COMMENTARY

D-Lib Magazine
March/April 2007

Volume 13 Number 3/4

ISSN 1082-9873

A Proposed Standard for the Scholarly Citation of Quantitative Data

Micah Altman
Associate Director, Harvard-MIT Data Center and
Senior Research Scientist, Institute for Quantitative Social Science
Center for Government and International Studies
1737 Cambridge Street
Harvard University
Cambridge MA 02138
<http://www.hmdc.harvard.edu/micah_altman/>
<micah_altman@harvard.edu>

Gary King
David Florence Professor of Government, Institute for Quantitative Social Science
1737 Cambridge Street
Harvard University
Cambridge MA 02138
<http://GKing.Harvard.Edu>
<King@Harvard.Edu>

Abstract

An essential aspect of science is a community of scholars cooperating and competing in the pursuit of common goals. A critical component of this 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 propose 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 data sets, 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.

1 Introduction

How much slower would scientific progress be if the near universal standards for scholarly citation of articles and books had never been developed? Suppose shortly after publication only some printed works could be reliably found by other scholars; or if researchers were only permitted to read an article if they first committed not to criticize it, or were required to coauthor with the original author any work that built on the original. How many discoveries would never have been made if the titles of books and articles in libraries changed unpredictably, with no link back to the old title; if printed works existed in different libraries under different titles; if researchers routinely redistributed modified versions of other authors' works without changing the title or author listed; or if publishing new editions of books meant that earlier editions were destroyed? How much less would we know about the natural, physical, and social worlds if the references at the back of most articles and books were replaced with casual mentions, in varying, unpredictable, and incomplete formats, of only a few of the works relied on?

Fortunately, these questions about written materials are purely counterfactual, and the influence of the simple idea of scholarly citation of printed works on scientific progress has been extraordinary. Indeed, since science is not merely about behaving scientifically, but also requires a community of scholars competing and cooperating to pursue common goals, scholarly citation of printed matter can be viewed as an instantiation of a central feature of the whole enterprise.

Unfortunately, no such universal standards exist for citing quantitative data, and so all the problems listed above exist now. Practices vary from field to field, archive to archive, and often from article to article.

The data cited may no longer exist, may not be available publicly, or may have never been held by anyone but the investigator. Data listed as available from the author are unlikely to be available for long and will not be available after the author retires or dies. Sometimes URLs are given, but they often do not persist. In recent years, a major archive renumbered all its acquisitions, rendering all citations to data it held invalid; identical data was distributed in different archives with different identifiers; data sets have been expanded or corrected and the old data, on which prior literature is based, was destroyed or renumbered and so is inaccessible; and modified versions of data are routinely distributed under the same name, without any standard for versioning. Copyeditors have no fixed rules, and often no rules whatsoever. Data are sometimes listed in the bibliography, sometimes in the text, sometimes not at all, and rarely with enough information to guarantee future access to the identical data set. Replicating published tables and figures even without having to rerun the original experiment, is often difficult or impossible (see Dewald et al. [1], Fienberg et al. [2], King [3], King [4], King [5], Altman and McDonald [6]).

In this article, we propose a standard for citing quantitative data, one that goes beyond the technologies available for printed matter and

responds to issues of confidentiality, verification, authentication, access, technology changes, existing subfield-specific practices, and possible future extensions, among others.

2 Quantitative Data

Although our citation standard puts no special restrictions on what constitutes a quantitative data set, a definition may be useful: A quantitative data set represents a systematic compilation of measurements intended to be machine readable. The measurements may be the result of scientific research or information produced by governments or others for any purpose, so long as it is systematically organized and described.

To fix ideas we note that many data sets include one or more rectangular tables of numbers or characters that systematically record information about research subjects. The rows refer to the units (such as survey respondents, countries, years, planets, metabolites, animals, test questions, or genes), and the columns represent variables coding attributes of these units (such as age, size, vote for president, percent correct, or numbers of legs, etc.). Cell entries are usually numbers but are sometimes alphanumeric. Data sets can include only a few rows or columns or may require terabytes of storage. Other data sets can be thought of as a (relational, non-relational, hierarchical, network, object, or other) data base, and may be stored in almost any digital format.

A data set must be accompanied by "metadata," which describes the information contained in the data set such as the meaning of the rows and columns, details of data formatting and coding, how the data were collected and obtained, associated publications, and other research information. Metadata formats range from a text "readme" file, to elaborate written documentation, to systematic computer-readable definitions based on common standards.

3 A Minimal Citation Standard

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 changes: a unique global identifier, a universal numeric

fingerprint, and a bridge service. They are also designed to take advantage of the digital form of quantitative data.

The unique global identifier 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.

Some examples of unique global identifiers include the Life-Science Identifier (LSID, see Clark et al. [7], lsid [8]), designed to identify biological entities; the Digital Object Identifier (DOI® namespace, see Paskin [9], DOI [10]), commonly used to identify commercial print publications in the CrossRef application [11]; and the Uniform Resource Name (URN), which is in practice more of a common syntax for identifier schemes. All are used to name data sets in some places, and under specific sets of rules and practices. For example, the International DOI Foundation's appointed Registration Agencies implement different business models using the DOI® System: CrossRef charges for each DOI created to name text documents, whereas the German National Library for Science and Technology registers DOIs for data sets for free but requires that all data registered be distributed without any charge or other restriction. Similarly, LSIDs are normally used to name entities with life science content.

For areas that do not already have their own established unique identifier schemes, we recommend LSIDs, DOIs, or other existing identifiers, if their rules and features fit the desired use. Otherwise, we suggest the widely used and openly documented Handle System® (see [12, 13]), which has a great deal of infrastructure in place and low barriers to adoption. In some very general sense, handles, DOIs, LSIDs, URNs, and other identifiers are competitors, but all are organized by public spirited standards-based organizations and are highly interoperable (e.g., DOIs are based on the handle protocol, share much of the Handle System® technology, and implement additional services; they can incorporate LSIDs; LSIDs follow URN syntax), and so the choice to have some persistent, globally unique identifier is considerably more important than the particular option chosen. The differences among these may be important for an archive or field but will usually be immaterial for a practicing scientist.

To fix ideas, consider this example of a handle: `hdl:1902.4/00754`, for which `hdl:` identifies the rest of the string as a handle, `1902.4` is the handle prefix that identifies the owner responsible for the persistence of the identifier and its connection to the associated content (followed by a

slash as a separator) and 00754 is the unique local data set name. Any data publisher, author, library, or other entity may register as a Resolution Services Provider (RSP), so that they may be assigned a unique naming authority that they can then use to assign unique global identifiers to data sets. All unique global identifiers are designed to persist (and remain unique) even if the particular RSP that created it goes out of business (transferring control of its data objects and handles to another organization) or changes names or location. Including such an identifier provides enough information to identify unambiguously and locate a data set, and to provide many value-added services, such as on-line statistical analyses, or forward citation to printed works that cite the data set, for any automated systems that are aware of the naming scheme chosen. Uniqueness is also guaranteed across naming schemes, since they each begin with a different identifying string.

We recommend that the unique global identifier resolve to a page containing the descriptive and structural metadata describing the data set, presented in human readable form to web browsers, instead of the data set itself. This metadata description page should include a link to the actual data set, as well as a textual description of the data set, the full citation in the format we describe below, complete documentation, and any other pertinent information.[1]

The advantage of this general approach is that identifiers in citations can always be resolved, even if the data are proprietary, require licensing agreements to be signed prior to access, are confidential, demand security clearance, are under temporary embargo until the authors execute their right of first publication, or for other reasons. Metadata description pages like these also make it easier for search engines to find the data. The metadata can follow emerging standards, or any other scheme.

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 fifth 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. (See [14], for the description of UNF properties and the original algorithm. Also see [15], for working UNF software and current algorithmic details.)

Finding an altered version of a data set that produces the same UNF as the original data is theoretically possible given enough time and

computing power, but the time necessary is so vast and the task so difficult that for good hash functions no examples (known as "collisions") have ever been found. Moreover, even in the unlikely event that they are eventually found, only a small subset will produce files that make any sense as data sets (e.g., some would have characters in numerical fields or more than two codes for gender, etc.) and so could be easily detected. This property, known as "second preimage resistance" in the cryptography literature, means that inadvertently altering the data and not knowing about it is almost impossible, and even doing so intentionally is no easier. The metadata page to which the global unique identifier resolves should include a UNF calculated from the data, even if the data are highly confidential, available only to those with proper security clearance, or proprietary. The one-way cryptographic properties of the UNF mean that it is impossible to learn about the data from its UNF and so UNFs can always be freely distributed.[2] Most importantly, this means that editors, copyeditors, or others at journals and book publishers can verify whether the actual data exists and is cited properly even if they are not permitted to see a copy. Moreover, even if they can see a copy, having the UNF as a short summary that verifies the existence, and validates the identity, of an entire data set is far more convenient than having to study the entire original data set.

An example of a UNF is `UNF:3:ZNQRI14053UZq389x0Bffg?==`, where `UNF:` identifies the rest of the string as a UNF, `:3` means that the fingerprint uses version 3 of the UNF and hash algorithm, and everything after the next colon is the actual fingerprint. For a particular algorithm and number of significant digits, the fingerprint is always the same length. Thus, the UNF includes enough self-identifying information so that the algorithm used may be updated to newer versions over time without disturbing old citations.

When a citation refers to a collection with several component data sets, we recommend that a UNF be calculated for each, all the UNFs be included on the metadata description page, and the formal citation include just one UNF that combines all the separate UNFs (in accordance with the UNF algorithm specification, by reapplying the UNF algorithm to the set of UNFs in Posix sort order). See also Section 6.

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 the sixth and 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 recommend that it have a domain name run by (and acknowledging) the organizational guarantor, followed by the unique global identifier translated into standard format. If the HTTP protocol in URLs is replaced someday, this component of the citation can be updated or dropped (even in new citations to the same material), but the global identifier should remain unchanged indefinitely. All major unique global identifier schemes have one or more of such bridge services. Some

implementations of this bridge service URL are examples of or follow the syntax of "Persistent URLs" (PURLs), see [16]. DOI name bridge services are implemented through their dx.doi.org service. An example of a bridge service for a handle identifier is: `http://id.thedata.org/hdl\%3A1902.4\%2F00754`, where `http://id.thedata.org` is a resolver service, in this case the Dataverse Network project at Harvard University (see King [17] and its predecessor, the Virtual Data Center project, Altman et al. [18]), and everything following the last slash is the translated handle. In citations to appear in printed matter, the bridge service URL would appear in full; when the citation is to be used on-line, it could optionally be used only to provide a hyperlink for the identifier, so that the user would not see the URL in the link directly.

An example of a complete citation, using this minimal version of the proposed standards, is as follows:

> Micah Altman; Karin MacDonald; Michael P. McDonald, 2005,
> "Computer Use in Redistricting",
> hdl:1902.1/AMXGCNKCLU UNF:3:J0PkMygLPflyT1E/8xO/EA==
> http://id.thedata.org/hdl%3A1902.1%2FAMXGCNKCLU

where we format the handle, UNF, and bridge service like current standards for URLs, such as breaking them without a dash to continue on the next line. We use a space to unambiguously separate the identifier, UNF, and bridge service elements. For display, we use a special typewriter font for these three items to clarify what we mean, but this is not necessary and can instead follow the style of the book or journal in which the citation appears. We recommend the given order for the citation components, but the components may be permuted (or added to existing citation practices) to suit different journal styles without loss of functionality.

4 Optional Citation Elements

The essential information provided by a citation is that which enables the connection between it and the cited object. The true minimum, therefore, must include just the persistent identifier. Other citation components are provided for the convenience of the reader or others. For example, *Science* magazine excludes titles of cited articles to save space, but most other publishers prefer to include the title so the reader can understand the subject of the cited article before deciding to retrieve it. In our proposed minimal quantitative data citation standard, any relevant additional information is available from the metadata description page, or from the data set itself. And even the author, date, and title information that we might prefer be in the proposed minimal citation standard, and any other relevant information, can be obtained from the associated metadata. Yet, authors, editors, publishers, data producers, archives, or others may still wish to add optional features to the citation, such as to give credit more visibly to specific organizations, or to provide advertising for aspects of the data set. They may also wish to choose

their own superset of our "minimal" standard in order to establish their own "required" citation rules, as a condition of using their data or publishing in their journal, for example. Adding this information in almost any way will not reduce the functionality of our basic citation elements. However, to enable these additional elements of the citation to be computer readable, and thus even more functional, we now offer a systematic way to add machine-readable information to data citations that also retains complete flexibility in added content.

For each added element, we recommend a two-part syntax composed of a field name that describes the content being added, preceded by the value of the field, and followed by an (optional) semicolon separator: "value [fieldname];". To encourage standardization we recommend that these terms be drawn from the DDI 2.1 specification elements for study and variable descriptions (see [19 and 20]), which are now widely used for rich cataloging of quantitative data in the social sciences.[3] For example, DDI elements can be used to identify the organizations in the chain of custody between the original authors and the researcher who used the data that were authorized to modify or document the data: "National Opinion Research Center [Producer]; Inter-university Consortium for Political and Social Research [Distributor]".

If descriptive elements needed are not in the DDI, additional items may be drawn from other metadata schemes and vocabularies, such as the widely used Dublin Core Metadata Initiative (see [21 and 22]) or the ISO 690-2 standard, by adding the identifier for that scheme in parentheses within the bracketed field name, such as "data set [Type (DC)]" and "Current Population Survey Supplements [Series (ISO 690-2)]".[4] In unusual cases, users could even easily add their own vocabulary if needed. (The six minimal elements of the proposed citation standard can also be classified under the Dublin Core, as Creator, Date, Title, Identifier, Identifier, Identifier, respectively, but these field names need not be specified in the citation.) Each added field name and scheme identifier serves to facilitate interpretation of the added elements and thus need not imply the existence of full metadata records in the other schemas.

An example of the use of the extended citation rules would be:

> Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," hdl:1902.4/00754 UNF:3:ZNQRI14053UZq389x0Bffg?== NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

where we have also suppressed the bridge service URL, and underlined the unique global identifier, to illustrate what the citation might look like on-line.

This extended standard can be used to create citations similar to and compatible with some existing approaches, such as ISO 690-2 [23], although some aspects of these approaches may now be obsolete. For example using "[Computer file]", "[magnetic tape]", or "[Link]" for a field

no longer distinguishes data sets from almost any other object, such as an article in a journal published only on the web. Similarly, the common practice of including the date a web site was "[Accessed]" provides little useful information for data sets.[5]

5 Institutional Commitment

The persistence of the connection from print citations to the correct physical copies depends on libraries keeping copies, or publishing concerns or sponsoring professional associations continuing both to exist and to provide information to the public. For example, a citation to a book from a major publisher is more likely to persist than one from a vanity publisher with no library sales.

Similarly, the persistence of the connection between data citation and the actual data ultimately must also depend on some form of institutional commitment. This means that, at least early on, readers, publishers, and archives will have to judge the degree of institutional commitment implied by a citation, just as with print citations. Obviously, if the citation is backed by a major archive, the Library of Congress, or a major university repository, there is less to worry about than there might otherwise be. Journal publishers may, in addition, wish to require that data be deposited in places backed by greater institutional commitment, such as established archives.

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. Even the Library of Congress, backed by the resources of the U.S. Government, cannot come close to keeping a copy of all printed matter. Moreover, even if the funds for such an organization could be amassed, a centralized solution would not address the political and institutional incentive problem of local archives needing to receive credit for their work and needing to retain some degree of organizational control over their intellectual property.

Fortunately, top-down archiving is not the only available solution. The LOCKSS project [24] has made great progress in creating a bottom up infrastructure for archiving, based on multiple copies held in libraries. Furthermore, hybrids of the top-down and bottom up approaches have started to emerge, where institutions have committed to partnerships with other archives to back each other up in the event that one fails. The Data Preservation Alliance for the Social Sciences (Data-PASS) [25] and CLOCKSS initiative [26] are institutional examples of this strategy. Since archives that receive credit for collecting and distributing data are more likely to be able to continue to do so, a hybrid solution has considerable benefits as well.

We also suspect that in the longer run, all three of these approaches will be used for archiving, and as data storage costs continue to drop, some

archives or organizations will develop projects to crawl the web, ingest data in usable and durable formats, and provide more centralized archives created in this fashion from the bottom up. It would certainly be a landmark opportunity for a major donor, company, government, or other organization to invest in the future of science. If we can establish standards now, useful for the decentralized web of archives and other data sources now in existence, this future possibility will be more likely.

6 Deep Citation

"Deep citation," or references to subsets of data sets, are analogous to page references in printed matter. Subsets, such as those used in a statistical analysis to generate a table or figure in a published work, are now often described verbally in printed publications, and sometimes also in computer programming code provided in replication data sets distributed along with some journal articles. Data may be subsetted by row (e.g., women between 18 and 24 who voted for Clinton in 1996), by column (e.g., using variables about support for the death penalty and education), or both. Subsets also often include additional processing, such as variable recodes or imputation of missing data.

Devising a simple standard for describing the chain of evidence from the data set to the subset would be highly valuable. The task of creating subsets is relatively easy and is done in a large variety of ways by researchers. However, describing the process in a simple enough way, tying it closely enough to the methods researchers use to create them, and convincing researchers to adopt these procedures and protocols will require considerably more research and development, as it may require changing the software tools and procedures used in empirical research (see [27] and the citations therein). We thus follow a simpler, less demanding, and more politically and institutionally feasible strategy that fits better into current research practices.

We suggest at a minimum that a citation be made to the entire data set as described above, and that scholars provide an explanation for how each subset is created in the text (as is current practice), and refer to a subset by reference to the full data set citation with the addition of a UNF for the subset (i.e., just as occurs now for page numbers in citations to printed matter). For example, if the citation above to the entire data set were in the references, we would describe the subset in the text for a particular analysis, figure, or table, and then write: see Verba (1998, subset UNF:3:1OxR51b05uUYq4V9p0P9f1+==). When the main citation refers to a collection of data sets, and as per our recommendation includes a UNF for each, referencing will be even more straightforward. We suggest that, when feasible, citations to subsets of data include a variable list. The extended syntax introduced in Section 4 can be used to accomplish this using DDI syntax to list the data set's variables. For example, Age,Sex,V4[VarGrp/@var]; where the field name in square brackets indicates that the variable names listed (Age, Sex, and V4) form a variable group (VarGrp) with variable names (@var) specified.

In a sense, the numerical results printed in published tables or figures represent a fingerprint that summarizes a data subset. However, as most who have tried to replicate the results of published research learn, this fingerprint is often insufficient for understanding what was actually done. In part this is because it reflects both the recoding and subsetting process as well as the statistical analysis performed on the subset. What the subset-UNF provides is a verification of the data subset, separate from the statistical analysis. This development thus enables researchers to devote less time in replicating ordinary subsetting processes that should be clear in textual descriptions of the research procedures but often is not as clear as they might be.

Huge data sets sometimes come with more specific methods of referencing data subsets, and can easily be added as optional elements. Any ambiguity in what constitutes a definable "data set," which may be an issue in very large collections, is determined by the author who creates the global unique identifier, UNF, and bridge service URL. If the subset includes substantial value-added information, such as imputation of missing data or corrections for data errors, then it will often be more convenient to store and cite the subset as a new data set, with documentation that explains how it was created.

7 Versioning

We recommend versions of the same data set be given new identifiers and treated as separate data sets, with links back to the prior version kept in the metadata describing that data set. Forward links to new versions from the original are easily accomplished via a metadata search on the unique global identifier.[6]

New versions of very large data sets (relative to available storage capacity) can be kept by creating a new object that contains only differences from the original, and describing how to combine the differences with the original on the object's metadata description page. Version changes should be reflected by a change in the date, and may also be noted in the title, or by using the extended citation elements.

8 Concluding Remarks

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. The proposed standard is flexible enough to accommodate some deep citation references, as well as any amount of

additional information of interest to archives, producers, distributors, publishers, or others, without losing functionality. This citation scheme enables forward referencing from the data set to subsequent citations or versions (through the persistent identifier) and even a direct search for all citations to any data set (by searching for the UNF and appropriate version number).

Archives connected to the Dataverse Network can automatically produce all elements of a complete citation for any data set. Authors may also go to any of the Dataverse Network sites to create elements of a data set citation for themselves, to use on-line tools, or to obtain open source downloadable software, or calculate UNFs (see [28]). Of course, the standards we offer herein can also be produced by other software systems and are not dependent on any specific choices of software, archive, data producer, publisher, or author.

Acknowledgements

Our thanks to Caroline Arms, Dale Flecker, Ann Green, Dave Kane, Gerome Miklau, Norman Paskin, Jeri Schneider, Karen Sullivan, Paul Uhlir, and Mary Vardigan for helpful comments; and the Library of Congress (PA#NDP03-1), the National Science Foundation (SES-0318275, IIS-9874747) and the National Institutes of Aging (P01 AG17625-01) for research support.

Notes

1. While a citation enables one to find the data of interest, the article itself may not contain enough information to make use of that data. Thus the need for complete documentation, sufficient for someone trained in the relevant discipline, but unfamiliar with the dataset, to understand and interpret the data itself.

2. In extremely sensitive cases, not publicly revealing the number of variables in the data set, or adding an extra randomly generated one, would eliminate even extremely far out possibilities of disclosure risk.

3. The DDI sections for file and document description do not describe the fundamental logical content of the data itself, so we avoid these elements.

4. For readability, we prefer to use the human readable names of the metadata elements where unambiguous. Where necessary, XPath 1.0 syntax (see [29]) can be used to precisely specify an element in a particular schema.

5. For example, ISO 690-2 requires the inclusion of two such elements. Our proposed citation standard can produce ISO 690-2 compliant citations that also have the advantage of being unambiguously machine interpretable by using ISO prescribed ordering and elements, explicitly labeling the element (date) that does not confirm to our proposed default

ordering, and placing the persistent identifier, UNF, and bridge service URL at the end of the citation.

6. Since persistent identifier systems do not support version semantics internally, the only practical alternative to assigning a new identifier for most datasets is to reuse the existing identifier to point to a new version of the data. This alternative would violate replicability, which would be signaled by a failure of the published UNF to match the UNF of the available data. For the special case of time-series databases that are subject to continuous incremental public updates, it may be practical to assign a single unchanging title and identifier to the data set, and to have the date of the citation reflect the last update of the database at time of citation. However, if the database does not support retrieval of the state of its contents given a particular citation date, replicability requires that each snapshot cited be treated and made available as a separate data set.

References

[1] William G. Dewald, Jerry G. Thursby, and Richard G. Anderson. "Replication in empirical economics: The journal of money, credit and banking project". *American Economic Review*, 76(4):587-603, September 1986.

[2] Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf. *Sharing Research Data*. National Academy Press, 1985.

[3] Gary King. "Replication, replication." *PS: Political Science and Politics*, 28(3):443-499, September 1995. <http://gking.harvard.edu/files/abs/replication-abs.shtml>.

[4] Gary King. "The future of replication." *International Studies Perspectives*, 4(1):443-499, February 2003. <http://gking.harvard.edu/files/abs/replvdc-abs.shtml>.

[5] Gary King. "Publication, Publication". *PS: Political Science and Politics*, 39(01) 119-125, 2006. <http://gking.harvard.edu/files/abs/paperspub-abs.shtml>.

[6] Micah Altman and Michael P. McDonald. "Replication with attention to numerical Accuracy". *Political Analysis*, 11(3):302-307, 2003.

[7] Tim Clark, Sean Martin, and Ted Liefeld. "Globally distributed object identification for biological knowledgebases". *Briefings in Bioinformatics*, 5(1):59-71, March 2004.

[8] Lsid project website. Web Site. URL <http://lsid.sourceforge.net>.

[9] Norman Paskin. "Digital object identifiers for scientific data." Data Science Journal, 28:12-20, April 2005. <http://www.doi.org/topics/050428CODATAarticleDSJ.pdf>.

[10] International DOI Foundation organizational website. Web Site. URL <http://www.doi.org>.

[11] CrossRef project website. Web Site. URL <http://www.crossref.org/>.

[12] Handle System® website. Web Site. URL <http://www.handle.net>.

[13] S. Sun, S. Reilly, L. Lannom, and J. Petrone. *Handle System Protocol (ver 2.1) Specification*. RFC 3652 (Informational), 2003. <http://www.ietf.org/rfc/rfc3652.txt>.

[14] Micah Altman, Jeff Gill, and Michael P. McDonald. *Numerical Issues in Statistical Computing for the Social Scientist*. John Wiley and Sons, New York, 2003.

[15] UNF software and documentation. Web Site. URL <http://purl.oclc.org/NET/UNF_PROJECT_WEBSITE>.

[16] Purl project website. Web Site. URL <http://purl.oclc.org>.

[17] Gary King. 2007. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," Unpublished Manuscript. URL <http://gking.harvard.edu/files/abs/dvn-abs.shtml>.

[18] Micah Altman, Leonid Andreev, Mark Diggory, Gary King, Daniel L. Kiskis, Elizabeth Kolster, M. Krot, and Sidney Verba. "A digital library for the dissemination and replication of quantitative social science research: The virtual data center." *Social Science Computer Review*, 19(4):458-470, Winter 2001. <http://gking.harvard.edu/files/abs/vdcwhitepaper-abs.shtml>.

[19] Grant Blank and Karsten B. Rasmussen. "The data documentation initiative: The value and significance of a worldwide standard." *Social Science Computer Review*, 22 (3):306-318, 2004.

[20] DDI project website. Web Site. URL <http://www.icpsr.org/DDI>.

[21] NISO. The dublin core metadata element set, 2001. <http://www.niso.org/standards/resources/Z39-85.pdf>.

[22] Dcmi type vocabulary. Web Site. URL <http://dublincore.org/documents/dcmi-type-vocabulary/>.

[23] ISO. The dublin core metadata element set, 1997. <http://www.collectionscanada.ca/iso/tc46sc9/standard/690-2e.htm>.

[24] LOCKSS project website. Web Site. URL <http://www.lockss.org/>.

[25] Data-PASS project website. Web Site. URL <http://www.icpsr.org/data-PASS/>.

[26] CLOCKSS project website. Web Site. URL <http://www.lockss.org/clockss/>.

[27] Gerome Miklau and Dan Suciu. "Managing integrity for data exchanged on the web." *Eighth International Workshop on the Web and Databases*, Baltimore, 16-17 June 2005. <http://webdb2005.uhasselt.be/papers/1-3.pdf>.

[28] Dataverse Network project website. Web Site. URL <http://thedata.org/>.

[29] Xpath 1.0 recomendation. Web Site. URL <http://www.w3.org/TR/1999/REC-xpath-19991116>.

Top | Contents
Search | Author Index | Title Index | Back Issues
Previous article | Conference Report
Home | E-mail the Editor

D-Lib Magazine Access Terms and Conditions

doi:10.1045/march2007-altman