

The Dataverse Network: An Infrastructure for Data Sharing

Gary King
Harvard University

February 5, 2008

- 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](https://hdl.handle.net/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

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

\Rightarrow 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.)

Annals of Applied Applications
The Journal for those who truly apply themselves

U.S. open-source group opens chapter in Europe
By Peter Judge
January 31, 2008, 6:37 AM PST

"More open-source projects have originated in Europe than anywhere else in the world," said Bernard Davis, chief executive officer of data-integration specialist Talend, a founding member of the Open Solutions Alliance. Founded a year ago, the OSA has had a U.S. focus until now.

The European chapter of the OSA will be formally incorporated in the next 90 days, and will then look for interoperability work required by European users.

"When the OSA got started, we saw our mission as a global one, but the critical mass of activity so far has been in the U.S.," said Dominic Sartorio, OSA president and senior director of product management at services company SchedulSource. "We were approached two months ago by a group of companies in Europe thinking of forming a global one but focused on Europe. This led us to think that while we are trying to achieve results globally, there are some parts of the world where we are doing best work."

Sartorio sees 2008 as turning point for open source. In late 2007, he promised that OSA would "out-Microsoft Microsoft" in response to a user survey that revealed a demand for interoperability among open-source tools.

"We have encouraged other groups with a similar mission, like the U.K.'s OpenForum, but they tend to focus on pure advocacy," said Sartorio. "The OSA focuses on business applications and their interoperability."

Open-source companies need interoperability because they are usually small and focused, he said. "When they get bought by larger companies, this helps them integrate, but not all open-source companies have that option."

Open source is growing rapidly in Europe, in France, later spending on open-source services and products jumped 50 percent to \$1.27 billion in 2007, according to Paris-based analysts France Author Consult.

Other OSA chapters are expected elsewhere in the world. "Our chapter system is intended to scale around the world," said Sartorio. "I would expect by this time next year we will have chapters up and running in other regions enjoying strong open-source adoption, including Asia and Latin America," he said but focused on Europe. "This led us to think that while we are trying to achieve results globally, there are some parts of the world where we are doing best work."

Sartorio sees 2008 as turning point for open source. In late 2007, he promised that OSA would "out-Microsoft Microsoft" in response to a user survey that revealed a demand for interoperability among open-source tools.

About AAA
Editorial Board
Contact Info
Current Issue
Letters
Dataverse

Your web site

Annals of Applied Applications
The Journal for those who truly apply themselves

U.S. open-source group opens chapter in Europe
By Peter Judge
January 31, 2008, 6:37 AM PST

"More open-source projects have originated in Europe than anywhere else in the world," said Bernard Davis, chief executive officer of data-integration specialist Talend, a founding member of the Open Solutions Alliance. Founded a year ago, the OSA has had a U.S. focus until now.

The European chapter of the OSA will be formally incorporated in the next 90 days, and will then look for interoperability work required by European users.

"When the OSA got started, we saw our mission as a global one, but the critical mass of activity so far has been in the U.S.," said Dominic Sartorio, OSA president and senior director of product management at services company SchedulSource. "We were approached two months ago by a group of companies in Europe thinking of forming a global one but focused on Europe. This led us to think that while we are trying to achieve results globally, there are some parts of the world where we are doing best work."

Sartorio sees 2008 as turning point for open source. In late 2007, he promised that OSA would "out-Microsoft Microsoft" in response to a user survey that revealed a demand for interoperability among open-source tools.

"We have encouraged other groups with a similar mission, like the U.K.'s OpenForum, but they tend to focus on pure advocacy," said Sartorio. "The OSA focuses on business applications and their interoperability."

Open-source companies need interoperability because they are usually small and focused, he said. "When they get bought by larger companies, this helps them integrate, but not all open-source companies have that option."

Open source is growing rapidly in Europe, in France, later spending on open-source services and products jumped 50 percent to \$1.27 billion in 2007, according to Paris-based analysts France Author Consult.

Other OSA chapters are expected elsewhere in the world. "Our chapter system is intended to scale around the world," said Sartorio. "I would expect by this time next year we will have chapters up and running in other regions enjoying strong open-source adoption, including Asia and Latin America," he said but focused on Europe. "This led us to think that while we are trying to achieve results globally, there are some parts of the world where we are doing best work."

Sartorio sees 2008 as turning point for open source. In late 2007, he promised that OSA would "out-Microsoft Microsoft" in response to a user survey that revealed a demand for interoperability among open-source tools.

About AAA
Editorial Board
Contact Info
Current Issue
Letters
Dataverse

STUDIES

Search:

- Supplemental: Volume 1, Number 1 (2007), pp. 1-200
- Supplemental: Volume 1, Number 2 (2007), pp. 201-400
- Supplemental: Volume 1, Number 3 (2007), pp. 401-600
- Supplemental: Volume 1, Number 4 (2007), pp. 601-800
- Supplemental: Volume 1, Number 5 (2007), pp. 801-1000
- Supplemental: Volume 1, Number 6 (2007), pp. 1001-1200
- Supplemental: Volume 1, Number 7 (2007), pp. 1201-1400
- Supplemental: Volume 1, Number 8 (2007), pp. 1401-1600
- Supplemental: Volume 1, Number 9 (2007), pp. 1601-1800
- Supplemental: Volume 1, Number 10 (2007), pp. 1801-2000
- Supplemental: Volume 1, Number 11 (2007), pp. 2001-2200
- Supplemental: Volume 1, Number 12 (2007), pp. 2201-2400
- Supplemental: Volume 1, Number 13 (2007), pp. 2401-2600
- Supplemental: Volume 1, Number 14 (2007), pp. 2601-2800
- Supplemental: Volume 1, Number 15 (2007), pp. 2801-3000
- Supplemental: Volume 1, Number 16 (2007), pp. 3001-3200
- Supplemental: Volume 1, Number 17 (2007), pp. 3201-3400
- Supplemental: Volume 1, Number 18 (2007), pp. 3401-3600
- Supplemental: Volume 1, Number 19 (2007), pp. 3601-3800
- Supplemental: Volume 1, Number 20 (2007), pp. 3801-4000
- Supplemental: Volume 1, Number 21 (2007), pp. 4001-4200
- Supplemental: Volume 1, Number 22 (2007), pp. 4201-4400
- Supplemental: Volume 1, Number 23 (2007), pp. 4401-4600
- Supplemental: Volume 1, Number 24 (2007), pp. 4601-4800
- Supplemental: Volume 1, Number 25 (2007), pp. 4801-5000
- Supplemental: Volume 1, Number 26 (2007), pp. 5001-5200
- Supplemental: Volume 1, Number 27 (2007), pp. 5201-5400
- Supplemental: Volume 1, Number 28 (2007), pp. 5401-5600
- Supplemental: Volume 1, Number 29 (2007), pp. 5601-5800
- Supplemental: Volume 1, Number 30 (2007), pp. 5801-6000
- Supplemental: Volume 1, Number 31 (2007), pp. 6001-6200
- Supplemental: Volume 1, Number 32 (2007), pp. 6201-6400
- Supplemental: Volume 1, Number 33 (2007), pp. 6401-6600
- Supplemental: Volume 1, Number 34 (2007), pp. 6601-6800
- Supplemental: Volume 1, Number 35 (2007), pp. 6801-7000
- Supplemental: Volume 1, Number 36 (2007), pp. 7001-7200
- Supplemental: Volume 1, Number 37 (2007), pp. 7201-7400
- Supplemental: Volume 1, Number 38 (2007), pp. 7401-7600
- Supplemental: Volume 1, Number 39 (2007), pp. 7601-7800
- Supplemental: Volume 1, Number 40 (2007), pp. 7801-8000
- Supplemental: Volume 1, Number 41 (2007), pp. 8001-8200
- Supplemental: Volume 1, Number 42 (2007), pp. 8201-8400
- Supplemental: Volume 1, Number 43 (2007), pp. 8401-8600
- Supplemental: Volume 1, Number 44 (2007), pp. 8601-8800
- Supplemental: Volume 1, Number 45 (2007), pp. 8801-9000
- Supplemental: Volume 1, Number 46 (2007), pp. 9001-9200
- Supplemental: Volume 1, Number 47 (2007), pp. 9201-9400
- Supplemental: Volume 1, Number 48 (2007), pp. 9401-9600
- Supplemental: Volume 1, Number 49 (2007), pp. 9601-9800
- Supplemental: Volume 1, Number 50 (2007), pp. 9801-10000

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

AA 1026 Databases >
Annals of Applied Applications
Search Issues User Guide Site Map Contact Us Login Harvard Archive

REPLICATION DATA FOR IMPROVING FORECASTS OF STATE FAILURES

Creating Information | Documentation, Data and Analysis

Creating Information

Publication Information

How to Cite Gary King, Langche Zeng, 2011, "Replication data for Improving Forecasts of State Failure," *AA 1026* (LIPJGZ30001) URL: [http://dx.doi.org/10.2139/ssrn.1948404](#) (Distribution)

Study Object ID 1026 LIPJGZ30001

Authors Gary King, Langche Zeng

Publication Date 2011

DOI/URL [http://dx.doi.org/10.2139/ssrn.1948404](#) **M R A**

Distribution Contact [http://help.aaapublishing.com](#)

Deposit Date 2010

Replication For King, Gary, Zeng, Langche, 2011, "Improving Forecasts of State Failure," *World Politics*, Vol. 52, No. 4, 2010-10. [http://www.tandf.co.uk/journals/0018-8844/52/4/2010-10](#)

Provenance Gary King Database

Abstract and Notes

We offer the first independent scholarly evaluation of the claims, forecasts, and causal inferences of the State Failure Task Force and their efforts to forecast when states will fail. State failure refers to the collapse of the authority of the central government to impose order, as in civil wars, revolutionary wars,

Annals of Applied Applications
The Journal for those who truly apply themselves

AA 1026 Databases > **Dataverse Network**
Annals of Applied Applications
Search Issues User Guide Site Map Contact Us Login Harvard Archive

REPLICATION DATA FOR IMPROVING FORECASTS OF STATE FAILURES

Creating Information | Documentation, Data and Analysis

Download all files in a single archive file (files that you cannot access will not be downloaded)

| File Name | Description | Owner/Permissions | Type | Available |
|-----------------------|---|-------------------|-----------------|-----------|
| 1. Documentation | Articles related to the study | | | |
| ImprovingForecast.pdf | Improving Forecasts of State Failure | | application/pdf | |
| 2. Replication Data | Articles related to the study | | | |
| 1026.zip | Sample open file for convenience | | zip | |
| 1026 | Full size with license file's ASCII editor | | application/zip | |
| 1026 | Full size with license file's | 11/01/11 | Text document | |
| 1026.txt | ASCII document describing the format of the replication file. | | text/plain | |
| 3. Related Data | Articles related to the study | | | |
| 1026.zip | Country names in original State Failure | | application/zip | |
| 1026 | Country names Date | 08/02/11 | Text document | |
| 1026.txt | Revised Full Data in Original Format, comma delimited text file | | text/plain | |
| 4. Original Data | Documentation for Original | | | |

Annals of Applied Applications
The Journal for those who truly apply themselves

44 2008 Observations
Annals of Applied Applications
Download Dataset Submit and Review Dataset Details Advanced Dataset Actions

REPLICATION DATA FOR CONSTITUENCY SERVICE AND INCUMBENCY ADVANTAGE
DATA FILE: CONSTITUENCYSERVICE.TM

Download Dataset Submit and Review Dataset Details Advanced Dataset Actions

Selected Variables

To rename (subset) a variable into a different one, first, select a variable from the selected variables list, click the arrow button below, and then the name and label of the variable you have chosen appear in the new variable name and label boxes for convenience. You must replace the old variable name with a unique variable name that is not used in the data file; the new variable label is optional and you can leave it blank.

New Variable Name
New Variable Label

Apply Selections

Select variables from table below (selected variables will be displayed above). You do not have access to the searching and analysis functionality for this selected data file. You can only view the variables and their summary statistics.

Show 20 Variables

| Variable Type | Variable Name/Variable Label | Quick Summary |
|-------------------------------------|------------------------------|--------------------------------|
| <input type="checkbox"/> Continuous | INCAD | Incumbency advantage |
| <input type="checkbox"/> Continuous | RESQ | Budget figures |
| <input type="checkbox"/> Continuous | SLC | Money |
| <input type="checkbox"/> String | CD | CONG DISTRICT & COUNTY NUMBERS |

The screenshot shows the website for the *Annals of Applied Applications*, which is described as "the Journal for those who truly apply themselves". The page features a navigation menu on the left with links for "About AAA", "Editorial Board", "Contact Info", "Current Issue", "Letters", and "Dataverse". The main content area displays a list of articles, with the first one titled "REPLICATION DATA FOR CONSTITUENCY SERVICE AND INCUMBENCY ADVANTAGE" and a sub-heading "DATA FILE: CONSTITUENCYSERVICE.TAB". A dropdown menu is open, showing a list of statistical models categorized into "Categorical Data Analysis", "Statistical Inference Models", "Event-Count Models", "Models for Continuous Bounded Dependent Variables", "Gamma Regression for Continuous, Positive Dependent Variables", and "State-Level Modeling Variables".

Annals of Applied Applications
the Journal for those who truly apply themselves

69 1000 References in
 Annals of Applied Applications

Search Options: [Full-Text](#) [Site Map](#) [Contact Us](#) [Log In](#) [Harvard Alerts](#)

REPLICATION DATA FOR CONSTITUENCY SERVICE AND INCUMBENCY ADVANTAGE [Back to Study](#)
 DATA FILE: CONSTITUENCYSERVICE.TAB

Downloaded Subject: [Subject and Previews](#) [Descriptive Statistics](#) [Advanced Statistical Analysis](#)

Selected Variables:

Choose a Statistical Model:

Categorical Data Analysis

- Chi-Square Tabulation

Statistical Inference Models

- Hierarchical Multilevel - Directed Statistical Inference Model for R x C Tables

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
- Serial Network-Poisson Regression for Event Count Dependent Variables

Models for Continuous Bounded Dependent Variables

- Case Proportional Hazard Regression for Duration Dependent Variables
- Exponential Regression for Duration Dependent Variables
- Gamma Regression for Continuous, Positive Dependent Variables
- General Estimating Equation for Gamma Regression

State-Level Modeling Variables

- DE state-level dummy variables
- SE state-level dummy variables
- NE state-level dummy variables
- ME state-level dummy variables
- MD state-level dummy variables
- SEIL state-level dummy variables
- SEIL state-level dummy variables



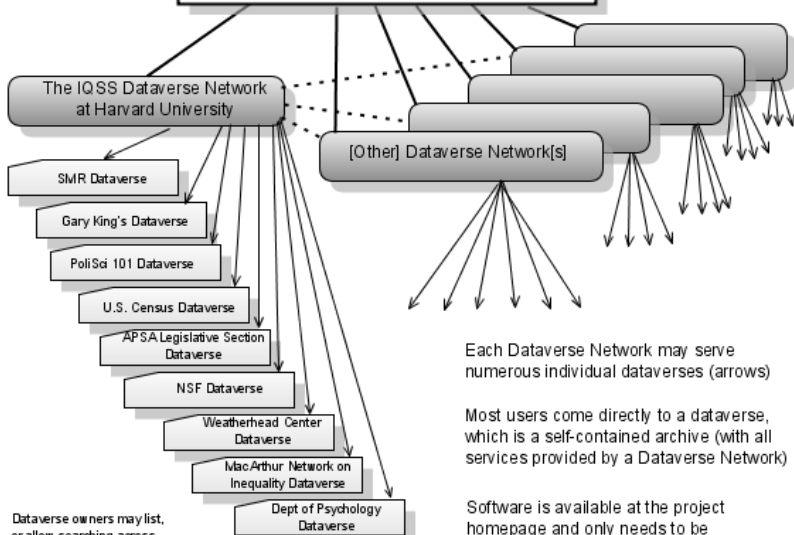
Your web site



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)



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 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
- 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

- **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

<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**