The Dataverse Network: An Infrastructure for Data Sharing

Gary King
Harvard University

February 5, 2008

Papers

- Gary King, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing, *Sociological Methods and Research*, 32, 2 (November, 2007): 173–199.
- Micah Altman and Gary King. A Proposed Standard for the Scholarly Citation of Quantitative Data, *D-Lib Magazine*, 13, 3/4 (March/April).
- Dataverse Network project: `http://TheData.org`

Infrastructure for Quantitative Data

- Accessibility:
  - Most large data sets: in public archives
  - Most data in published articles: not accessible, results not replicable without the original author
  - Most data sets from NSF & NIH grants: not publicly available
- Problems even with professional archives:
  - Data in different archives have different identifiers
  - One major archive renumbered all its acquisitions
  - Changes to data are made; identifiers are reused or deaccessioned; old data are lost
- Data sets are not like books
  - Static data files (even if on the web): unreadable after a few years
  - When storage methods change: some data sets are lost; others have altered content!

What About a Centralized Data Access Solution?

- Highly desirable when feasible
- Works great in genomics, astronomy, etc., when data formats are universal, goals are common, and agreements are in place
- Impossible when data are heterogeneous in format, origin, size, effort needed to collect or analyze, IRB access rules, etc.
- Why don't researchers put data in public archives?
  - The Archive gets the credit
  - Upon questioning: they want credit, control, and visibility
  - (So why don't they worry about print publishers getting all the credit? Lack of data citations!)
- We propose: technological solutions to these political and social problems

Requirements for Effective Data Sharing Infrastructure

- Recognition, for authors, journals, etc. in (1) citations to data, (2) citations to associated articles, and (3) visibility on the web.
- Public Distribution, without permission from the author
- Authorization: fulfill requirements the author originally met
- Validation: check that data exists, without authorization
- Persistence Decades from now. . . .
- Verification: data remains unchanged, even if converted from SPSS to Stata to R, from a PC to a Mac to Linux, and from 8 inch magnetic tape to 5.25 inch floppies to a DVD.
- Ease of Use Neither editors nor authors employ professional archivists
- Legal Protection:
  - Journals have liability protection for print; none for data
  - *If you put data on the web without IRB approval, you are violating federal regulations*
  - (IRB approval must be for data distribution, not merely for the study)
  - Solution must not require lawyers (we've automated the IRB)

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 Third author Year Article title Journal (no longer exists) Volume number Issue number Season Pages Special formatting codes Special indentation Citations: rule-based, precise, redundant Print Citations Work: authors don't think publishers get all the credit; cited articles can be found; copyeditors don't need to see the original to know it exists; the link from citation to print persists

A New Citation Standard for Numeric Data

Sidney Verba, 1998, "Political Participation Data", hdl:1902.4/00754, UNF:3:6:ZNQRI14053UZq389x0Bffg?== Annals of Applied Statistics [Distributor]; NORC [Producer].

1. Author
2. Year
3. Title
4. Unique Global Identifier: will work after URLs stop working
5. Linked to a Bridge Service (presently a URL: `http://id.thedata.org/hdl%3A1902.4%2F00754`)
6. Universal Numeric Fingerprint (UNF)
7. Standard rules for adding citation elements

Data to Universal Numeric Fingerprints

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

Advantages of UNFs

- UNF is calculated from the content not the file: Its the Same UNF regardless of changes in computer hardware, storage medium, operating system, statistical software, database, or spreadsheet software.
- Cryptographic technology: any change in data content changes the UNF. (cannot tinker after the fact!)
- Noninvertible properties
  - UNFs convey no information about data content
  - OK to distribute for highly sensitive, confidential, or proprietary data
  - Copyeditor can validate data's existence even without authorization
- The citation refers to one specific data set that can't ever be altered, even if journal doesn't keep a copy
- Future researchers can quickly check that they have the same data as used by the author: merely recalculate the UNF

Web 2.0 Terminology

- Software: find CD, install locally, hit next, hit next, hit next...
- Web application software: no installation; load web browser and run (Dataverse Network Software)
- Host: The computers where the web application software runs (universities, archives, libraries)
- Virtual host: Where the web application software *seems* to run, but does not (web sites of: authors, journals, granting agencies, research centers, universities, scholarly organizations, etc.)

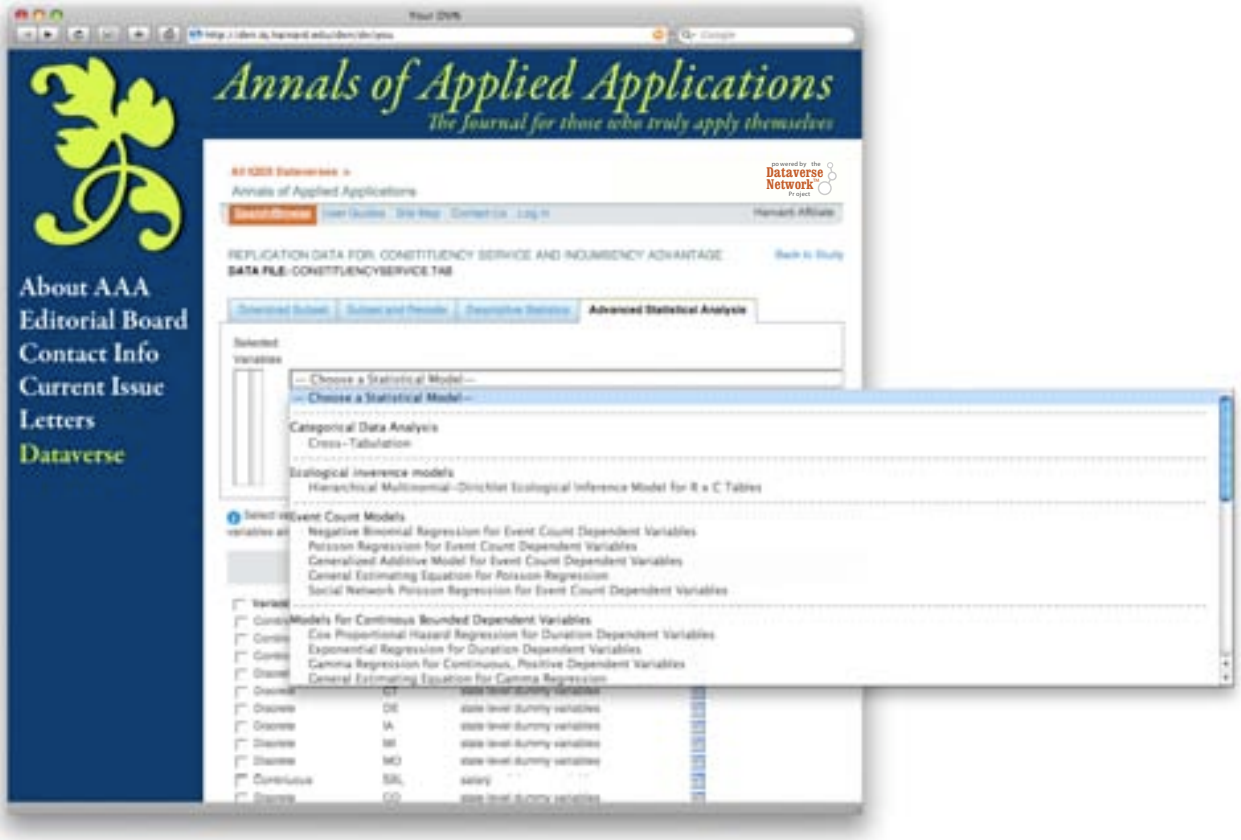
Your web site

Your dataverse branded as your web site but served by the Dataverse Network, therefore requiring no local installation and providing an enormous array of services


Annals of Applied Applications
The Journal for those who truly apply themselves
About AAA
Editorial Board
Contact Info
Current Issue
Letters
Dataverse
Annals of Applied Applications
REPLICATION DATA FOR: IMPROVING FORECASTS OF STATE FAILURE
Cataloging Information
Documentation, Data and Analysis
MRA

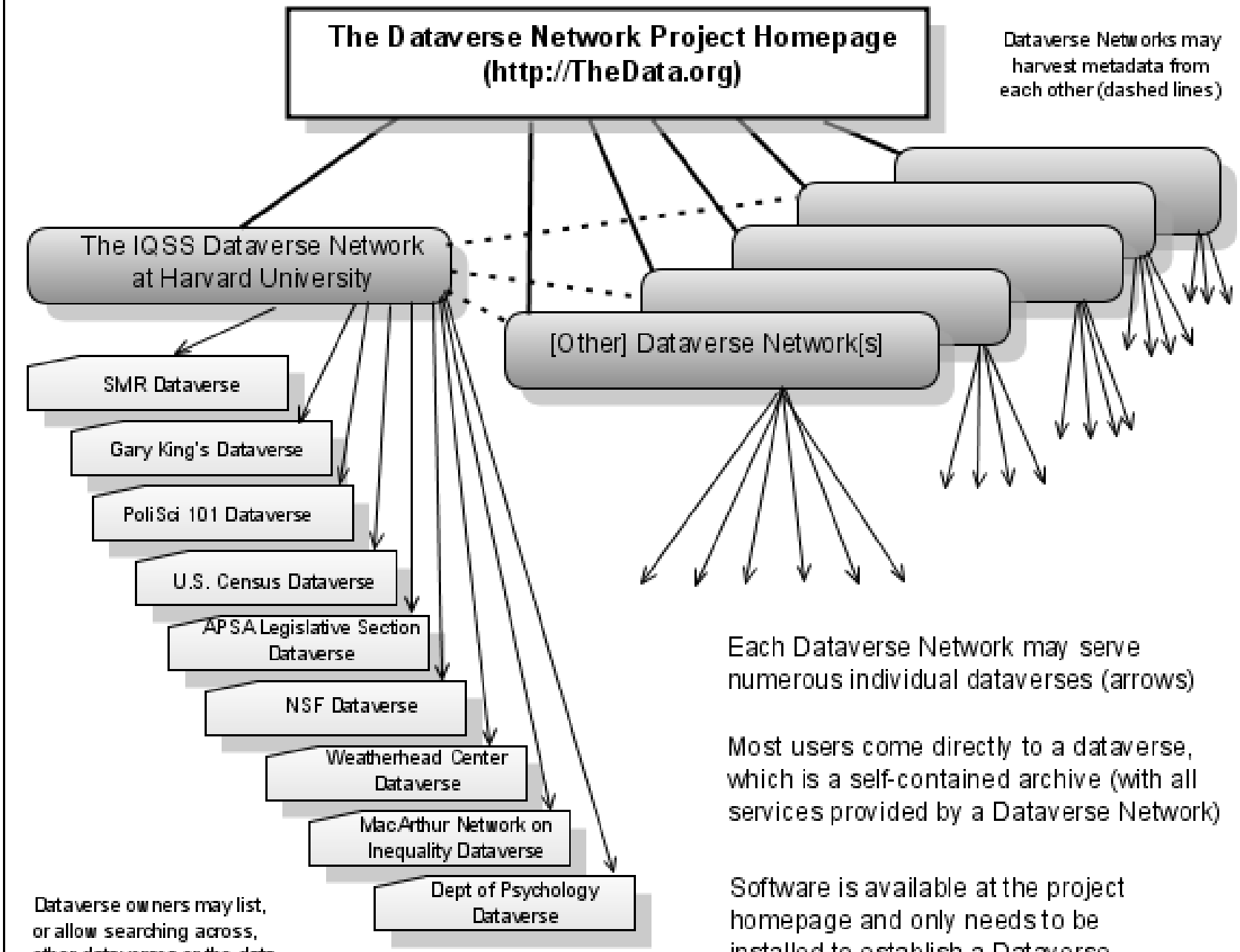

Annals of Applied Applications
The Journal for those who truly apply themselves
About AAA
Editorial Board
Contact Info
Current Issue
Letters
Dataverse
Annals of Applied Applications
REPLICATION DATA FOR: IMPROVING FORECASTS OF STATE FAILURE
Cataloging Information
Documentation, Data and Analysis


Annals of Applied Applications
The journal for those who truly apply themselves
About AAA
Editorial Board
Contact Info
Current Issue
Letters
Dataverse


Annals of Applied Applications
The Journal for those who truly apply themselves
About AAA
Editorial Board
Contact Info
Current Issue
Letters
Dataverse
Annals of Applied Applications
Advanced Statistical Analysis
-- Choose a Statistical Model --
Categorical Data Analysis
Cross-Tabulation
Ecological inference models
Event Count Models
Negative Binomial Regression for Event Count Dependent Variables
Poisson Regression for Event Count Dependent Variables
Generalized Additive Model for Event Count Dependent Variables
General Estimating Equation for Poisson Regression
Social Network Poisson Regression for Event Count Dependent Variables
Models for Continous Bounded Dependent Variables
Exponential Regression for Duration Dependent Variables
Gamma Regression for Continuous, Positive Dependent Variables

Your web site

Your dataverse branded as your web site but served by the Dataverse Network, therefore requiring no local installation and providing an enormous array of services


The Dataverse Network Project Homepage
(http://TheData.org)
Dataverse Networks may harvest metadata from each other (dashed lines)
The IQSS Dataverse Network at Harvard University
[Other] Dataverse Network[s]
SMR Dataverse
Gary King's Dataverse
PoliSci 101 Dataverse
U.S. Census Dataverse
APSA Legislative Section Dataverse
NSF Dataverse
Weatherhead Center Dataverse
MacArthur Network on Inequality Dataverse
Dept of Psychology Dataverse
Dataverse owners may list, or allow searching across, other dataverses or the data sets in them
Each Dataverse Network may serve numerous individual dataverses (arrows)
Most users come directly to a dataverse, which is a self-contained archive (with all services provided by a Dataverse Network)
Software is available at the project homepage and only needs to be installed to establish a Dataverse Network. Dataverses are virtual hosts.

A Journal Dataverse for a Replication Data Archive

- Full service virtual archive, with numerous data services (citation, metadata, archiving, subsetting, conversion, translation, analysis, . . .)
- Hierarchical list of data: year/volume/issue/article/dataset
- Branded as yours: with the look and feel of your site
- Easy to setup: give DVN your style, and include a link to your new dataverse
- Easy to manage: no software or hardware installation, backups, worry about archiving standards, or data format transations; still exists if you move; easy to rebrand
- Does not disrupt workflow: copyeditor ensures data is cited; author is given password to upload data. Everything else is self-service
- High acceptability: experiments indicate $> 90\%$ uptake for authors (without publication at stake)
- Reuse: data may appear on author's dataverse too
- Results: Journals with replication policies have three times the impact factor! (with dataverse, it should be more)

Dataverse Uses

- Journals, for replication data archives
- Authors, for their own data
- Future Researchers: browse or search for a dataverse or dataset; forward citation search; verification via UNFs; subsetting; read metdata, abstract, & documentation; check for new versions; translate format; statistical analyses; download
- Teachers, a list or for in depth analysis
- Sections of scholarly organizations, to organize existing data
- Granting agencies
- Research centers
- Major Research Projects
- Academic departments, universities, data centers, libraries
- Data archives
- Using DVN tools with outside data

The Universe of Data meets the Universe of Methods

- R Project for Statistical Computing
  - nearly 1000 packages; most new methods appear in R first
  - Highly diverse examples, syntax, documentation, and quality
  - Difficult for statistics-types; harder for applied researchers
- Zelig: Everyone's Statistical Software
  - An ontology we developed of almost all statistical methods
  - Users incorporate original packages via simple bridge functions via a simple model description language
  - Result: Unified Syntax, the same 3 commands to use any method
  - Automatically generated Graphical User Interface with all the world's methods
- R + Zelig + Dataverse Network
  - Greatly reduced time from methods development to use
  - Easy for applied researchers even if non-programmers
  - Can save R script for replication or further analysis

Licensing

- Web application software
  - Pros: easy to use, no installation costs
  - Cons: the software can vanish or change at any time
- Dataverse Network Software
  - Affero GPL License
  - Open source, public ownership
  - If you don't like the new version: you can make a new one
  - You own the software & the underlying code
  - The license guarantees that this will remain true in the future
- Licensing data
  - DVN automates the IRB process; no lawyers necessary
  - Data may be restricted in many ways, while metadata is available

For more information

http://TheData.org

Technology used in DVN Software

- Language: Java Enterprise Edition 5 (with EJB3 and JSF) (team picked for JavaOne; Sun engineers regularly call for advice)
- Application server: GlassFish (wrote press release on our project)
- Database: we use PostgreSQL (can substitute others)
- Statistical computing: R and Zelig