometrics for evaluating scholarship (Declaration on Research Assessment, 2013; Furner, 2014; Rafols, de Rijcke, & Wouters, 2014). This festschrift chapter, which is informed by several decades of discussion with Blaise Cronin, explores the thorny relationships between citing publications and citing data, asking how theories of bibliometrics might be applied to the use of research data and vice versa.

In a festschrift chapter for Eugene Garfield, edited some 15 years ago by Blaise Cronin and Helen Barsky Atkins (Borgman, 2000b), I expressed concerns about the slow uptake of bibliometrics to study scholarly communication and about the lack of understanding about how bibliometrics could be applied to electronic publishing. In the time since, the use of bibliometrics and webometrics to study scholarly communication in digital environments has blossomed. Theory and method in these areas also is far more mature (Almind & Ingwersen, 1997; Borgman, 1990; Borgman & Furner, 2002; Cronin & Sugimoto, 2014a, 2014b; Thelwall, Vaughan, & Bjorneborn, 2005). At this juncture, my concerns address how little is understood about the implications of data citation for the theory, method, and practice of bibliometrics – and conversely, how theories of bibliometrics can inform the design of citation mechanisms for data.

A Short History of Data Citation

The open access movement, writ large, is about facilitating the movement of publications, software code, government data, research data, and other intellectual content with minimal licensing restrictions and minimal costs (Kelty, 2008; "Open Knowledge Foundation," 2013; Suber, 2012). In the realm of scholarly communication, open access to publications and to data is being promoted or required by funding agencies, journals, universities, and other stakeholders. Adoption of these open access policies varies widely among fields, countries, institutions, and individuals. The biosciences, especially the "omics" fields, have adopted open data policies most fully. Genomic sequence data, for example, are submitted to repositories in concert with submitting articles for publication. Journals may require evidence of deposit, such as a record number, to consider the article for review. In most other fields, deposit of data is uneven at best, whether due to a lack of repositories, resources, skills, or incentives (Borgman, 2015; Fecher, Friesike, & Hebing, 2015; Kratz & Strasser, 2015; Wallis, Rolando, & Borgman, 2013).

Communities that value data as scholarly products to be shared, disseminated, recombined, and reused need ways to describe those data. The first method proposed, naively, was simply to map established mechanisms of bibliographic citation to data citation. The primary problem with this approach, as discussed below, is the lack of agreement on what constitutes data. A second problem, also discussed below, is the distinction between credit and attribution for data, hence the broader title of the research agenda workshop held by the U.S. National Academies of Science, "For Attribution—Developing Data Attribution and Citation Practices and Standards" (Uhlir, 2012). The workshop explicated a wide range of conceptual issues involved in the citation and attribution of data, allowing the work of the international CODATA-ICSTI Task Group to move forward. The report of that group promulgated a set of ten principles but did not establish an implementation plan. The diversity of constituencies and practices was deemed too great to be resolved by the Task Group alone (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013). However, other parties joined the effort quickly. The several reports, in combination with a manifesto (Crosas et al., 2013), provided the foundation for the community to implement the recommendations. Members of the Task Group, most of whom were practitioners from libraries, archives, data repositories, policy and standards agencies, and publishers, joined other stakeholders to refine the ten principles into a more succinct list of eight (Datacitation Synthesis Group, 2014). These principles, now finalized and endorsed by many parties, are quoted in full as a means to explore the comparisons between data and publications:

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 life span 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 verifying 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.

These principles map to the general functions of bibliographic citation, with concerns for documenting evidence, accommodating variant practices among communities, identifying cited items unambiguously, and improving access to the cited objects. Two functional differences have notable ramifications for the theory, method, and practice of bibliometrics. One is the assumption that referencing and cited objects are in digital form and available online. A data citation is much more than descriptive metadata; it should support machine action (principle 4). The second, made most explicit in principle 7, is that a data citation should facilitate access to related objects. Data may be interpretable only in combination with contextual information, and perhaps with software code, instrumentation, and other technologies. In contrast, publications are presumed to be interpretable as independent units.

Working groups on the dissemination and implementation of the data citation principles were established under the auspices of *Force11*, "a community working together in support of the goal of advancing scholarly communication" (Force11, 2015). Data citation is but one of their topics of interest. This volunteer community has weekly conference calls and daily flows of email. Some of their activities overlap with that of *Research Objects for Scholarly Communication* (ROSC), a burgeoning eScience community established in 2014 under the auspices of the World Wide Web Consortium (Bechhofer, De Roure, Gamble, Goble, & Buchan, 2010; "Research Object for Scholarly Communication Community Group," 2014). The *Research Data Alliance* (RDA), a more formal organization that has funding from public agencies in the U.S., Europe, and Australia, has working groups, interest groups, and birds-of-a-feather groups that intersect with the concerns of Force11 and ROSC. RDA, established in 2013, has more than 1600 members from 70 countries. While the overlap in membership is considerable, RDA draws practitioners, technologists, and policy makers interested in building infrastructures for data management; Force11 is concerned with reforming scholarly communication generally; and Research Objects for Scholarly Communication is concerned with technical approaches for managing data, publications, software, and other objects created in scientific research.

Among the interests common to these groups, and to others within individual domains, are the desire to redesign scholarly communication for networked environments, the changing relationships among stakeholders, the changing criteria for evaluating scholarship, and the complexity of data management and stewardship. A critical mass of stakeholders now considers these to be urgent problems. Data citation is but one mechanism to address these issues, albeit a fairly central one to the extent that it facilitates credit, attribution, discovery, access, retrieval,

management, use, and stewardship of scholarly content. These developments should be watched closely by bibliometricians, given the broad implications for scholarly communication. "Data citation" has become a catchword that encompasses a larger array of issues involved in managing the many digital objects that are created or used in research.

Theoretical Problems of Data Citation

Scholarly authors are expected to document their evidence by citing their sources. Bibliographic referencing, the traditional means to do so, matured in an era of print publication. Books, articles, and other scholarly products were stable entities. Once published, they stayed published. Given adequate bibliographic description, most cited documents could be located in research libraries, or perhaps in archives. As publication moved to digital formats, first as duplicates to print publication, later as a primary format, the stability of documents and citations no longer could be assumed.

Data are much different entities than publications, introducing many new features and requirements for citations. In turn, these different characteristics require a new set of theoretical premises for bibliometrics. Modeling the flows of data alone would be hard enough. To the extent that data are cited as objects on par with publications, bibliometric analyses will draw upon heterogeneous pools of cited entities. Thus it is useful to consider how citation practices differ between genres of publication and of data.

Generally speaking, authors cite sources that are accessible to their readers. In most cases, they cite other publications, providing enough information that readers can locate those sources in a library or online. The publication to which a citation refers may exist in many copies. Metadata elements such as volume, issue, and page numbers usually suffice to identify the item uniquely, whether in a print issue or online. Even if the document was obtained online, citations may reference the page numbers of the printed copy. When cited objects are available only online, location information such as URLs, or unique and persistent identification such as digital object identifiers (DOI) that can resolve to a location, are required. Publications usually are assumed be static objects, which facilitates identity and location. In cases where cited objects are not assumed to be stable, a specific version can be cited. Although links to online publications may break fairly quickly, these objects tend to remain available somewhere, and discovery mechanisms are improving (Klein & Nelson, 2010; Van de Sompel et al., 2012; Van de Sompel, Nelson, & Sanderson, 2013)

Once outside the realm of formal publication, citations become less reliable means to locate sources of evidence. Authors may cite rare or original sources but not include the name and location of the archive in their bibliographic descriptions. Authors rarely provide bibliographic citations for their own data unless they are depositing those data in a place accessible to readers. Rather, most authors describe their methods and data to the degree expected by their field and publication venue, providing tables, figures, and supplementary materials as appropriate. In cases where a publication draws on data from external sources, such as those from an archive, repository, or colleague, those data may or may not be referenced. Data in repositories are most easily cited, as these institutions usually offer suggested citation formats that include unique and persistent identifiers. However, if external data were obtained to calibrate instruments or to "ground truth" a field site, they may not be cited because they were considered background to the research or implicit in the methods (Wallis et al., 2013; Wynholds, Wallis, Borgman, Sands, & Traweek, 2012). In other cases, authors might cite a "data paper" associated with a data release, as in astronomy (Ahn et al., 2012), an entire archive (e.g., Sloan

Digital Sky Survey), or publisher of data sources (e.g., OECD). References to data often are informal, such as a URL, a footnote, a figure caption, or an oblique mention in a sentence (Pepe, Goodman, Muench, Crosas, & Erdmann, 2014). Links to data decay even more quickly than do links to publications, as researchers are much less likely to curate data for long periods of time. The eternal quest for bibliographic control (Borgman, 2000a) is even more ephemeral for data than for publications.

Stakeholders and Styles

A particular challenge in building bibliometric theory for data citation is the number of stakeholders involved. These include, for example, scholars, publishers, librarians, funders, repository managers, policy makers, and technologists. Each has different interests in the forms that data citation will take. Some would make credit and attribution the highest priorities; others would focus on data citation as a means to improve discovery and access. The diversity of publication manuals and bibliographic citation styles suggests that achieving unity in data citation is highly unlikely. Bibliographic referencing tools such as Zotero, Endnote, Refworks, and Mendeley provide style sheets that will render citations in the formats of individual journals, conferences, and publishers. For example, Zotero currently supports 7,429 citation styles (Zotero, 2015). Only a few fields and journals have established citation styles for data.

The tensions are many. As discussed further below, scholars want credit for their scholarly work, but do not necessarily desire separate credit for their data. Most lack the skills, resources, and often motivation to invest in curating their own data well enough to make them citable. Search engines would like to add value to existing assets by making them more discoverable. Funding agencies may require that data resulting from projects they support be shared and reused, but few such agencies have been willing to invest heavily in data stewardship. Overall, better knowledge infrastructures are needed to manage, discover, and exploit research data and information (Borgman, 2015; CrossRef, 2013; Edwards et al., 2013).

Commercial interests see opportunities in hosting and providing access to data. Cloud computing services will host data, but do not wish to be in the curatorial business. Publishers may provide access to data as value-added services, but few are willing to host data except as a for-profit venture. Data repositories, which typically are non-profit consortial organizations, are concerned about their long-term ability to curate resources in the face of commercial competition that may have a shorter view. Universities seek better records of the scholarly output of their faculty, students, and research staff for use in promoting their reputations, managing their resources, and evaluating people and departments. Research libraries see a role in curating the data produced by researchers in their universities or other organizations, but may not wish to compete with repositories. Rather, libraries are more likely to apply their expertise in information organization, curation, and discovery to orphan data.

Each of these stakeholders addressed their concerns independently until 2005 or so. As interest grew in data management plans, data sharing, reuse, and citation, competing stakeholders began to see some common ground. Influential policy documents helped to lay foundations for further discussion (Atkins et al., 2003; Boulton et al., 2012; Bourne et al., 2011; CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013; Hey & Trefethen, 2005; National Science Board, 2005; Uhlir, 2012; Wood et al., 2010). Some of these documents were consensus reports; others resulted from conferences and workshops on several continents. Coalitions such as Force11 and RDA bring competing stakeholders to the same table to discuss the future of scholarly communication, including access to data (Borgman, 2015).

Defining Data

At the core of the data citation problem is the lack of agreement on what constitutes data. Despite the plethora of policies and press about data, big and small, little effort is devoted to defining these terms. This is not a new problem. As Rosenberg (2013) comments, histories of science and epistemology tend to mention data only in passing, if at all (Blair, 2010; Daston, 1988; Poovey, 1998; Porter, 1995). Foundational works on the making of meaning in science discuss facts, representations, inscriptions, and publications, with little attention to data per se (Bowker, 2005; Latour, 1987, 1988, 1993; Latour & Woolgar, 1979). Bibliometricians, as members of the information sciences, are well aware of the difficulties in defining "information" (Buckland, 1991; Case, 2002, 2012; Furner, 2010). Precise operational definitions of the units being cited are necessary for bibliometrics, and particularly for machine discovery of cited objects. Attempts to distinguish between data and datasets have not achieved much clarity, as notions of the identity of datasets pose other theoretical challenges (Agosti & Ferro, 2007; Renear & Dubin, 2003; Renear, Sacchi, & Wickett, 2010).

The definition proposed elsewhere is suitable for discussions of bibliometrics and data citation: *Data* refers to entities used as evidence of phenomena for the purposes of research or scholarship (Borgman, 2015). The advantages of this definition are several. It recognizes the degree to which data may exist in the eye of the beholder. One person's signal is another's noise. Thus, one set of entities could be used for evidence of different phenomena for different purposes. In scientific publications, authors may consider their data to be the tables and figures presented, the cleaned and analyzed data set from which those tables and figures were derived, the initial "raw" observations from the field or instrument—or all or none of these. Collaborators may reasonably disagree on what were the data from any given field site, experiment, or study (Borgman, Wallis, & Mayernik, 2012). Thence comes the problem of granularity. A data citation might refer to one or a few observations, to a dataset assembled over the course of a career, or anything in between. The essence of principle 7 is that citations can be made to whatever unit of data is appropriate evidence in a particular case. The citation should be unique, as stated in principle 4.

The granularity problem also arises in bibliographic citations. Scientific styles tend to cite entire documents, whereas humanities styles tend to cite individual page numbers or passages. Often these variant forms can be reconciled if enough metadata is provided; e.g., author, title, date, page numbers. While bibliometric analyses often aggregate documents by author, institution, journal, date range, or other elements, the unit of analysis is usually the cited document (Borgman, 1990). Similarly, most namespaces for publications are based on the publication as the basic unit—ISBN, LCCN, DOI, etc. As DOIs are assigned to articles, to data, and to individual tables and figures within articles, identification and retrieval are further complicated. Determining the "version of record" is ever more difficult in digital environments. The technological solution may be to reconcile "versions of the record" (Van de Sompel, 2013).

Provenance

Principles 4 and 5, on unique identification and access, and principle 8, on interoperability and flexibility, indicate the need for provenance information. Data citations can facilitate provenance, but may not be able to incorporate all the necessary content and context. Provenance is both more and less than metadata. It involves the origin and history of something, and documentation of the chain of evidence, custody, and relationships to other entities (Borgman, 2015; Buneman,

Khanna, & Tan, 2001; Carata et al., 2014; Groth, Gil, Cheney, & Miles, 2012; Groth & Moreau, 2013).

Rarely can data be interpreted without provenance information such as research methods, protocols, and the software necessary to open a file or run the program. Data continue to change form and meaning as they are processed, mined, aggregated, and disaggregated. The farther that reusers are from the origins of data, whether in terms of time, theory, geography, domain, or other factors, the more reliant they may be on provenance documentation. Provenance records may provide the information necessary to interpret, trust, or determine the legal rights to reuse, repurpose, or combine datasets—the evidentiary chain. If data creators are to receive credit through citation, that credit must carry forward through subsequent reprocessing. Sustaining the provenance chain is a daunting technical challenge. Provenance chains will evolve over time as more relationships are accrued and as links break. Provenance may also pose the greatest theoretical challenge, as authors, readers, and later analysts encounter substantially different aggregations of objects over time (Pepe, Mayernik, Borgman, & Van de Sompel, 2010).

Releasing, Sharing, and Reusing Data

Authors cite evidence that is available to readers, so readers can evaluate that evidence. Thus, determining what data to cite is partly a function of what data are released and made publicly available. Little is understood about what data scholars choose to share or about how, when, and why they reuse data. Data sharing and reuse are topics ripe for research and theorizing (Borgman, 2012, 2015).

Theoretical questions persist about what objects scholars choose to cite in any given publication or about the meaning of individual citations, despite decades of empirical research and theoretical development (Cronin, 1981). Citation practices are more learned than taught. Publication manuals and "instructions to authors" in journals provide explicit instructions on how to cite sources in specific styles, but offer little guidance on what to cite. One commonality among data citation practice, data sharing, and reuse is that these are localized behaviors that are difficult to articulate.

Data sharing and reuse rest heavily on trust between the parties involved. Data repositories are intermediaries in the trust relationship between those who give data and those who receive. Data citation is one mechanism to document those relationships. Citing data already stored in repositories is the low hanging fruit for data citation, and a starting point for initiatives such as DataCite (DataCite, 2013). Unique and persistent identifiers, stable and persistent links between related digital objects, digital signatures that verify the integrity of digital objects, and similar mechanisms contribute to the trust fabric. No matter how sophisticated the technology, trust is based in the individuals and the social institutions involved (Blanchette, 2012).

The ability to share and reuse data rests on early decisions about how to describe and manage them. The earlier in the process that scholars document data in ways that make them reusable, the better they can represent them as citable objects. Data citation mechanisms can support these functions, although individual citations, per se, are unlikely to carry enough information to interpret data or to document provenance.

Credit

Assigning credit for data is even more problematic than is assigning credit for authoring publications. Contemporary authorship is negotiated with collaborators or determined by the policy of the parent organization. Policies of publishers and journals also may influence the

designation of author or contributor roles. Notions of authorship credit appear to vary widely between domains, as Cronin has shown (Cronin, 1984, 1995, 1998, 2001, 2005, 2008; Davenport & Cronin, 2001). Policies at CERN, for example, are intended to provide authorship credit for early contributions to data collection, thus conflating credit for data and for publication (Mele, 2013). In space-based astronomy missions, decisions about what data to collect, how to collect them, and how to process them are made many years before researchers use those data in publications. Data papers and instrument papers are the means by which those involved early in the process get credit for their contributions. By the time those data are used by later astronomers, individuals responsible for creating the data may be invisible, anonymous, or departed (Borgman, 2015).

In smaller teams, authorship is negotiated, but credit for data is not usually part of the discussion. "Authorship" is not terminology that resonates with scholars when thinking about their data (Wallis, 2012; Wallis & Borgman, 2011). Data may not be released because the responsibility for data is so diffuse that no individual is empowered or motivated to do so. The larger the collaboration, the less familiarity the principal investigator (PI) may have with the specifics of the data collection, and the greater likelihood that the PI has the long-term responsibility for a diffuse organization. The students and post-doctoral fellows who have the most intimate knowledge of the data have the highest turnover rate as team members. The PI may be responsible for stewardship of the data, which deserves credit. Those who have the most intimate familiarity with the data possess tacit knowledge that is necessary for interpretation, which also deserves credit. Expertise, responsibility, and authorship are not equivalent with respect to data; it is unclear how credit should be allocated in each instance

The workshop conducted by the National Academies and the CODATA-ICSTI Task Group sought input from many stakeholders about how to assign credit for data. While the starting assumption was that scholars cared the most about receiving credit for their data, it became clear over the two days of discussion that many other parties also wanted credit: funding agencies who supported the research; data repositories who acquire, curate, and release data; university research officers; and other data providers (Uhlir, 2012). Authors want credit for citations to their publications, as these are currency for hiring and advancement; thus, citing publications as surrogates for the data reported in them suits the interests of most authors. If datasets are cited instead of publications, authors may have disincentives for citing data. Researchers usually receive more credit for citations to peer-reviewed publications than for other activities such as teaching, editorial work, or service. Where citations to data, or to other non-peer-reviewed objects, fall on this credit spectrum is unknown, but it appears that any practice that risks diluting credit for publications may be viewed with suspicion.

Attribution of Sources

Attribution of the sources for data is equally problematic to credit. Agencies providing data commonly do so under licenses that constrain who can use the data, for what purposes, for how long, and with what attribution (Pearson, 2012). Researchers often place restrictions on the sharing and reuse of their data, whether by licensing or other means. They may require a specific citation to data. If they use Creative Commons licenses, they may specify whether the dataset (or other object) may be used only as a whole or whether in parts, for commercial or non-commercial purposes, and the form of attribution required (Creative Commons, 2013). While the desire for control is understandable, due to concerns for intellectual property, credit, and misuse or misinterpretation, attribution requirements complicate reuse considerably. Licensing also

makes the process of combining and reusing data more complex, if attributions must be carried forward in provenance records (Guibault, 2013).

Due to these complications, many have argued for the open release of data without licensing restrictions, or for direct release into the public domain (Murray-Rust, Neylon, Pollock, & Wilbanks, 2010; Nielsen, 2011; Wilbanks, 2013). Releasing data openly without restrictions and without requiring credit or attribution certainly simplifies data sharing and reuse. However, it runs counter to the interests of most scholars. Documenting data for reuse often requires considerable investment of resources. Data can be assets to be controlled, protected, exchanged, and bartered for other resources, including academic posts (Borgman, 2015; Hilgartner & Brandt-Rauf, 1994; Latour, 1987; Latour & Woolgar, 1979). Credit and attribution may be insufficient rewards for scholars to relinquish those assets or to expose themselves to potential liabilities associated with reuse.

Discovery

While discovery is not mentioned explicitly in the data citation principles, it is implicit throughout. Describing data in sufficient detail to ensure unique identification, persistence, specificity, verifiability, and interoperability will improve their discovery. Discovery is a precondition for gaining access to the desired data or other objects (principle 5). Similar requirements apply in using bibliometrics to discover, locate, and retrieve publications (Cronin, 2014; Kousha & Thelwall, 2014). In principle 4, "machine actionability" implies that a citation should support machine discovery of data. Rather than a citation providing enough information to search the shelves of a library, it should have embedded links that allow direct access to the referenced object. Digital object identifiers (DOIs), in combination with technical standards such as OpenURL and publisher-led initiatives such as CrossRef, facilitate machine actionability from citations in online articles to cited articles. However, many such links, whether for publications or for data, lead to a landing page where a human can identify the object of interest. When the searcher is a computer, these discovery mechanisms fail (Van de Sompel, 2012). A goal for the next generation of search technologies, especially for data discovery, is to support machine actionable links for entire provenance chains (Bechhofer et al., 2010; CrossRef, 2009, 2014; Klein et al., 2014; Pepe et al., 2010; Sanderson & Van de Sompel, 2012; Simons, 2012; Van de Sompel, 2015; Van de Sompel, Hochstenbach, & Beit-Arie, 2000; Van de Sompel & Lagoze, 2009).

At present, most data discovery appears to be a fairly manual process. Individuals identify data of interest by reading papers or by searching repositories. Discovery can be improved by more extensive description of data, figures, tables, and other elements in publications. Such descriptions can be accommodated by open annotation systems that facilitate interoperability across systems, which includes synchronizing links to related resources (Ciccarese, Ocana, & Clark, 2012; Das et al., 2009; Foster & Moreau, 2006; Hunter, 2009; Van de Sompel et al., 2012). These approaches may be effective to the extent that document enrichment survives the publication process. Publishers tend to "flatten out" submitted documents by reducing them to portable document format (PDF), which is a proprietary standard, for ingest to their systems. In that process, they usually strip annotations, citation records stored in the document by bibliographic referencing tools, and other features that support machine actions. New platforms, unencumbered by legacy publishing systems, may enrich research objects in ways that support more robust discovery and more complex document structures.

Initiatives such as DataCite, Schema.org, and Object Reuse and Exchange are developing metadata schema for data. Search engines, which have largely ignored metadata and other forms of document enrichment, are implementing more structured methods. Provenance graphs are essential for data management at scale (DiLauro, 2013). Approaches that publish graphs of object relationships will aid both discovery and bibliometrics. Proprietary control of these graphs will hide those chains of evidence from other programs, users, and bibliometricians. To the extent that these metadata schema are adopted, and especially to the extent that graphs are open, they should aid in discovery of research data (DataCite, 2014; "Object Reuse and Exchange," 2014; Schema.org, 2012; World Wide Web Consortium (W3C), 2013). However, individual researchers tend to invest very little effort in providing metadata or other curatorial description to their data to make them usable for others. Labor and skill requirements to do so are high and incentives are low. Because digital data are far less self-describing than are textual objects such as publications, discovery depends heavily on metadata. Thus data citation is implicitly a means to improve the discovery of data (Borgman, 2015).

Discussion and Conclusion

Data are not equivalent to publications, hence data citation is not equivalent to bibliographic citation. However, theories of bibliographic citation are useful in thinking about what data citation is, or could be. The most fundamental distinction between bibliographic objects and data is the degree of independence. Bibliometrics – including scientometrics, informetrics, webometrics, altmetrics, and other variants – are used to model relationships between objects that can be treated as independent entities, whether web pages or tweets. Data, however, rarely can be interpreted as independent objects. Most are meaningless without links to contextual information, software, and related objects. The scope and identity of a dataset vary along multiple dimensions. Without agreements on what constitutes data in any given instance, it is difficult to count or compare the uses of those data. Empirical and theoretical work on what are data, how those data are used, how data are aggregated and disaggregated, and when and how they are cited as sources of evidence are avenues ripe for exploration at the intersection of scholarly communication and bibliometrics.

Blaise Cronin was among the first to call for a theory of citation behavior, as so little is understood about the purposes for which an object is cited (Cronin, 1981). Data citation exacerbates that theoretical challenge. References to articles are sometimes surrogates for citing the data within them. When data sets are accessible, those can be cited. Later authors who use those data may cite the dataset, a larger dataset or repository from which the data were drawn, articles in which the datasets were discussed, or some combination of these. Early efforts to classify the purposes for individual citations revealed that article citations are sometimes data citations (Lipetz, 1965; White, 1982). When data are cited, it is often not in the reference list, but buried in footnotes, URLs, or mentions in text. The difficulty of identifying citations to data is not new, but demands to standardize and promulgate data citation increase the urgency of addressing the problem.

Distinguishing between data citation and data use is another thorny theoretical challenge. The few studies on data reuse indicate that scientifically important uses of data may not be mentioned or cited in publications (Palmer, Weber, & Cragin, 2011; Wallis et al., 2013; Wynholds et al., 2012). The reasons for lack of citation are many. One is that data citation is not (yet) common scholarly practice. Another may be that many sources are used in research, but few are cited. Views, downloads, library reshelving statistics, and other measures of use tend not

to correlate well with citations of those same items (Bollen, Van de Sompel, & Rodriguez, 2008; Haustein, 2014; Kousha & Thelwall, 2014; Thelwall, Haustein, Larivière, & Sugimoto, 2013). Again, it has long been known that reading, library use, and citation are different behaviors. How those differences translate to the use, reuse, and citing of data is unexplored territory.

Another opportunity for theory building in bibliometrics posed by data citation is the changing notion of authorship. While authorship was never a stable concept, as Blaise Cronin has shown (Cronin, 1981, 1995, 2001, 2003, 2005, 2008, 2013, 2014; Davenport & Cronin, 2000), practical concerns for credit and attribution have focused largely on the roles of individuals in scholarly communication. Some journals ask for precise descriptions of the contributions of each named author; e.g., writing, data collection, data analysis, instrumentation, and management (Committee on Publication Ethics, 2013; Harvard University and Wellcome Trust, 2012). The work associated with collecting, cleaning, analyzing, managing, and reporting data is essential to the conduct of scholarly research, but that work is not necessarily equivalent to authorship. How these roles should be credited in data citation, and how they should be weighted in contributions to scholarship are open questions. The labor associated with data management and software engineering tends to be lower in status than the scientific work that leads to peer-reviewed papers (Darch et al., 2015).

Lastly, the intersection of data citation and scholarly communication is an example of the uneasy fit between structure and process in scholarly communication. While research on process should inform research on structure, and vice versa, rarely do these approaches intersect (Lievrouw, 1990). Bibliometrics and their brethren address structural relationships in scholarly communication. The validity of these analyses rests on understanding the processes by which these structures arise and evolve. A better understanding of the processes associated with the creation, use, and reuse of data, should lead to the design of better data citation mechanisms.

Bibliometricians are in an ideal position to contribute to—and to learn from—the development of theory and practice in data citation. The caveat is in the title of this chapter. Bibliometrics, strictly speaking, are based on publications. Data are not publications, therefore data citation is something other than bibliometrics. However, data most certainly are objects exchanged in scholarly communication. Theoretical approaches to data citation must accommodate the ways in which data differ from publications. Data tend to be compound objects with unclear boundaries, whereas publications can be treated as independent objects with clear boundaries, at least for the purpose of bibliometrics. Data usually consist of multiple objects that are interdependent, with relationships that often are unstable and difficult to document. Theory and methods from bibliometrics, scientometrics, and webometrics can be used to study the characteristics of these relationships and how they evolve over time. The "catch-22" is that it will be difficult to model these relationships until units of data are sufficiently documented to be traceable. This is an opportune moment for those concerned with data, scholarly communication, knowledge infrastructures, and bibliometrics to explore common ground. Blaise Cronin has laid the foundation that allows this conversation to move forward.

Cited References

Agosti, M., & Ferro, N. (2007). A formal model of annotations of digital content. *ACM Transactions on Information Systems*, *26*(1). http://doi.org/10.1145/1292591.1292594

Ahn, C. P., Alexandroff, R., Allende Prieto, C., Anderson, S. F., Anderton, T., Andrews, B. H., … Zinn, J. C. (2012). The Ninth Data Release of the Sloan Digital Sky Survey: First

Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey. *Astrophysical Journal*, *203*, 21. http://doi.org/10.1088/0067-0049/203/2/21

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to "webometrics." *Journal of Documentation*, *53*(4), 404–426. http://doi.org/10.1108/EUM0000000007205

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., … Wright, M. H. (2003). *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure*. Washington, DC: National Science Foundation. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*. http://doi.org/10.1038/npre.2010.4626.1

Blair, A. M. (2010). *Too Much to Know: Managing Scholarly Information before the Modern Age*. New Haven, CT: Yale University Press.

Blanchette, J.-F. (2012). *Burdens of Proof: Cryptographic Culture and Evidence Law in the Age of Electronic Documents*. Cambridge, MA: The MIT Press.

Bollen, J., Van de Sompel, H., & Rodriguez, M. A. (2008). Towards Usage-based Impact Metrics: First Results from the MESUR Project. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 231–240). Pittsburgh, PA: Association for Computing Machinery.

Borgman, C. L. (1990). Editor's introduction. In *Scholarly Communication and Bibliometrics* (pp. 10–27). Newbury Park, CA: Sage.

Borgman, C. L. (2000a). *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: MIT Press.

Borgman, C. L. (2000b). Scholarly communication and bibliometrics revisited. In *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 143–162). Medford, NJ: Information Today.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

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

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge MA: MIT Press.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. In *Annual Review of Information Science and Technology* (Vol. 36, pp. 3–72).

Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2012). Who's Got the Data? Interdependencies in Science and Technology Collaborations. *Computer Supported Cooperative Work*, *21*(6), 485–523. http://doi.org/10.1007/s10606-012-9169-z

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., … Walport, M. (2012). *Science as an Open Enterprise*. The Royal Society. Retrieved from http://royalsociety.org/policy/projects/science-public-enterprise/report/

Bourne, P. E., Clark, T., Dale, R., de Waard, A., Hovy, E. H., & Shotton, D. (Eds.). (2011). Force 11 Manifesto: Improving Future Research Communication and e-Scholarship. Retrieved from http://www.force11.org/white_paper

Bowker, G. C. (2005). *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science*, *42*, 351–360.

Buneman, P., Khanna, S., & Tan, W.-C. (2001). Why and Where: A Characterization of Data Provenance. In J. V. den Bussche & V. Vianu (Eds.), *Database Theory — ICDT 2001* (Vol. 1973, pp. 316–330). Berlin: Springer.

Carata, L., Akoush, S., Balakrishnan, N., Bytheway, T., Sohan, R., Selter, M., & Hopper, A. (2014). A primer on provenance. *Communications of the ACM*, *57*(5), 52–60. http://doi.org/10.1145/2596628

Case, D. O. (2002). *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. San Diego: Academic Press.

Case, D. O. (2012). *Looking for Information: a survey of research on information seeking, needs and behavior* (3rd ed.). Bingley, UK: Emerald Group Publishing.

Ciccarese, P., Ocana, M., & Clark, T. (2012). Open semantic annotation of scientific publications using DOMEO. *Journal of Biomedical Semantics*, *3*(Suppl. 1), S1. http://doi.org/10.1186/2041-1480-3-S1-S1

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, *12*, 1–75. http://doi.org/10.2481/dsj.OSOM13-043

Committee on Publication Ethics. (2013). [Home page]. Retrieved September 9, 2013, from http://publicationethics.org/

Creative Commons. (2013). Creative Commons License Choices. Retrieved October 2, 2013, from http://creativecommons.org/choose/

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation*, *37*(1), 16–24. http://doi.org/10.1108/eb026703

Cronin, B. (1984). *The Citation Process: The Role and Significance of Citations in Scientific Communication*. London: Taylor Graham. Retrieved from http://garfield.library.upenn.edu/cronin/citationprocess.pdf

Cronin, B. (1995). *The scholar's courtesy: The role of acknowledgement in the primary communication process. London: Taylor Graham.* London: Taylor Graham.

Cronin, B. (1998). Metatheorizing citation. *Scientometrics*, *43*(1), 45–55. http://doi.org/10.1007/BF02458393

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569. http://doi.org/10.1002/asi.1097

Cronin, B. (2003). Scholarly communication and epistemic cultures. In *Scholarly Tribes and Tribulations: How Tradition and Technology are Driving Disciplinary Change*. Washington, DC: Association of Research Libraries. Retrieved from http://www.arl.org/scomm/disciplines/Cronin.pdf

Cronin, B. (2005). *The Hand of Science: Academic Writing and its Rewards*. Lanham, MD: Scarecrow Press.

Cronin, B. (2008). Toward a Rhopography of Scholarly Communication. *Studia Humaniora Ouluensis*.

Cronin, B. (2013). Self-plagiarism: An odious oxymoron. *Journal of the American Society for Information Science and Technology*, *64*(5), 873–873. http://doi.org/10.1002/asi.22966

Cronin, B. (2014). Scholars and scripts, spoors and scores. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Metrics-Based Evaluation of Research* (pp. 3–21). Cambridge, MA: MIT Press.

Cronin, B., & Sugimoto, C. R. (2014a). *Beyond Bibliometrics: Metrics-Based Evaluation of Research*. Cambridge, MA: MIT Press.

Cronin, B., & Sugimoto, C. R. (Eds.). (2014b). *Scholarly Metrics Under the Microscope: Citation Analysis and Academic Auditing*. Medford, NJ: Information Today.

Crosas, M., Carpenter, T., Shotton, D., & Borgman, C. L. (2013). Amsterdam Manifesto on Data Citation Principles. Presented at the Force11: Beyond the PDF 2 Conference, Amsterdam. Retrieved from https://www.force11.org/AmsterdamManifesto

CrossRef. (2009). *The Formation of CrossRef: A Short History*. CrossRef. Retrieved from http://www.crossref.org/01company/02history.html

CrossRef. (2013). FundRef. Retrieved September 30, 2013, from http://www.crossref.org/fundref/

CrossRef. (2014). Home page. Retrieved May 26, 2014, from http://www.crossref.org/

Darch, P. T., Borgman, C. L., Traweek, S., Cummings, R. L., Wallis, J. C., & Sands, A. E. (2015). What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. *International Journal on Digital Libraries*, 1–17. http://doi.org/10.1007/s00799-015-0137-3

Das, S., Girard, L., Green, T., Weitzman, L., Lewis-Bowen, A., & Clark, T. (2009). Building biomedical web communities using a semantically aware content management system. *Briefings in Bioinformatics*, *10*(2), 129–138. http://doi.org/10.1093/bib/bbn052

Daston, L. J. (1988). The Factual Sensibility. *Isis*, *79*(3), 452–467. http://doi.org/10.2307/234675

Datacitation Synthesis Group. (2014). Joint Declaration on Data Citation Principles - Final. Retrieved February 12, 2014, from http://www.force11.org/datacitation

DataCite. (2013). [Home page]. Retrieved September 10, 2013, from http://www.datacite.org/

DataCite. (2014). DataCite Schemas repository. Retrieved February 12, 2014, from http://schema.datacite.org/meta/kernel-3/index.html

Davenport, E., & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 517–534). Medford, NJ: Information Today.

Davenport, E., & Cronin, B. (2001). Who dunnit? Metatags and hyperauthorship. *Journal of the American Society for Information Science and Technology*, *52*(9), 770–773. http://doi.org/10.1002/asi.1123

Declaration on Research Assessment. (2013). [Home page]. Retrieved May 24, 2013, from http://am.ascb.org/dora/

DiLauro, T. (2013). *Research data management experience at Johns Hopkins University Sheridan Libraries*. Retrieved from https://docs.google.com/file/d/0B1X7I2IVBtwzVTgxczJzVUFMMnM/edit

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Ann Arbor: University of Michigan. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/97552

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? *PLoS ONE*, *10*(2), e0118053. http://doi.org/10.1371/journal.pone.0118053

Force11. (2015). Home page. Retrieved August 6, 2014, from https://www.force11.org/about

Foster, I., & Moreau, L. (2006). *Provenance and Annotation of Data*. Heidelberg: Springer. Retrieved from http://www.w3.org/2011/prov/wiki/Connection_Task_Force_Informal_Report

Furner, J. (2010). Philosophy and information studies. *Annual Review of Information Science and Technology*, *44*(1), 159–200. http://doi.org/10.1002/aris.2010.1440440111

Furner, J. (2014). The ethics of evaluative bibliometrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Metrics-Based Evaluation of Research* (pp. 85–107). Cambridge, MA: MIT Press.

Groth, P., Gil, Y., Cheney, J., & Miles, S. (2012). Requirements for Provenance on the Web. *International Journal of Digital Curation*, *7*(1), 39–56. http://doi.org/10.2218/ijdc.v7i1.213

Groth, P., & Moreau, L. (2013). PROV-Overview. Retrieved April 14, 2014, from http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Guibault, L. (2013). Licensing research data under open access conditions. In D. Beldiman (Ed.), *Information and Knowledge: 21st Century Challenges in Intellectual Property and Knowledge Governance*. Cheltenham: Edward Elgar.

Harvard University and Wellcome Trust. (2012). *International Workshop on Contributorship and Scholarly Attribution*. Retrieved from http://projects.iq.harvard.edu/files/attribution_workshop/files/iwcsa_report_final_18sept12.pdf

Haustein, S. (2014). Readership metrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Metrics-Based Evaluation of Research* (pp. 327–344). Cambridge, MA: MIT Press.

Hey, T., & Trefethen, A. (2005). Cyberinfrastructure and e-Science. *Science*, *308*, 818–821. http://doi.org/10.1126/science.1110410

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data access, ownership and control: Toward empirical studies of access practices. *Knowledge*, *15*, 355–372.

Hunter, J. (2009). Collaborative semantic tagging and annotation systems. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 43, pp. 187–239).

International Council for Scientific and Technical Information. (2015). Retrieved February 17, 2015, from http://www.icsti.org/

Kelty, C. M. (2008). *Two Bits: the Cultural Significance of Free Software*. Durham, NC: Duke University Press.

Klein, M., & Nelson, M. L. (2010). Evaluating methods to rediscover missing web pages from the web infrastructure (p. 59). ACM Press. http://doi.org/10.1145/1816123.1816133

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE*, *9*(12), e115253. http://doi.org/10.1371/journal.pone.0115253

Kousha, K., & Thelwall, M. (2014). Web impact metrics for research assessment. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Metrics-Based Evaluation of Research* (pp. 289–306). Cambridge, MA: MIT Press.

Kratz, J. E., & Strasser, C. (2015). Researcher Perspectives on Publication and Peer Review of Data. *PLoS ONE*, *10*(2), e0117619. http://doi.org/10.1371/journal.pone.0117619

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.

Latour, B. (1988). Drawing things together. In M. E. Lynch & S. Woolgar (Eds.), *Representation in Scientific Practice* (pp. 19–68). Cambridge, MA: MIT Press.

Latour, B. (1993). *We Have Never Been Modern*. Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The Construction of Scientific Facts*. Beverly Hills, CA: Sage.

Lide, D. R., & Wood, G. H. (2012). *CODATA @ 45 Years: 1966 to 2010*. Paris: CODATA. Retrieved from http://www.codata.org/about/CODATA@45years.pdf

Lievrouw, L. A. (1990). Reconciling structure and process in the study of scholarly communication. In *Scholarly Communication and Bibliometrics* (pp. 59–69). Newbury Park, CA: Sage.

Lipetz, B.-A. (1965). Improvement of the Selectivity of Citation Indexes to Science Literature through Inclusion of Citation Relationship Indicators. *American Documentation*, *16*(2), 81–90.

Mele, S. (2013, May 23). Higgs Boson discovery at CERN: Physics and Publishing. Retrieved September 7, 2013, from http://www.oii.ox.ac.uk/events/?id=598

Murray-Rust, P., Neylon, C., Pollock, R., & Wilbanks, J. (2010). Panton Principles. Retrieved August 30, 2013, from http://pantonprinciples.org/

National Science Board. (2005). *Long-Lived Digital Data Collections*. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

Nielsen, M. (2011). *Reinventing Discovery: The New Era of Networked Science*. Princeton, NJ: Princeton University Press.

Object Reuse and Exchange. (2014). Retrieved from http://www.openarchives.org/ore/

Open Knowledge Foundation. (2013). Retrieved July 22, 2013, from http://okfn.org/

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–10. http://doi.org/10.1002/meet.2011.14504801174

Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*, WDS32–WDS46. http://doi.org/10.2481/dsj.WDS-042

Pearson, S. H. (2012). Three legal mechanisms for sharing data. In P. F. Uhlir (Ed.), *For Attribution—Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop* (pp. 71–76). Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=13564&page=71

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLoS ONE*, *9*(8), e104798. http://doi.org/10.1371/journal.pone.0104798

Pepe, A., Mayernik, M. S., Borgman, C. L., & Van de Sompel, H. (2010). From artifacts to aggregations: Modeling scientific life cycles on the semantic web. *Journal of the American Society for Information Science and Technology*, *61*, 567–582. http://doi.org/10.1002/asi.21263

Poovey, M. (1998). *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. Chicago: University of Chicago Press.

Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Pritchard, A. (1969). Statistical Bibliography or Bibliometrics. *Journal of Documentation*, *25*(4), 348–9.

Rafols, I., de Rijcke, S., & Wouters, P. (2014). The Leiden Manifesto in the making. Retrieved October 30, 2014, from http://www.cwts.nl/News/

Raisig, L. M. (1962). Statistical Bibliography in the Health Sciences. *Bulletin of the Medical Library Association*, *50*(3), 450–461. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC197860/

Renear, A. H., & Dubin, D. (2003). Towards identity conditions for digital documents. In *Proceedings of the 2003 international conference on Dublin Core and metadata applications: supporting communities of discourse and practice*. Seattle, Washington: Dublin Core Metadata Initiative.

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem* (Vol. 47, pp. 1–4). Medford, NJ: Information Today. http://doi.org/10.1002/meet.14504701240

Research Object for Scholarly Communication Community Group. (2014). Retrieved August 6, 2014, from http://www.w3.org/community/rosc/

Rosenberg, D. (2013). Data before the fact. In L. Gitelman (Ed.), *"Raw Data" is an Oxymoron* (pp. 15–40). Cambridge, MA: MIT Press.

Sanderson, R., & Van de Sompel, H. (2012). Cool URIs and dynamic data. *IEEE Internet Computing*, *16*(4), 76–79. http://doi.org/10.1109/MIC.2012.78

Schema.org. (2012). Describing Datasets with schema.org. Retrieved August 19, 2013, from http://blog.schema.org/2012/07/describing-datasets-with-schemaorg.html

Simons, N. (2012). Implementing DOIs for research data. *D-Lib Magazine*, *18*(5/6). http://doi.org/10.1045/may2012-simons

Suber, P. (2012). *Open Access*. Cambridge, MA: MIT Press.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE*, *8*(5), e64841. http://doi.org/10.1371/journal.pone.0064841

Thelwall, M., Vaughan, L., & Bjorneborn, L. (2005). Webometrics. In *Annual Review of Information Science and Technology* (Vol. 39, pp. 81–135).

Uhlir, P. F. (Ed.). (2012). *For Attribution—Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: National Academies Press.

Van de Sompel, H. (2012). The Web-Based Scholarly Record: Identification, Persistence, Actionability. *Libraries in the Digital Age (LIDA) Proceedings*, *12*(0). Retrieved from http://ozk.unizd.hr/proceedings/index.php/lida2012/article/view/60

Van de Sompel, H. (2013, April). *From the Version of Record to a Version of the Record*. Presented at the Coalition for Networked Information. Retrieved from http://www.youtube.com/watch?v=fhrGS-QbNVA&feature=youtube_gdata_player

Van de Sompel, H. (2015). hiberlink. Retrieved February 17, 2015, from http://hiberlink.org/

Van de Sompel, H., Hochstenbach, P., & Beit-Arie, O. (2000). OpenURL Syntax Description. Retrieved from http://www.exlibrisgroup.com/sfx_openurl_syntax.htm

Van de Sompel, H., & Lagoze, C. (2009). All Aboard: Toward a Machine-Friendly Scholarly Communication System. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery* (pp. 1–8).

Van de Sompel, H., Nelson, M. L., & Sanderson, R. (2013). HTTP framework for time-based access to resource states -- Memento draft-vandesompel-memento-09. Retrieved September 30, 2013, from https://datatracker.ietf.org/doc/draft-vandesompel-memento/

Van de Sompel, H., Sanderson, R., Klein, M., Nelson, M. L., Haslhofer, B., Warner, S., & Lagoze, C. (2012). A perspective on resource synchronization. *D-Lib Magazine, 18*(9/10). http://doi.org/10.1045/september2012-vandesompel

Wallis, J. C. (2012). *The Distribution of Data Management Responsibility within Scientific Research Groups* (PhD Dissertation). University of California, Los Angeles, Los Angeles, CA. Retrieved from http://escholarship.org/uc/item/46d896fm

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

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? data sharing and reuse in the long tail of science and technology. *PLoS ONE, 8*(7), e67332. http://doi.org/10.1371/journal.pone.0067332

White, H. D. (1982). Citation analysis of data file use. *Library Trends*, *30*(2), 467–478. Retrieved from https://www.ideals.illinois.edu/handle/2142/7222

Wilbanks, J. (2013). Licence restrictions: A fool's errand. *Nature, 495*(7442), 440–441. http://doi.org/10.1038/495440a

Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D. R., … Hudson, R. L. (2010). *Riding the wave: How Europe can gain from the rising tide of scientific data*. Final report of the High Level Expert Group on Scientific Data. Retrieved from http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

World Wide Web Consortium (W3C). (2013). WebSchemas/Datasets. Retrieved February 15, 2014, from http://www.w3.org/wiki/WebSchemas/Datasets

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A. E., & Traweek, S. (2012). Data, data use, and scientific inquiry: two case studies of data practices. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (pp. 19–22). New York: ACM. http://doi.org/10.1145/2232817.2232822

Zotero. (2015). Zotero Style Repository. Retrieved August 6, 2014, from https://www.zotero.org/styles