2012

Why are the attribution and citation of scientific data important? In: Uhlir, Paul and Cohen, Daniel (eds.). Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop.

Christine L Borgman, *University of California, Los Angeles*

Why Are the Attribution and Citation of Scientific Data Important?

Christine L. Borgman[1]
University of California at Los Angeles

Borgman, C.L. (2012, forthcoming). Why are the attribution and citation of scientific data important? In: Uhlir, Paul and Cohen, Daniel (eds.). *Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop*. National Academy of Sciences' Board on Research Data and Information. National Academies Press: Washington DC. http://www.nap.edu

Table of Contents

Introduction

My roles as Symposium Chair and as Keynoter are to frame the problems to be addressed in two days of discussion. This is a very sophisticated set of speakers and participants. Each of you has been concerned with research data, in some way, for some years. By now, all of us are familiar with the data deluge metaphor. We are being drowned in data, much of which is runoff. Valuable research data often are not captured, cited, or reused. Our challenge is to identify

[1] Presentation available at: http://sites.nationalacademies.org/PGA/brdi/PGA_064019

what part of these resources should be kept, the right way to keep them, and the right tools and services to make them useful.

Data have become a critical focus for scholarly communication, information management, and research policy. We cannot address the full array these issues, fascinating though they may be. Our two days will focus closely on questions of attribution and citation of scientific research data, although we frame *scientific* broadly enough to include most areas of scholarship.

We will devote little time to definitional issues (e.g., what are data), except to acknowledge that *data* often exist in the eyes of the beholder. Our principle concerns are how to assign credit for data (attribution) and how to reference data (citation) in ways that others can identify, discover, and retrieve them. Among the questions to be explored are what a community considers to be data, what data might be shared, what data should be shared, when data can be shared, and in what forms can data be shared. We will consider what approaches may be generic across disciplines and what practices may be field-specific.

Data citation and attribution are not new topics[2]. We have had standards for cataloging data files since the 1970s. Objects that can be cataloged also can be cited. Similarly, data archives have been promoting data citation practices for several decades. However, over this same period, very few journal editors required data citations, disciplines did not instill data citation as a fundamental practice of good research, granting agencies did not reward the data citations of applicants, tenure and reward committees did not recognize data citations in annual performance reviews, and researchers did not take responsibility for citing data sources. What have we learned from the past? What seems to be new today?

Several developments contribute to the renewed interest in data citation and attribution, all of which are topics of this Symposium. One is the growth in data volume relative to storage and analytic capacities. Fields such as astronomy, physics, and genomics are producing more data than investigators can investigate themselves. By sharing and combining data from multiple sources, other researchers can ask new questions. Another factor is advances in the technical infrastructure for generating, managing, analyzing, and distributing data. Tools are more sophisticated, bandwidth capacity is greater, and transfer speeds continue to improve. Third, and by no means least, are associated shifts in research policy. Data are now viewed as significant research products in themselves, more than just adjuncts to publications[3]. Funding agencies now expect investigators to capture, manage, and share their data. When viewed as research products, data deserve attribution similar to that of publications. Attribution, in turn, requires mechanisms for references to be made and citations to be received. Yet data are very different entities than publications. They take many more forms, both physical and digital, are far more malleable than publications, and practices vary immensely by individual, by research team, and by research area. Institutional practices to assure stewardship of data are far less mature than are practices to sustain access to publications. All of these factors contribute to the complexity of data citation and attribution. It is the many interacting dimensions of data

[2] Thanks to an anonymous reviewer for suggesting a fuller discussion of drivers for data citation and attribution than was included in the oral presentation at the Symposium. Some of the reviewer's comments are included in this text.

[3] 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://dx.doi.org/10.1002/asi.22634

attribution and citation that make it a problem worthy of this Symposium and of the multi-year effort with which the Symposium is associated.

Scholarly Infrastructure

Questions of data citation and attribution are best framed in terms of the infrastructure for digital objects. For our purposes, scholarly infrastructure is captured by the eight dimensions of infrastructure identified by Susan Leigh Star and Karen Ruhleder (1996)[4], as mapped in Figure 1, taken from Bowker, et al (2010)[5]:

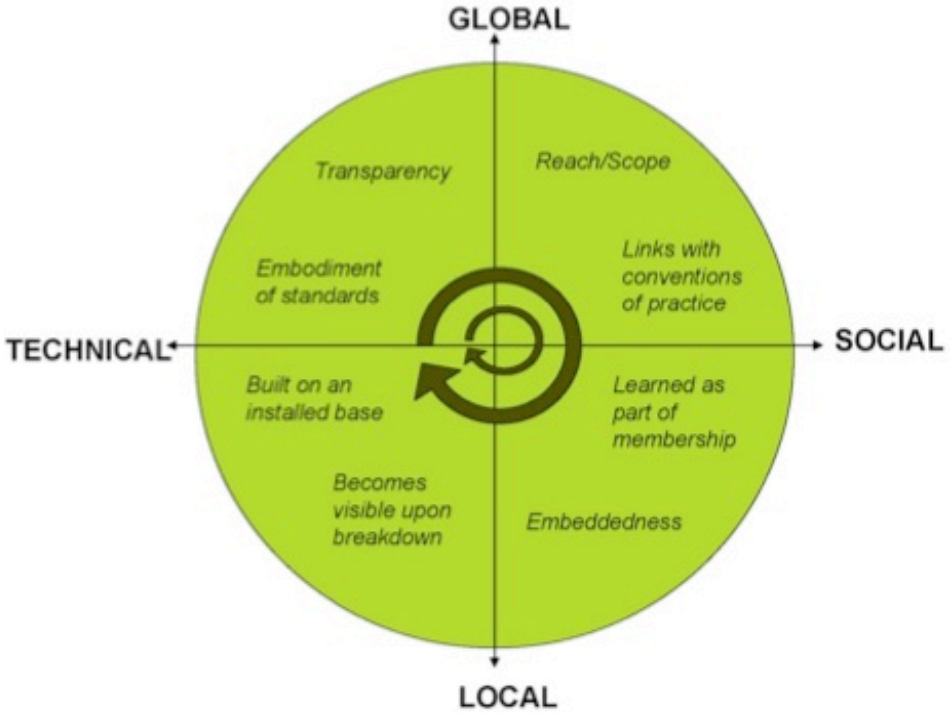

Fig. 1: Dimensions of Infrastructure

Our Symposium presentations will touch upon each part of this model. At the technical edge of the model, infrastructure is the embodiment of standards, which in turn are built on an installed base. Among the installed bases that influence data citation are Internet protocols, publishing practices, and library cataloging methods. At the social end, infrastructure is linked to

[4] Star, S. L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1): 111-134.

[5] Bowker, G. C., Baker, K., Millerand, F., Ribes, D., Hunsinger, J., Klastrup, L. & Allen, M. (2010). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In Hunsinger, J., Klastrup, L. & Allen, M. (Eds.). International Handbook of Internet Research. Dordrecht, Springer Netherlands: 97-117.

conventions of practice – whether cataloging or data management – and learned as part of membership in a community (e.g., librarians or astronomers). A social topic of particular interest is the relationship of reward systems to data citation. At the local edge of the model are individual practices for managing data and library practices for data stewardship. The global edge represents the inherently international character of scientific scholarship. Data practices, data exchange, and citation and attribution all must work effectively across political, institutional, and disciplinary boundaries.

Data in the global-technical quadrant of Figure 1 are most amenable to automated capture, management, and discovery. These are data, for example, from shared instruments such as space-based telescopes, and are associated with established data structures, analytical tools, and repositories. It is these types of data that are most readily cited. Conversely, data in the local-social quadrant tends to be more heterogeneous in form and content, more artisanal in data collection methods, and more varied in practices for management, use, and reuse. These data are much less amenable to established methods of data ingest, stewardship, citation, and attribution.

The infrastructure for digital objects has many features. We are concerned at this Symposium with how they apply to data attribution and citation, but we must remember that they are part of a larger Internet architecture of digital objects. The list of features below, around which the rest of my presentation is organized, is neither exhaustive nor mutually exclusive. Rather, it is a useful starting point to assess how these infrastructure features are applied to data citation and attribution:

- Social practice
- Usability
- Identity
- Persistence
- Discoverability
- Provenance
- Relationships
- Intellectual property
- Policy

Social Practice

Among the drivers for this Symposium are renewed interest in data citation due to increases in data volume, to advances in technical infrastructure, and to shifts in research policy associated with data. These developments still beg the questions of why data should be attributed and cited. Those questions have at least as many answers as there are persons attending this event. At the highest level, most of these answers can be grouped into categories of reproducing research, replicating findings, or more generally, to reuse data. To reuse data, it is necessary to determine what useful data exist, where, in what form, and how to get them. In turn, data must be described in some way if they are to be discoverable. For people to invest effort in making data discoverable, they should receive credit for creating, cleaning, analyzing, sharing, and otherwise making data available and useful. To get credit, some means must exist to associate names of individuals and organizations with specific units of data.

This Symposium is titled with the awkward phrase “Developing Data Attribution and Citation Practices and Standards” to make the point that these are not equivalent concepts. The

distinction is both subtle and important. *Attribution* is made to the responsible party. Attribution might thus be given to an individual investigator, to a research team, to a university, to a funding agency, to a data repository, to a library, or to another party responsible for gathering, assembling, curating, or otherwise contributing to the availability of data for others to use. Attribution is more closely associated with the notion of contribution, or contributor, than with author, which is among the differences between handling data and handling publications. Like publications, however, attribution implies social responsibility to give credit where credit is due. When we write journal articles and books, we reference other publications and the evidence on which they are based to attribute our sources.

Citation, in contrast, is the mechanism by which one makes references to other entities. In bibliometric parlance, references are made and citations are received. Even in the bibliographic world, reference/citation formats are many and varied: the *American Psychological Association* standard is popular in the social sciences, the *Modern Language Association* standard in the humanities, the *Association for Computing Machinery* in computer science and engineering, and the *Blue Book* in law, for example. These standards vary by the units they reference (e.g., full publications or individual pages), presentation (e.g., numerical references to the bibliography or author names and dates in text), choice of data elements (e.g., author, title, date, volume, issue, page numbers, DOI, URL, legal jurisdiction), and other factors. The multiplicity of bibliographic standards reflects the diversity of practices within and between research areas. None of them map easily to data or datasets, which have yet more diversity in form and practice.

Usability

Data citation and attribution must be considered in the context of the usability of data as digital objects. While data in the form of physical objects (e.g., samples, artifacts, lab notebooks on paper) also must be referenced, digital descriptions of those objects typically serve as surrogates. Among the actions people – or machines – may wish to perform on digital objects are to interpret, evaluate, open, read, compute upon, reuse, combine, describe, and annotate. This incomplete list suggests the range of capabilities that must be accommodated by a successful system for citing and attributing data.

Identity

To be citable and attributable, data must be identified uniquely. Identity and identification are well known problems in computer science and in epistemology. Our speakers on these topics bring those fields to bear on the question of identity for units of data. Identity is complex when we think in terms of people reading books and reading data. Humans can disambiguate similar objects, such as different editions of a book. Identity questions are even more complex when computers are discovering, reading, and interpreting data. Identity also is closely intertwined with usability and with trust. Among the questions to ask are: What are the dimensions of data identity? What identity levels are necessary to open, to interpret, to read, to compute upon, to combine, and to trust data as digital objects? An effective set of identity mechanisms for data citation and attribution must incorporate a trust fabric.

Persistence

The next session in this symposium is on identity and persistence of digital objects. Identity and persistence tend to be more concerned with containers of the data than about the data per se – how we package, name, and label data will influence the ability to identify them, to ensure they persist, or to dispose of them accurately. The data may exist, but unless we have labeled them

and stored them in a place to which others can return, their usability will be negatively affected. A variety of persistent identifier systems already exist, including Uniform Resource Identifiers (URI), Digital Objects Identifiers (DOIs) and other types of Handles, and other namespaces. While all are useful, none addresses all of the needs for data identity and persistence. Much remains to be learned about which systems are best, for which types of data, and for what purposes.

Discoverability

Discoverability is a broad topic, most researched in information retrieval. For the purposes of data citation and attribution, discoverability is the ability to determine the existence of a set of data objects with specified attributes or characteristics. The attributes of interest include the producer of the data, the date of production, the method of production, a description of an object's contents, and its representation. Discoverability may also include aspects such as levels of quality, certification, and validation by third parties. Discoverability depends both on the description and representation of data and on tools and services to search for data objects. Description and representation usually take the form of metadata, some of which may be automated if data are generated by instruments such as sensor networks or telescopes. Even for these types of data, metadata creation may require considerable human effort, making it an expensive process that is often avoided by researchers.

Human intervention is necessary to add metadata and description to most other kinds of data. As data move from one place to the next, those metadata may be augmented and incremented. Unlabeled bits are equivalent to books shorn of their covers and title pages. Data generally are discoverable via the metadata that describe them.

A variety of approaches to discovery are possible. Web search engines are one possibility, assuming that data descriptions are reachable via standard web protocols. With the introduction of semantic web technologies and associated search engines, location of datasets of interest based on semantic content becomes possible. Alternatively, more discipline-specific and structured catalogs can be created.

Data are discoverable only as long as someone keeps them, somewhere. Library and archival practice tends toward saving forever anything that is worth saving, although both professions also have long histories of weeding collections and of scheduling record disposal. Individual investigators are less likely, and less able, to maintain data permanently for discovery at some unknown later date.

Discoverability is thus associated with economics, a topic largely beyond the scope of this Symposium. Many research libraries and archives view data as important special collections, but also are concerned that data stewardship is an unfunded mandate. Data retention schedules will influence data discovery. Some data may be discoverable only in the short term, such as a scratch space for other people to use. Other data will be kept at least until the associated reports are published, and for some time thereafter. Yet other data will be "long-lived," usually defined long enough to be concerned about migration from one format to the next[6,7]. Data

[6] Reference Model for an Open Archival Information System (2002). Recommendation for Space Data System Standards: Consultative Committee for Space Data Systems Secretariat, Program Integration Division (Code M-3), National Aeronautics and Space Administration. http://public.ccsds.org/publications/archive/650x0b1.pdf

citation and attribution practices and standards may vary considerably depending on the period of time data are expected to remain available. Considerations also will vary between raw data, observations, models, physical samples, the predicted life span of utility, and many other factors.

Provenance

Provenance is particularly important for data citation. In citing data, it is important to reference the correct version, and where possible, to cite prior states of data and the transformations made between states. Provenance was once the exclusive concern of museum curators, archivists, and forensic specialists, all of whom view provenance as the chain of custody of an object. For example, The Getty Museum trusts the authenticity of an artwork only if the custody of that object can be documented at all steps since its origin. This linear model of provenance is less applicable to digital objects. In computing, provenance is the ability to track all transformations from the original state. Data provenance is becoming an active research area, and one to which we devote a session of the Symposium.

Relationships

While data can be discrete digital objects, they usually are related to other objects such as publications. Often multiple types of data have relationships to each other, providing context, calibration, and comparisons. Data citation and attribution mechanisms thus must facilitate linking of related objects and be able to refer to groups of objects, as well as to individual items. The choice of units for reference is a particularly contentious topic in data citation and attribution. When does citing a dataset associated with a journal article provide sufficient granularity? When is it necessary to cite each observation, each cell in a table, or each point on a graph? Identifying units, relationships between units, and types of relationships are all aspects of data citation and attribution.

Intellectual property

Intellectual property is a broad topic even if confined to scientific data. The Symposium session on intellectual property will focus on rights associated with data, such as the rights to use, reuse, combine, publish, and republish. Discovering data is but a first step. Once discovered and retrieved, users need to be able to identify what rights are associated with those data. For example, may we use the data for commercial purposes? May we share them with others? May we use them for teaching? Research teams, especially small teams, may not have documented ownership or rights associated with their data. Until data came to be viewed as valuable research products, ownership was an issue rarely discussed. Data often are not shared for the simple reason that it is not possible to determine who in a collaborative project has the rights to release them.

Open access, albeit an overused term with many meanings, has sensitized researchers to the value of making their research products available. From an intellectual property perspective, making a reference to data should be no different than a reference to a book or published paper. Including bibliographies in published works does not violate the copyright of the works cited. Rather, the bibliography is the form of attribution most central to scholarly practice.

---

[7] Long-Lived Digital Data Collections. (2005). National Science Board. http://www.nsf.gov/pubs/2005/nsb0540/

Policy

Both data citation and attribution have policy components. Many stakeholders are concerned with scholarly information policy, including funding agencies, publishers, data repositories, universities, investigators, and students. Each has policy concerns, thus we must ask what policy, what kinds of policy, and whose policy? Data management plan requirements and data sharing policies are case examples. Many funding agencies have established such policies, the specifics of which vary widely between the National Science Foundation and the National Institutes of Health in the U.S., the Wellcome Trust and the Economic and Social Research Council in the U.K., and others in the United Kingdom, the European Union, and Asia. These requirements may evolve to become more explicit about who is to receive what kinds of attribution for what kinds of data contributions, and how such contributions are to be cited.

Symposium Themes

All of these infrastructure issues, and more, will be explored in our program. The two days of the Symposium are organized around these driving questions from the project's task statement (emphasis added):

1. What are the **major technical issues** that need to be considered in developing and implementing scientific data citation standards and practices?
2. What are the **major scientific issues** that need to be considered in developing and implementing scientific data citation standards and practices? Which ones are **universal** for all types of research and which ones are **field- or context- specific**?
3. What are the **major institutional, financial, legal, and socio-cultural** issues that need to be considered in developing and implementing scientific data citation standards and practices? Which ones are universal for all types of research and which ones are field- or context-specific?
4. What is the **status of data attribution and citation practices** in individual fields in the natural and social (economic and political) sciences in the United States and internationally? Case Studies.
5. **Institutional Roles and Perspectives:** What are the respective roles and approaches of the main actors in the research enterprise and what are the similarities and differences in disciplines and countries? The roles of research funders, universities, data centers, libraries, scientific societies, and publishers will be explored.

Next Steps

This summary report of the Symposium, published by the National Academy of Sciences' Board on Research Data and Information, is the first formal product of the overall initiative on data attribution and citation. The CODATA- ICSTI Task Group on Data Citation Standards and Practices is conducting a survey, literature review, and gathering other materials for a white paper on best practices. Task Group members are giving presentations at international meetings over the next several years. Future efforts of this Task Group are expected to lead to standardization work. These efforts also will continue via Symposium participants' dissemination of the ideas generated here.

Acknowledgements

Many people devoted many months of effort to organizing this event. Paul Uhlir, Dan Cohen, and Cheryl Levey of the National Academies devoted much of their 2011 summer to the Symposium project. Paul, Dan, and I had conference calls (arranged by Cheryl) with each and every panel to ensure synthesis and continuity. The Symposium is part of the larger CODATA Task Group, whose co-chairs and members are Bonnie Carroll (Co-Chair), Jan Brase (Co-Chair), Sarah Callaghan (Co-Chair), Micah Altman, Elisabeth Arnaud, Christine Borgman, Dora Ann Lange Canhos, Todd Carpenter, Vishwas Chavan, Nathan Cunningham, Michael Diepenbroek, John Helly, Jianhui Li, Brian McMahon, Karen Morgenroth, Yasuhiro Murayama, Soren Roug, Helge Sagen, Eefke Smit, Martie van Deventer, John Wilbanks, and Koji Zettsu. Paul Uhlir, Dan Cohen, and Franciel Linares are staff consultants. Special thanks also are due to the Symposium Steering Committee: Christine Borgman (Chair), Allen Renear, Herbert van de Sompel, Gary King, Steven Jackson, David Kochalko, and John Wilbanks. The Symposium sponsors made this event possible: the Alfred P. Sloan Foundation, the Institute for Museum and Library Services, CODATA, Microsoft Research, and the National Science Foundation.