The Dataverse Network: An Infrastructure for Data Sharing

Gary King
Institute for Quantitative Social Science
Harvard University

(8/14/08 talk at “UseR! 2008”, Technische Universität, Dortmund, Germany)
- More information: [http://TheData.org](http://TheData.org)
Accessibility:
- Most large data sets: in public archives
- Most data in published articles: not accessible, results not replicable without the original author

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!

Connection to analysis software (like R)
- uncertain, time consuming, annoying, error prone
What About a Centralized Data Access Solution?

- Highly desirable when feasible
- Works great in 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 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
  - *In the U.S., 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)
<table>
<thead>
<tr>
<th>Author</th>
<th>Sidney Verba</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>1998</td>
</tr>
<tr>
<td>Title</td>
<td>“Political Participation Data”</td>
</tr>
<tr>
<td>Linked to a Bridge Service (presently a URL:</td>
<td><a href="http://id.thedata.org/hdl%3A1902.4%2F00754">http://id.thedata.org/hdl%3A1902.4%2F00754</a>)</td>
</tr>
<tr>
<td>Universal Numeric Fingerprint (UNF)</td>
<td></td>
</tr>
<tr>
<td>Standard rules for adding citation elements</td>
<td></td>
</tr>
</tbody>
</table>
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} \implies \text{ZNQRI14053UZq389xBffg?}==
Advantages of UNFs

- UNF is calculated from the content not the file: It's 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 dataverse branded as your web site but served by the Dataverse Network, therefore requiring no local installation and providing an enormous array of services.

http://www.peterson.com

http://dvn.iq.harvard.edu/peterson
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

- 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

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.
Your Dataverse

- Full service virtual archive, with numerous data services (citation, metadata, archiving, subsetting, conversion, translation, analysis, ...)
- List of your data, or your view of the universe of data
- 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
- High acceptability: experiments indicate > 90% uptake for authors
- Reuse: same data may appear on different dataverses
- Results: Articles with data available have twice the impact factor! (with dataverse, it should be more)
Dataverse Uses

- Authors, for their data or their view of the universe of data
- Journals, for replication data archives
- Future Researchers: browse or search for a dataverse or dataset; forward citation search; verification via UNFs; subsetting; read metadata, 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
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
  - Can be difficult for us; harder for applied researchers

- **Zelig: Everyone’s Statistical Software**
  - An ontology we developed of almost all statistical methods
  - Users incorporate original packages a simple model description language (and R bridge functions)
  - Result: Unified Syntax, the same 3 commands to use any method
  - Easy for applied data analysts who use R

- **R + Zelig + Dataverse Network**
  - Write Zelig bridge function \( \rightarrow \) your method appears in the DVN GUI
  - Greatly reduced time from methods development to widespread use
  - Easy for applied researchers who don’t use R
  - (GUI time not wasted: save R code for replication or further analysis)
How to participate

- To increase citations to your data (& web visibility), choose:
  - Sign up for a free dataverse for your web site (no installations, branded as yours, citations for all your data)
  - or install DVN software & you can also give out dataverses

- To increase use of your R package through Zelig and the DVN GUI:
  - Write a simple Zelig bridge function

- To join us:
  - DVN and Zelig are open source projects; contributions welcome!

- 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