

GOOD PRACTICES IN CITATION IN THE OUTPUTS OF
INTERNATIONAL STATISTICAL OFFICES

The Citation of Datasets

Report prepared by UNESCO

Background

Proper citation and bibliographic reference of material have taken place for many years and have a number of well-developed and effective styles. Each of these styles have been updated so that they contain the necessary guidelines to effectively cite many forms of traditional information sources as well as newer information sources, such as the internet.

There are many citation and bibliographic reference styles currently in widespread use. The most popular of these styles include:

- APA: generally used in psychology, education, and other social sciences;
- MLA: literature, arts, and humanities
- AMA: medicine, health, and biological sciences
- Turabian: designed for college students to use with all subjects
- Chicago: used with all subjects by books, magazines, newspapers, and other publications.
- Harvard: commonly used in UK academia and in the legal profession.

There has also been two ISO standards created in the area of bibliographic references:

- ISO 690: Information and documentation – Bibliographic references – Content, form, and structure;
- ISO 690-2: Information and documentation – Bibliographic references – Part 2: Electronic documents or parts thereof.

ISO 690-2 specifies the elements to be included in bibliographic references to electronic documents. It sets out a prescribed order for the elements of the reference and establishes conventions for the transcription and presentation of information derived from the source electronic document.

ISO 690-2 is intended for use by authors and editors in the compilation of references to electronic documents for inclusion in a bibliography, and in the formulation of citations within the text corresponding to the entries in that bibliography. It does not apply to full bibliographic descriptions as required by librarians, descriptive and analytic bibliographers, indexers, etc.

Even with all of this work in the area of citation, citation of datasets is still relatively unexplored. The complexity of dataset citation is increasing due to the ease in which data is redistributed and reused so that the original source may be a number of stages back. Data may be transformed accidentally or deliberately at any of these stages. Data may also be delivered embedded in software and will require metadata for informed understanding. Data may also be very dynamic or provided via a database environment, which could make it difficult in the future to reproduce the state of the data at the time that it was cited. In international organizations, how to credit the data sources within countries for the provision of their data is also an issue.

Reasons for Citation

There are many reasons why citation of data is as important as citation of other published sources of material. These reasons include:

1. It helps in the evaluation of the value of datasets to be able to track usage accurately. Including a feedback and tracking mechanism as part of a data citation policy is very useful in this regard.
2. The reliability of the information may be assessed on the basis of its provenance and the context and additional information provided in the source may permit the reader to go more deeply into the subject and to verify sources and authenticity;
3. The importance of giving appropriate credit to the producers of datasets. This is particularly the case in the increasingly competitive academic sector whereby credit needs to be attached to the production of high quality, well-documented datasets, but we are aware that it also happens in our own environment whereby different agencies re-use one another's data.
4. It can facilitate other researchers to locate the exact version of the data used so that they might re-analyze the data to amplify, extend, confirm or refute the author's interpretation of it (all of which is an important part of the scientific process);
5. It can facilitate other researchers to locate current versions of the same dataset or similar datasets from the same source.
6. It is important that, as the producers of the data, we should be able to locate quickly and accurately the exact version of the data we supplied so that we can answer queries quickly and can also resolve problems with the data.

These reasons underline the fact that an effective citation places an obligation on the data user to follow common citation best practices. To encourage effective citation, an obligation is also placed upon the data provider to provide the necessary information (metadata) in conjunction with the dataset to respect these citation best practices.

There is a more fundamental issue however. Even with all of the problems and challenges mentioned previously, it is possible for users to provide a basic citation for a dataset based on the guidelines that have been provided in the various citation styles referenced previously. This basic citation may not address all of the concerns that have been identified but that is another matter. The fact is, citation of datasets is not viewed in the same manner as citation of other materials. Unfortunately if one reads through publications which use data from our agencies one finds a huge array of practices with data sources often not cited, or cited incompletely (for example not specifying the agency concerned, or the access point within that agency, or the version of the dataset used). Of even more concern this occurs even within publications produced by our own agencies.

There are examples of two UN online databases that take datasets from many different data sources and try to effectively cite the source of all of the data. These two examples are the UN Common Database containing 477 data series from 52 different data sources and the UNEP Geo Data Portal containing 415 data series from 39 different data sources. Each dataset is essentially defined as a single variable or indicator, which has the original data provider as the data source. Even after the excellent efforts to identify the source organization and the source dataset, the contact information, release date, and dataset version, are not included in the citation information making the citation information only partially useful. Even with these citation shortcomings, they are two excellent examples of initiatives that are putting citation information into their databases and providing references back to the data providers on a very large scale.

If citation of datasets is to be taken seriously outside of our organizations, a concerted effort must be made within our organizations to:

- Place a data citation policy in an obvious place on our websites and provide the policy in conjunction with any electronic datasets. This citation policy could be as simple as the following:

Citation Information

In the event that data from the <organization name> are incorporated into your research or publication, please supply the following acknowledgment within your published work: "These data are distributed by the <organization name> <organization website>"

If possible, please e-mail or send us reprints/citations of papers or oral presentations founded on <organization name> data (see below for email and mailing address). This will help us to stay informed of how our data are being utilized.

There are no restrictions for use of data from the <organization name> unless otherwise expressly stated.

If you have any questions, please contact:
<contact information>

- Secondly, encourage a culture of data citation both inside our organizations as well as outside wherever our data is being used. This awareness can be raised by contacting all known users of our data, all editors of publications known to use our data, etc. and requesting that they follow the citation policy for the organization in future publications.

A simple but effective citation style for use within our organizations is to include:

- the unambiguous name of the dataset;
- the author of the dataset;
- the agency (or part of the agency) responsible for the dataset;
- the date of the dataset (or version number);
- the contact details for queries;
- the address of the archive or other place of storage or system for accessing the data;
- the publisher (if this is different from the author though for most of our agencies' publications the author and publisher are the same);
- if appropriate, the paragraph, table or page number.

This citation style should be followed for any data that is published internally or externally as well as for the documentation of any datasets that are created or modified.

Traditionally, a citation only cites the most recent use of a reference even though it may have passed through a number of different organizations since the responsible organization first created it. That is, hypothetically, if UIS data was provided to the World Bank who then provided it to another organization, the World Bank would cite the UIS and the other organization would cite the World Bank thus creating a chain of citations. The rationale is that by following the chain of citations, the original source of the reference and the responsible organization can be found. While this may not be the preferred approach for datasets, it is the most manageable approach. Adhering to this common citation practice would be the recommendation for dataset citations.

The challenge of effectively citing data sources in countries can be addressed by following this common citation practice of citing the most recent source. If the data is simply collected from countries by an international organization, not modified in any way, and put into the dataset, then the country should be cited as the source of the data. If the data is collected as part of a survey or statistical activity, which acts upon the data and subsequently generates a dataset, then the documentation of the survey activity should credit the data sources in the countries for providing all of the data. The dataset itself should reference the survey activity as the source of the data since it has gone through a lot more than simply a collection process. However, if a publication is produced by the same organization that has managed the data collection process, then thanking and providing credit to the countries for the original data would be appropriate.

If an organization takes a dataset and modifies it in some way before redistributing the dataset or in whole or as part of a publication, the citation for the modified dataset must indicate the source or the original dataset but the citation must also indicate that the data was modified from its original state.

Metadata

In earlier works that discussed the citation of datasets, the matter of codebooks and whether to cite the codebook or the dataset was raised as an issue. It was generally decided at that time that the citation should focus on the dataset, not the codebook. In today's environment, there is a lot of focus on metadata, or data about data and the metadata systems needed to manage it. Generally speaking, metadata can be described as information that describes the data.

There are many metadata initiatives underway. A small subset of these initiatives:

- the Data Documentation Initiative (DDI)
- the GDDS and SDDS data dissemination standards initiatives;
- METIS;
- ISO11179
- CBS Cristal Model
- plus many others

The metadata work that is taking place covers a broad spectrum but in all cases, the information needed to effectively cite a dataset, as per the citation style recommendation above, is included as part of the metadata work. As the implementation of metadata systems and standards become commonplace, the information needed for a more effective citation of datasets will be more readily available as well. Managing this metadata and making it available will involve changes in both systems and processes. The effective management of metadata however, will become an important component in determining the overall quality of datasets.

Data Management

The objectives numbered four (4) and six (6) in the "Reasons for Data Citation" stated above imply that:

- A historical copy is being maintained of datasets in an organization;
- The exact version of the dataset can be located based upon the information available to the user at the time they initially accessed the dataset.

In a highly dynamic environment where the data is constantly changing, these items become very complex. The introduction of databases also complicates the matter.

How to effectively cite the information in a highly dynamic environment will depend upon how the organization providing the data can recreate the environment at the time of data retrieval. The notion of recording both the date and time of dataset access, as part of the dataset citation may be exactly what is required. Ultimately however, the

recommendation of how to cite this dataset effectively will have to be provided by the organization that provides the dataset. Different technical implementations may require different information be included in the citation.

As well, datasets are copied and redistributed in many forms to meet the needs of the moment. For Internet dissemination, a dataset may be placed in an online database environment with interactive access to the data. The same dataset may be used as the foundation for analytical papers. The same dataset may also be placed in reference databases, cdroms, or publications. The end result is that there are multiple uses of the original dataset both internal and external to the organization that created the dataset.

The information (metadata) that accurately describes the datasets is important in order to identify that the source dataset in these instances is common and to manage data retention and archival activities.

Organizations need to have an effective data management and data archival policy that will keep a historical record of the datasets. The retention period for various datasets will be different depending upon the data and requirements.

Summary

In order to encourage effective citation on the part of data consumers, the data provider has a number of obligations that should be met.

- Provide a clear data citation policy and examples of proper citation in conjunction with the dataset irregardless of where the data is obtained (website, email, ftp site, ...). This citation policy and examples should be readily available to all data users.
- Provide the necessary descriptive information (metadata) along with datasets in order to meet the requirements of the organizations citation policy.
- Encourage a culture of data citation both internally and externally;
- Where possible, link the citation policy to the data lifecycle or archive policy for the organization in order to provide an indication of the long-term availability of the dataset.

Users of datasets should strive to respect best practices in effective citation as well as to respect the citation policy as set forth by the data provider organizations. This should be looked at from two perspectives.

- to ensure that proper credit is provided to the data provider;
- to ensure that there is no attempt to 'validate' data which was not provided as part of the original dataset.

Some useful references

We have been very grateful for assistance from Robert Johnston of the UNSD who has fought for good practice in this area for many years and for access to an internal note he has written on this topic.

Dodd, Sue A., “Bibliographic References for Computer Files in the Social Sciences: a Discussion Paper” IASSIST Quarterly 4(2) Summer 1990

Dodd, Sue A., “Bibliographic References for Numeric Social Science Data Files; Suggested Guidelines” Journal of the American Society of Information Science 30, 1979: pp 77- 82

Dodd, Sue A., “Cataloguing Machine-Readable Data Files: An Interpretive Manual” Chicago: American Library Association pp169 – 172

Ruus, Laine and Bombak Anna “Bibliographic Citations for Computer Files” Social Science Data Archive News, ANU, Canberra, Australia, Issue 31 March 1995

Sieber, Joan E., editor. “Sharing Social Science Data :advantages and challenges” Sage Publications, Newbury Park, California, 1991

See also <http://www.mun.ca/library/media/dlicit3.html>
and <http://www.census.gov/main/www/citation.html>
and <http://dpls.dacc.wisc.edu/cite.html>
and <http://www.nara.gov/publications/leaflets/gil17.html>