

Overview of a Proposed Standard for the Scholarly Citation of Quantitative Data

This extended abstract summarizes the proposed standard. For full details see:
Micah Altman, Gary King, 2006. "A Proposed Standard for the Scholarly Citation of
Quantitative Data",
[Working Paper]
<http://gking.harvard.edu/files/cite.pdf>

A critical component of the scholarly and library community is the common language of and the universal standards for scholarly citation, credit attribution, and the location and retrieval of articles and books. We present a proposal for a similar universal standard for citing quantitative data that retains the advantages of print citations, adds other components made possible by, and needed due to, the digital form and systematic nature of quantitative datasets, and is consistent with most existing subfield-specific approaches. Although the digital library field includes numerous creative ideas, we limit ourselves to only those elements that appear ready for easy practical use by scientists, journal editors, publishers, librarians, and archivists.

We propose that citations to numerical data include, at a minimum, six required components. The first three components are traditional, directly paralleling print documents. They include the author(s) of the data set, the date the data set was published or otherwise made public, and the data set title. These are meant to be formatted in the style of the article or book in which the citation appears. The author, date, and title are useful for quickly understanding the nature of the data being cited, and when searching for the data. However, these attributes alone do not unambiguously identify a particular data set, nor can they be used for reliable location, retrieval, or verification of the study. Thus, we add three components using modern technology, each of which is designed to persist even when the technology inevitably changes. They are also designed to take advantage of the digital form of quantitative data.

The fourth component is a unique global identifier, which is a short name or character string guaranteed to be unique among all such names, that permanently identifies the data set independent of its location. We allow for any naming scheme to be chosen, so long as it (1) unambiguously identifies the data set object, (2) is globally unique, and (3) is associated with a naming resolution service that takes the name as input and shows how to find one or more copies of the identical data set. Long-term persistence of the resolution service is meant to be guaranteed by the organization that operates it, although as is now becoming common redundant multiple naming resolution services can be set up so that archives can back each other up in case one goes out of business

Unique global identifiers thus guarantee persistence of the link from the citation to the object, but we also need to guarantee and independently verify that the object does not change in any meaningful way even when data storage formats change. Thus, we add as the next component a Universal Numeric Fingerprint or UNF. The UNF is a short, fixed-length string of numbers and characters that summarize all the content in the data set, such that a change in any part of the data would produce a completely different UNF. A UNF works by first translating the data into a canonical form with fixed degrees of numerical precision and then applies a cryptographic hash function to produce the short string. The advantage of canonicalization is that UNFs (but not raw

hash functions) are format-independent: they keep the same value even if the data set is moved between software programs, file storage systems, compression schemes, operating systems, or hardware platforms. Finally, since most web browsers do not currently recognize global unique identifiers directly (i.e., without typing them into a web form), we add as a final component of the citation standard a bridge service, which is designed to make this task easier in the medium term. Given how web services are accessed presently, the bridge service should be a URL, which can thus be recognized by any browser.

We also offer a systematic way to add information to data citations that also retains complete flexibility in added content. For each added element, we recommend a three-part syntax composed of a field name that describes the content being added, preceded by the value of the content, and followed by an (optional) semicolon separator: “value [fieldname];” or for example “data set [Type];”. To encourage standardization, field names should come from the widely used Dublin Core Metadata Initiative. If others are needed, additional items may be drawn from other metadata schemes and vocabularies by adding the identifier for that scheme in parentheses within the bracketed field name, such as “Interuniversity Consortium for Political and Social Research [Distributor (DDI)]” or “Current Population Survey Supplements [Series (ISO 690-2)]”. In unusual cases, users could even easily add their own vocabulary if needed. This extended standard can be used to create citations similar to and compatible with some existing approaches, such as ISO 690-2 (see ISO, 1997) (although some aspects of these approaches may now be obsolete).

Together, the global unique identifier, UNF, and bridge service ensure permanence, verifiability, and accessibility even in the situations where the data are confidential, restricted, or proprietary; the sponsoring organization changes names, moves, or goes out of business; or new citation standards evolve. Together with the author, title, and date, which are easier for humans and search engines to understand, all elements of the proposed full citation for quantitative data should achieve what print citations do and, in addition to being somewhat less redundant, take advantage of the special features of digital data to make it considerably more functional.