For Attribution:
Developing Data Attribution and Citation Practices and Standards

Board on Research Data and Information
Policy and Global Affairs Division
National Research Council
in collaboration with
CODATA-ICSTI Task Group on Data Citation Standards and Practices
and
The Data Cite Group

SUMMARY

The Board on Research Data and Information proposes to establish a steering committee that would organize a symposium and workshop on scientific data attribution and citation practices and standards pursuant to the following statement of task:

1. What is the status of data attribution and citation practices in the natural and social sciences in United States and internationally?
2. Why is the attribution and citation of scientific data important and for what types of data? Is there substantial variation among disciplines?
3. What are the major scientific, technical, 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 are some of the options for the successful development and implementation of scientific data citation practices and standards, both across the natural and social sciences and in major contexts of research?

BACKGROUND

The growth of electronic publishing of literature has created new challenges, such as the need for mechanisms for citing online references in ways that can assure discoverability and retrieval for many years into the future. The explosion of scientific datasets online presents related, yet more complex challenges. Data citation standards and good practices can form the basis for increased incentives, recognition, and rewards for scientific data activities that in many cases are currently lacking in most fields of research. The rapidly-expanding universe of online digital data holds the promise of allowing peer-examination and review of conclusions or analysis based on experimental or observational data, as well as the ability for subsequent users to make new and unforeseen uses and analyses of the same data – either in isolation, or in combination with other datasets.

This promise, however, depends upon the ability to reliably identify, locate, access, interpret and verify the version, integrity, and provenance of the digital datasets. The problem of citing online data is complicated by the lack established practices for referring to portions or subsets of data. As funding sources for scientific research have begun to require data management plans as part of their selection and approval processes, it is important that the necessary standards, incentives, and conventions to support data citation, preservation, and accessibility be put into place. There are already a number of initiatives in different organizations, countries, and disciplines already underway. An important set of technical and policy approaches have already been launched by the Internet Engineering Task Force (IETF), the U.S. National Information Standards Organization (NISO), and other standards bodies regarding persistent identifiers and online linking, including the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) and InfoURI. The World Data System is also focusing on these issues.

These, and a variety of other initiatives in the United States and elsewhere, however, have not been well coordinated. Recently, two groups have been formed to help promote a coordinated effort internationally. One is the CODATA-ICSTI Task Group on Data Citation Standards and Practices, The other, with is also represented in the CODATA Task Group, is the DataCite Group, which has been organized at the Technische Informations Bibliothek (TIB)—the German National Library of Science and Technology.

This proposed project would collaborate with these latter two focused umbrella groups, the CODATA-ICSTI Task Group and the Data Cite Group. The project would examine a number of key issues related to data identification, attribution, citation, and linking, which would promote coordination of activities in this area at the national level, and examine areas where further research is needed for common practices, standards, and technologies for use by the scientific community.

Issue Areas in the Development and Implementation of Scientific Data Attribution and Citation

There are many issues that would need to be addressed in establishing data attribution and citation standards and good practices. Below is a description of some of the topics that the project would address.

Technical

1. Interoperability and Facilitation of Re-use. There is already considerable diversity in database formats, such as various flat-file, hierarchical, relational, object-oriented, and XML-based databases. There is every reason to expect that new modalities and formats for storing and manipulating digital data will continue to emerge.

2. Citation Formats. What data citation conventions have been developed already? How are they similar and how do they differ? Can they be standardized and if so, how? It should be noted, however, that citation formats are not major considerations compared to the difficulty of determining the unit of data or the identity of that which is to be linked (Cole 2008).
3. Metadata. How do metadata conventions or standards affect attribution and citation of data?
4. Database Versioning. Datasets are more dynamic than documents, and this creates additional challenges for citation practice. When should the dataset as a whole be cited? How can a specific, time-fixed version be cited? What changes to the data constitute a new contribution or added value? How should this be acknowledged? How are database versions controlled and labelled?

A crucial dimension in this regard is provenance and how it is related o the need for attribution and citation. What attributions are needed given the complex provenance that is common for many types of data? How does one cite data that has been through many stages of transformation, some of them adding significant value and some trivial?

Scientific

The creators and users of online scientific datasets may have diverse needs that should be considered in the development, management, and use of scientific data in different discipline or research contexts. They also may have different needs regarding persistent identifier standards and models. For example, different disciplines may have disparate needs for granularity at which digital "objects" are identified. Some need geospatial metadata while others do not. What are the differences among disciplines that need to be addressed distinctly?

Institutional Roles

Successfully developing and implementing data citation practices and standards requires the participation by all major groups with the research community. What are the roles in this regard of the respective stakeholders in the system—the data managers, researcher umbrella groups, universities, libraries, publishers, and research funders? What are the implications for these stakeholders? Does this vary by major field of science or type of research?

Intellectual Property Rights and Licensing

Any registry system must accommodate traditional intellectual property rights (IPRs), such as those established through copyright, as well as emerging mechanisms of "some rights reserved", such as Creative Commons and Science Commons licensing.

Various important issues arise from data ownership, control, and IPRs. These are key drivers behind the different attitudes and practices toward data attribution and citation in different fields and countries, but relatively little work has focused on sorting out these issues. Principles and practices that have been tried in

different contexts need to be identified, and approaches that are more appropriate in the digital age should be explored. A recent OECD study (OECD 2008) addressed publicly funded data in this way, but both public and private data need to be considered. This is important because the willingness of individuals and institutions to accept and use different attribution, citation, and reuse frameworks will depend to a large extent on the real and perceived ownership, control, and IPRs associated with various databases.

Socio-cultural and Community Norms

A major reason for promoting the adoption of standard data citation practices is to develop a common basis and community of practice for recognizing and rewarding data work and incentivizing disclosure of data in interoperable and quality controlled ways. What are the factors that need to be considered in this area? Of particular interest is how such data management activities might impact the personal performance evaluations of scientists and the reward and promotion structures in science.

Attribution is not quite the same as citation, although citation is one of the ways of giving attribution. Licences akin to Creative Commons may require attribution, but this can result in "attribution stacking", where the work of hundreds or even thousands may need to be acknowledged. The route through this may be by establishing community norms for what are acceptable levels of attribution for datasets. Creative Commons and Science Commons recently added cut-and-paste citation support to their new version of the CC0 deed and to our norms documents (see http://labs.creativecommons.org/demos/pd/ and click through to the ones with metadata to see examples).

Persistent Digital Identifiers and Financial Sustainability

In a field that requires a lot of granularity in data use, even nominal registration fees per object can quickly become cost-prohibitive. In order for a data citation system to be useful, it must be accessible and its costs affordable by all necessary user communities.

It is important to consider data citations in the context of the semantic web. Online, the reference becomes "actionable"—the user wants to link directly to the item being cited. Distributed, linked technologies actually take us back to the original intention of citations, which is to enable the reader to discover, retrieve, and verify the identity of the referenced item. Bibliographic references presume that the desired object exists in multiple printed copies, and that any copy will do. In a digital world, only one "copy" exists. It is that copy that must be discoverable and retrievable. That sought item needs a persistent identifier. Normally, that persistent identifier is a URI.

The semantic web is predicated on linking of persistent identifiers. That model would be more inclusive and forward looking than the present framing. Within the semantic web framework, progress is being made on modelling relationships

between scholarly objects. The technical standards now in place are the Open Archives Initiative – Object Reuse and Exchange protocol (OAI-ORE) (Pepe, Mayernik, Borgman & Van de Sompel, 2010).

As noted above, there is a need for registration and persistent identification for online digital datasets. Some registry and resolution models for this function have already emerged, but the various models – for-profit vs. not-for-profit, public vs. private, etc. – must be examined to assure that they are sustainable in the long term. Moreover, just as the persistence of the connection from print citations to the correct physical copies depends on libraries or publishers keeping, the persistence of the connection between data citation and the actual data ultimately must also depend on some form of commitment by durable institutions to preserving data that is cited. Although a top down, centralized archive that keeps and organizes all data is an obviously attractive concept and works in some fields, creating such a trustworthy structure is probably not feasible universally, especially given the huge increases in the amount and types of data being generated or used by the scientific community. Distributed approaches to preservation such as institutional repositories, the Data Preservation Alliance for Social Science, and LOCKSS are emerging examples of alternatives to the centralized archiving model.

Other Issues

There are certain to be other important elements to the proper development and implementation of data attribution and citation standards and good practices, especially discipline-specific ones that may be identified by the steering committee.

PRELIMINARY WORK PLAN

A steering committee of about 6 experts will be appointed by the Chairman of the NRC having the following expertise: computer science and engineering, library and information science, database management, scientific publishing, science policy and research administration, and the sociology of science.

The steering committee will oversee all the project activities, including the planning and organization of the symposium and workshop. The project will be completed and the report published within 12 months.

Symposium/Workshop: A one-day symposium of invited expert speakers, organized in collaboration with the CODATA-ICSTI Task Group on Data Citation Standards and Practices, and with the Data Cite Group, will give presentations and discuss the issues pertaining to tasks 1-3 of the task statement. The symposium sessions will be moderated by the steering committee members. A workshop will be held on the second day with the invited speakers and the committee members. The purpose of the workshop will be to discuss the issues raised in the symposium in greater detail and to identify issues that were not

addressed, but still need to be studied. Both the symposium and workshop will be held in open session and the discussions of both will be recorded and subsequently transcribed. Members of the symposium audience will be asked to actively participate in and contribute to the discussions.

Following the meeting, a professional editor and the project staff will summarize the key points from the transcriptions. The report will be reviewed and edited in accordance with the established procedures of the Report Review Committee of the National Academies.

The steering committee members, staff, and members of the collaborating organizations will present the results of the project to various stakeholder groups in the year following the publication of the report.

Responsible National Research Council Staff

Paul F. Uhlir, Director, Board on Research Data and Information
National Academies, Keck-511
500 Fifth Street NW
Washington, DC 20001
E-mail: puhlir@nas.edu