Finding, Analyzing, Disseminating, and Preserving Quantitative Data

Gary King
Harvard University

Joint work with Micah Altman and Sidney Verba

Rate of scientific progress without print citations?

- You can read my article, if you don't criticize me
- You can read my book, if you make me a coauthor
- Titles of books and articles change unpredictably, with no link to the old title
- Libraries have different titles for the same books
- You can't find articles I cite
- Researchers make "corrections" to books; leave title and author the same
- References replaced with casual mentions of a few in unpredictable formats
- For articles and books, this is FICTION
- For quantitative data, this is FACT

Data Access is the Key to Science

- Science is not (only) about being scientific
- Scientific progress requires community: Competition and cooperation in the pursuit of common goals
- Without access to the same materials: no community exists
- The value of an article that can't be replicated: ?
- Scholarly articles are summaries, not the actual research results
- But: Data access is spotty by field
- Movement to require data access with publication
- Finding the data is still hard
- Hard for journal editors to verify
- If you find it, how do you know it's the same?
- Class replication projects: most published articles cannot be replicated

Data Access is also the Key to Democracy

- Statistics = state-istics
- The state tax authority: counting people, estimating wealth
- Reformers use data to get the goods on the state
- In modern democracy: the public needs a direct source of information
- (Partnership with U.S. Census Bureau I'll describe later)

What is Quantitative Data For?

- Ready reference: What is the percent of women 18-24 who voted for Clinton in Massachuetts?
- Replication: validation & extension of scientific results
- Secondary analysis: Using data for purposes not originally envisioned
- Dissemination and Preservation: important for science, often a requirement of grants and journals

Rules for Citing Printed Matter

Kim, Jae-On, Norman Nie, and Sidney Verba. 1977. "A Note on Factor Analyzing Dichotomous Variables: The Case of Political Participation," Political Methodology, Vol. 4: No. 2 (Spring): Pp. 39–62.

First author (last name first) Second author My coauthor! Year Article title Journal (no longer exists) Volume number Issue number Season Pages Special formatting codes Special indentation Citations: rule-based, precise, redundant

Lack of Rules for Citing Numeric Data

- No consistency in practice
- No fixed rules for copyeditors
- Sometimes in the list of references; sometimes a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
- Sometimes the author is noted
- Date of creation is sometimes given
- URLs often given, rarely persist
- Dates of access: protect the researcher, do not help find the data
- The data may not be available publicly
- The data may no longer exist
- The data may not have ever been held by anyone but the investigator

Lack of Rules for Preserving Data

- A major archive renumbered all its acquisitions
- The same data distributed by different archives have different identifiers
- Publishers sometimes withdraw data from some archives, but it remains in others. Study numbers rendered invalid or ambiguous.
- When a dataset is expanded, the old study number is sometimes "deaccessioned" and a new one assigned. (Data remains available, but citation is invalid.)
- Researchers sometimes distribute modified (or corrected) versions of data as in archives, using the same identifiers.
- Changes to datasets are made and existing identifier is "reused"; old data lost.
- When storage media changes, are the data the same?

A New Citation Standard for Numeric Data

Sidney Verba, 1998, "Political Participation Data", hdl:1902.4/00754, http://id.thedata.org/hdl%3A1902.4%2F00754, UNF:3:6:ZNQRI14053UZq389x0Bffg?==

1. Author
2. Year
3. Title
4. VDC Unique Global Identifier (handle)
5. Bridge Service (presently a URL)
6. Universal Numeric Fingerprint (UNF)

Data to Universal Numeric Fingerprints

$$\begin{pmatrix} 1 & 4 & 4 & 21 & \cdots & 121 \\ 1 & 2 & 2 & 91 & \cdots & 212 \\ 1 & 9 & 2 & 72 & \cdots & 104 \\ 0 & 2 & 2 & 2 & \cdots & 321 \\ 1 & 6 & 2 & 12 & \cdots & 204 \\ 1 & 9 & 4 & 52 & \cdots & 311 \\ 0 & 3 & 2 & 23 & \cdots & 92 \\ 0 & 2 & 5 & 91 & \cdots & 212 \\ 0 & 5 & 8 & 91 & \cdots & 91 \\ 1 & 9 & 1 & 72 & \cdots & 104 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 2 & 91 & \cdots & 212 \end{pmatrix} \Longrightarrow \text{ZNQRI14053UZq389x0Bffg?==}$$

Same UNF regardless of hardware, operating system, statistical software, database, or spreadsheet software.

The Data Center When We Came to Harvard

Give me my data!!!!

The Harvard-MIT Data Center Today

- The VDC has automated most previously uninteresting activities
- Its more fun to work here
- We're become a research organization (part of the Institute for Quantitative Social Science)

Who the VDC Serves

- used in production for data delivery to Harvard and MIT
- 1000s of users annually, from every Harvard school
- 10,000s of quantitative studies available through system
- Provides virtual access to local and remote data collections
- Disseminates Murray Research Archive collection
- Can now be installed at other sites at Harvard and around the world; most will federate

What the VDC Does: For the User

- Imagine sitting in your dorm room or office
- Do a structured search for data: locally, at other archives, and at other VDC sites
- Find data, see abstract, read documentation
- (Or with a existing citation, go straight to its meta-data)
- Authenticate yourself and get access authorization
- Run descriptive statistics and graphics
- Run cutting-edge statistical analyses (with replication code)
- Subset data (only men from Western countries)
- Translate to a convenient format
- Download subset
- Citation for subset provided

What the VDC Does: For Science

- Replication and Citation (creation and management of persistent identifiers for datasets, UNF generation, replication code generation for analyses)
- Sophisticated, Replicable On-line Analyses (Large array of statistical procedures available)
- Instant, Automated Inclusion of New Statistical Procedures (interface with R and Zelig)
- Preservation (preservation formatting, preservation metadata)
- Distribution and Federation (federated searching and browsing, distributed virtual collections, metadata harvesting, repository caching, and federated authentication and authorization)

What the VDC Does: For the Archive

- Study Preparation (ingest; conversion of data and documentation formats; catalog record creation)
- User Interfaces (data users, data producers, data archive administrators, data curators, librarians)
- Study Management (file-format independent storage, archival formatting, cataloging)
- Metadata Search and Harvesting (DC, MARC and DDI metadata import and export; OpenArchives and Z39.50 protocol gateways)
- Dissemination (download packaging, format conversion, subset selection and generation).
- Curator's Collections (share expertise, make collections virtual, cross-institution)

What the VDC Does: For Data Providers

- Include your study in a specific archive
- Include your collection in that archive
- Have your own branded collection on your web page, in your page's style, served by your archive, with full VDC services
- Have your own fully customized VDC Server

Partnership: VDC and U.S. Census Bureau's DataWeb

Data is the Intersection of Science and Democracy

- VDC: Scientific Research Data
  - Unifying access to scientific data
  - Easy access for academics
  - Allowing access to all official U.S. Data through Census
  - Statistical analysis through Zelig
- Census: Government Data
  - Unifying access to all official Governmental data
  - Easy access to the general public
  - Access to scientific data through the VDC
  - Statistical analysis through Zelig

Development Principles

- Web-based, light client for users, administrators, curators
- Built with off-the shelf components E.g.: Apache web server, OpenLDAP, R, Zelig, PostgresSQL Integration: Perl, Java Servlets, XSL/XML
- Open Source
  - Source code is included
  - You own the program; if you don't like what we do, you can go in a different direction, or add to the project
  - Modifiable & Redistributable
  - Does not restrict use of commercial data services
- Follows Open Source Standards Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- Completely distributed
  - Simple components-based architecture
  - Any component can be on any computer hardware
  - Distributed catalog: harvesting, distributed search
  - Distributed data: proxying, caching, replication
- Considerable Resources Marshalled

Next at the VDC

- First public version just released
- DATA-PASS Preservation and cataloging agreement, under Library of Congress auspices, among
  - ICPSR (U Michigan), Odum Institute (UNC), Roper Center (UConn), NARA, HMDC, Murray
- Integration with U.S. Census Bureau's DataWeb Project
- Integration with GenePattern at the Broad Institute
- Many other technical developments
- Interest from many universities and other organizations

For more information

http://GKing.Harvard.edu