An Introduction to the Dataverse Network as an Infrastructure for Data Sharing

Gary King
Harvard University

December 10, 2007

Papers

- Gary King, "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," *Sociological Methods and Research*, forthcoming.
- Micah Altman and Gary King. "A Proposed Standard for the Scholarly Citation of Quantitative Data," *D-Lib Magazine*.

Infrastructure for Quantitative Data

- Most large data sets: in public archives
- Data used in most published articles: not publicly available
- Many articles cannot be replicated without the original author
- Most data sets from NSF & NIH grants: not publicly available
- Recently, a major archive renumbered all its acquisitions
- Data in different archives have different identifiers
- Changes to data are made; identifiers are reused or deaccessioned; old data are lost
- 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 not put data in public archives?
  - Researchers infrequently use archives
  - They worry about the Archive getting all 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 will propose technological solutions to these political and sociological problems

Requirements for Effective Data Sharing Infrastructure

- Recognition, for authors, journals, etc. in (1) citations to data, (2) citations to 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
- Verification that data associated with an article is unchanged into the future, and 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.
- Persistence Decades from now. . . .
- Ease of Use Neither editors nor authors employ professional archivists
- Legal Protection Publishers have liability procedures for print, but not data. Need to be able to use the expertise of archives or others.

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?== Interuniversity Consortium for Political and Social Research [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

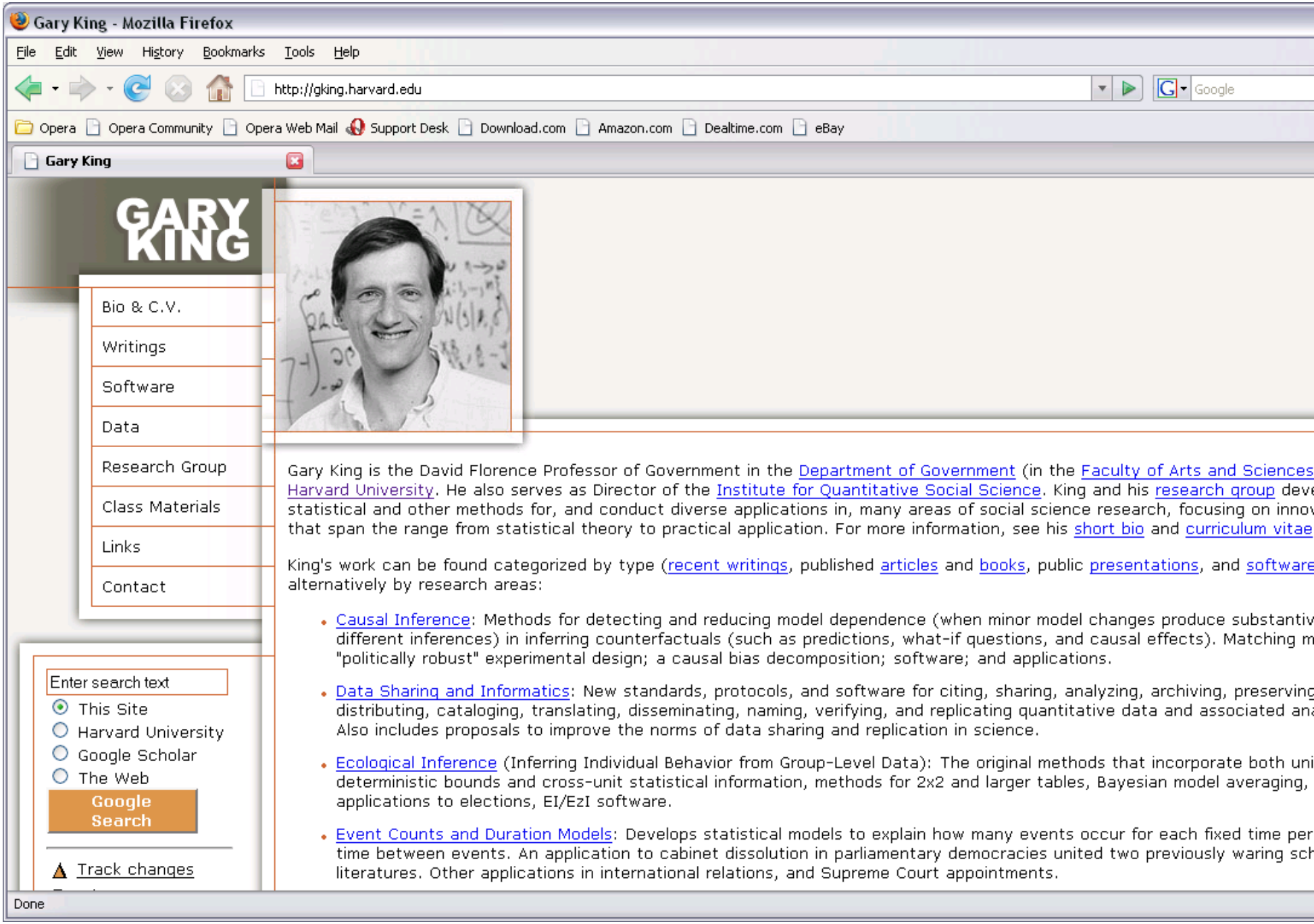
$$\begin{pmatrix} 1 & 4 & 4 & 21 & \cdots & 121 \\ 1 & 2 & 2 & 91 & \cdots & 212 \\ 1 & 9 & 2 & 72 & \cdots & 104 \\ 0 & 2 & 2 & 2 & \cdots & 321 \\ 1 & 6 & 2 & 12 & \cdots & 204 \\ 1 & 9 & 4 & 52 & \cdots & 311 \\ 0 & 3 & 2 & 23 & \cdots & 92 \\ 0 & 2 & 5 & 91 & \cdots & 212 \\ 0 & 5 & 8 & 91 & \cdots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

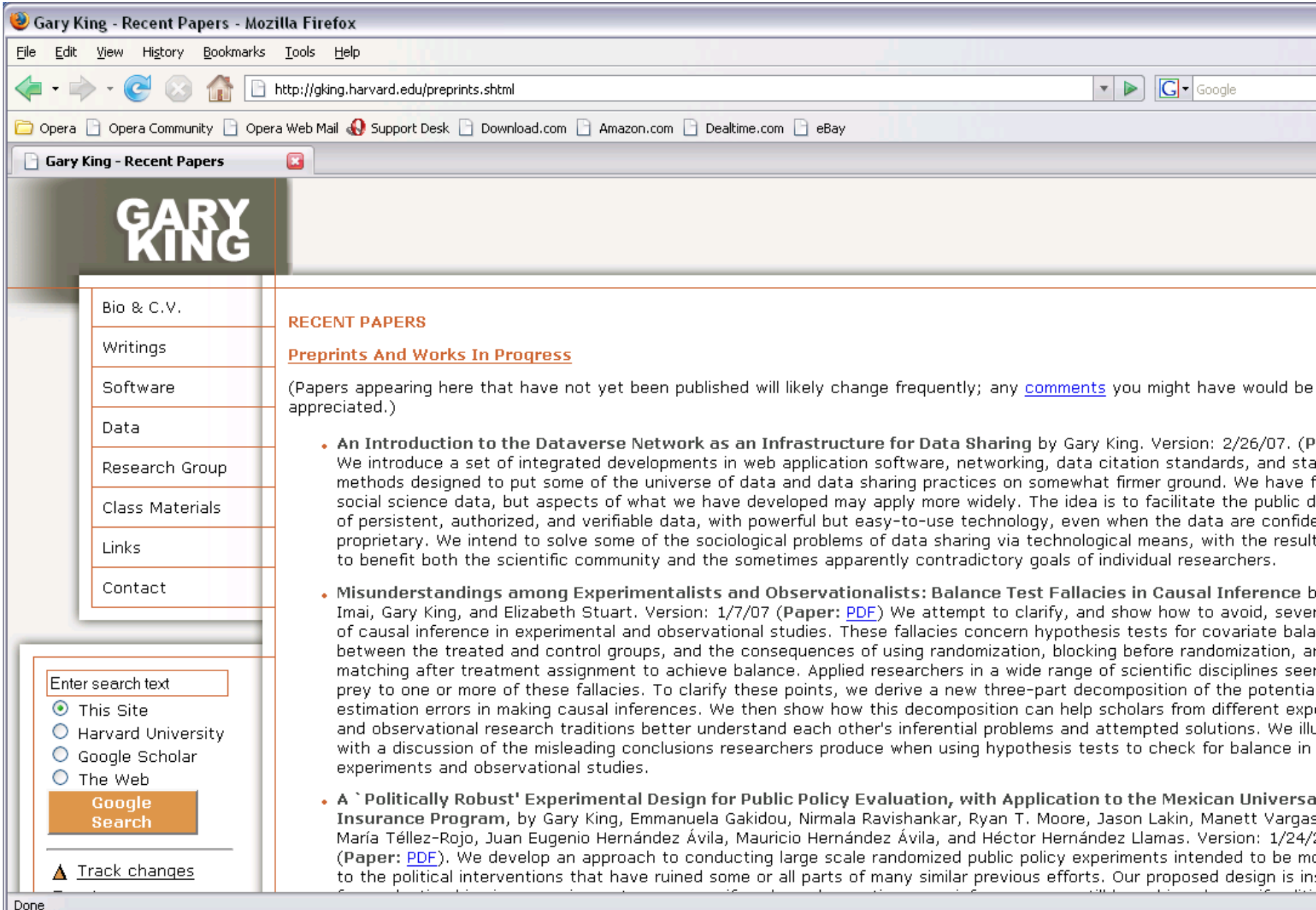
Advantages of UNFs

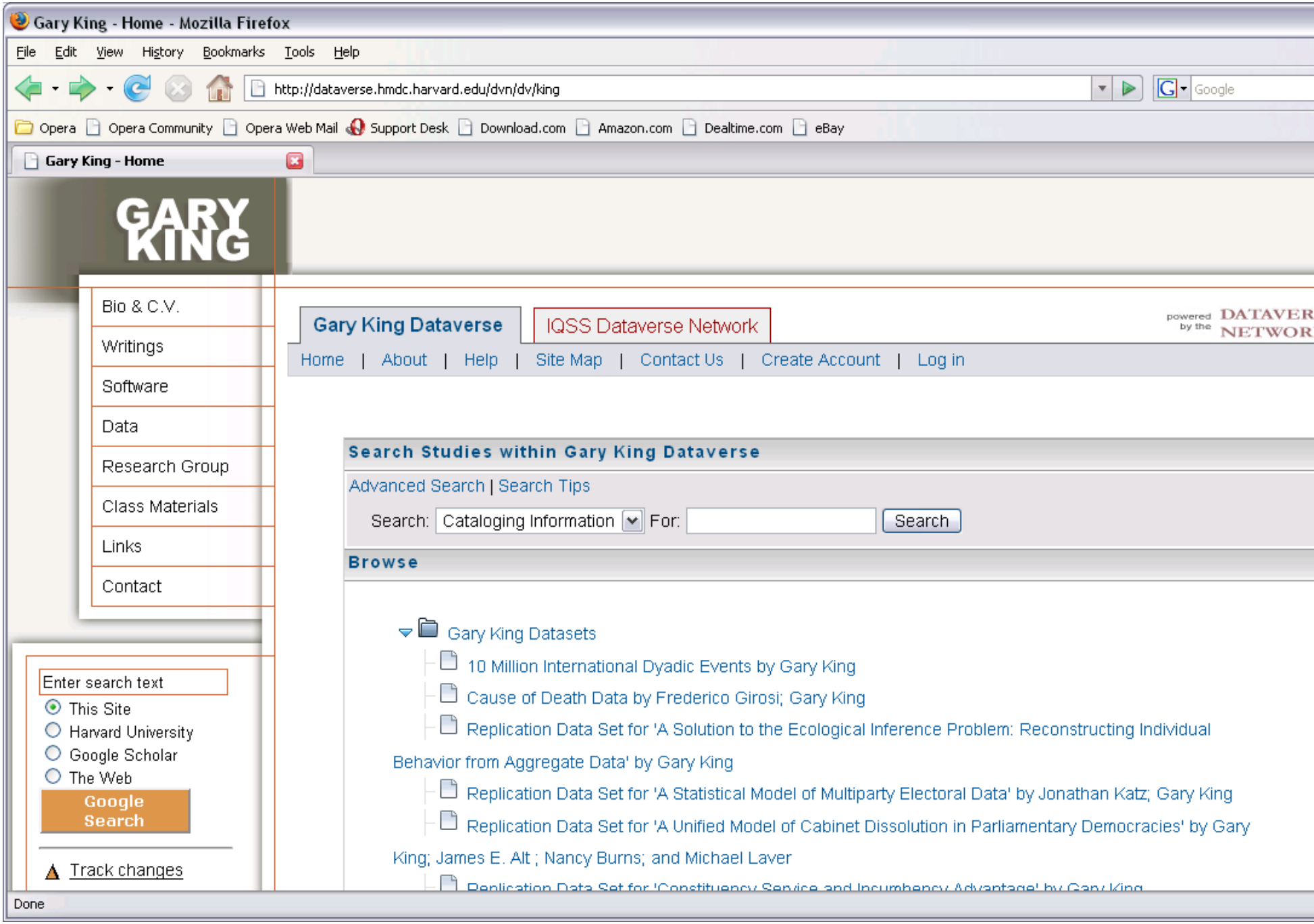
- UNF is calculated from the content not the file: Its the Same UNF regardless of changes in computer hardware, storage medium, operating system, statistical software, database, or spreadsheet software.
- Cryptographic technology: any change in data content changes the UNF. (cannot tinker after the fact!)
- 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

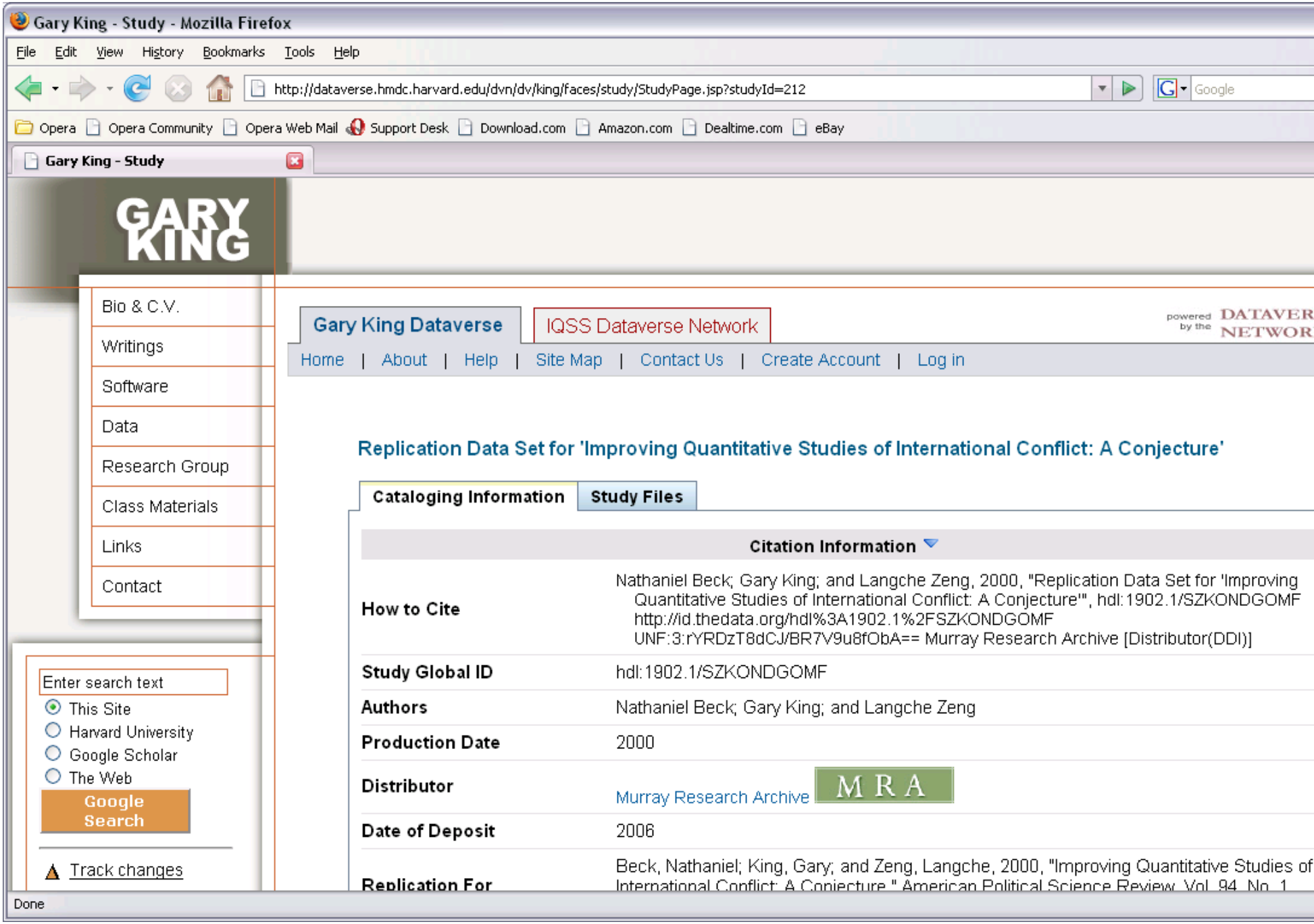
Web 2.0 Terminology

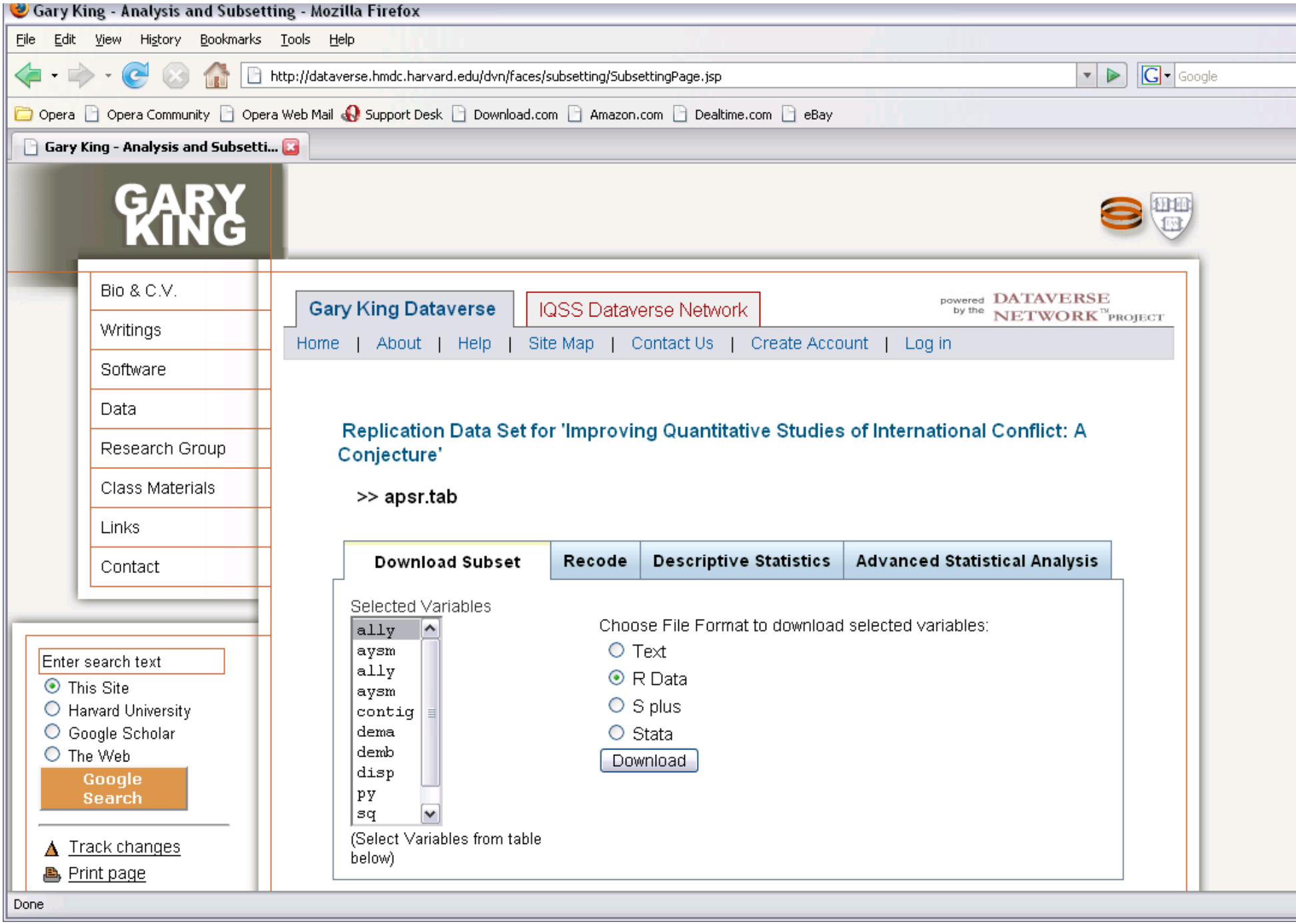
- Software: find CD, install locally, hit next, hit next, hit next...
- Web application software: no installation; load web browser and run (Dataverse Network Software)
- Host: The computers where the web application software runs (universities, archives, libraries)
- Virtual host: Where the web application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

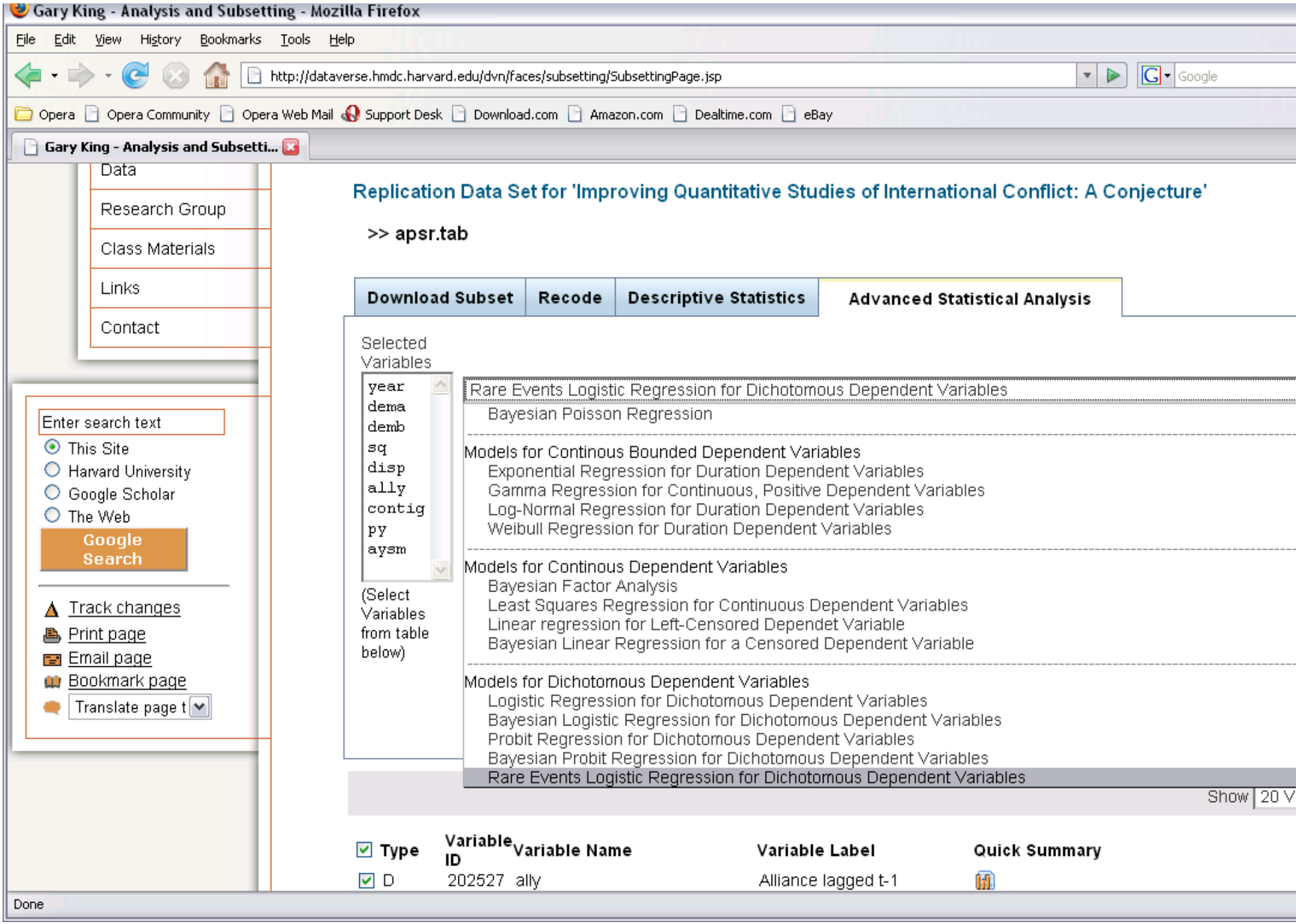

Gary King - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://gking.harvard.edu
Google
Opera Opera Community Opera Web Mail Support Desk Download.com Amazon.com Dealtime.com eBay
Gary King
GARY KING
Bio & C.V.
Writings
Software
Data
Research Group
Class Materials
Links
Contact
Enter search text
This Site
Harvard University
Google Scholar
The Web
Google Search
Track changes
Gary King is the David Florence Professor of Government in the Department of Government (in the Faculty of Arts and Sciences Harvard University. He also serves as Director of the Institute for Quantitative Social Science. King and his research group deve statistical and other methods for, and conduct diverse applications in, many areas of social science research, focusing on innov that span the range from statistical theory to practical application. For more information, see his short bio and curriculum vitae
King's work can be found categorized by type (recent writings, published articles and books, public presentations, and software alternatively by research areas:
Causal Inference: Methods for detecting and reducing model dependence (when minor model changes produce substantiv different inferences) in inferring counterfactuals (such as predictions, what-if questions, and causal effects). Matching m "politically robust" experimental design; a causal bias decomposition; software; and applications.
Data Sharing and Informatics: New standards, protocols, and software for citing, sharing, analyzing, archiving, preservinc distributing, cataloging, translating, disseminating, naming, verifying, and replicating quantitative data and associated ana Also includes proposals to improve the norms of data sharing and replication in science.
Ecological Inference (Inferring Individual Behavior from Group-Level Data): The original methods that incorporate both uni deterministic bounds and cross-unit statistical information, methods for 2x2 and larger tables, Bayesian model averaging, applications to elections, EI/EzI software.
Event Counts and Duration Models: Develops statistical models to explain how many events occur for each fixed time per time between events. An application to cabinet dissolution in parliamentary democracies united two previously waring sch literatures. Other applications in international relations, and Supreme Court appointments.
Done

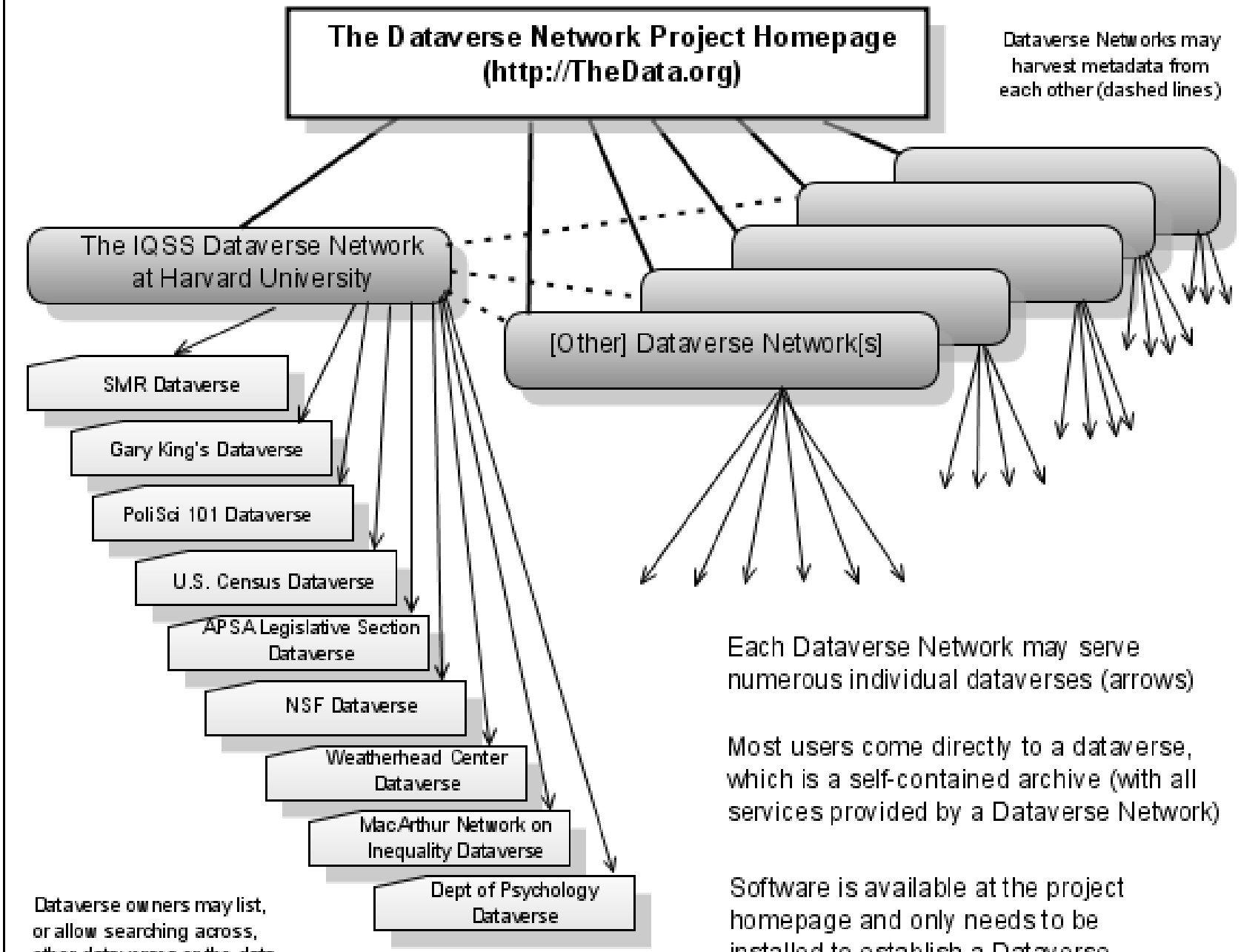

Gary King - Recent Papers - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://gking.harvard.edu/preprints.shtml
Google
Opera Opera Community Opera Web Mail Support Desk Download.com Amazon.com Dealtime.com eBay
Gary King - Recent Papers
GARY KING
Bio & C.V.
Writings
Software
Data
Research Group
Class Materials
Links
Contact
Enter search text
This Site
Harvard University
Google Scholar
The Web
Google Search
Track changes
RECENT PAPERS
Preprints And Works In Progress
(Papers appearing here that have not yet been published will likely change frequently; any comments you might have would be appreciated.)
An Introduction to the Dataverse Network as an Infrastructure for Data Sharing by Gary King. Version: 2/26/07.
We introduce a set of integrated developments in web application software, networking, data citation standards, and sta
methods designed to put some of the universe of data and data sharing practices on somewhat firmer ground. We have f
social science data, but aspects of what we have developed may apply more widely. The idea is to facilitate the public d
of persistent, authorized, and verifiable data, with powerful but easy-to-use technology, even when the data are confide
proprietary. We intend to solve some of the sociological problems of data sharing via technological means, with the result
to benefit both the scientific community and the sometimes apparently contradictory goals of individual researchers.
Misunderstandings among Experimentalists and Observationalists: Balance Test Fallacies in Causal Inference b
Imai, Gary King, and Elizabeth Stuart. Version: 1/7/07 (Paper: PDF) We attempt to clarify, and show how to avoid, sever
of causal inference in experimental and observational studies. These fallacies concern hypothesis tests for covariate bala
between the treated and control groups, and the consequences of using randomization, blocking before randomization, a
matching after treatment assignment to achieve balance. Applied researchers in a wide range of scientific disciplines seer
prey to one or more of these fallacies. To clarify these points, we derive a new three-part decomposition of the potentia
estimation errors in making causal inferences. We then show how this decomposition can help scholars from different exp
and observational research traditions better understand each other's inferential problems and attempted solutions. We illu
with a discussion of the misleading conclusions researchers produce when using hypothesis tests to check for balance in
experiments and observational studies.
A `Politically Robust' Experimental Design for Public Policy Evaluation, with Application to the Mexican Universa
Insurance Program, by Gary King, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas
María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. Version: 1/24/
(Paper: PDF). We develop an approach to conducting large scale randomized public policy experiments intended to be m
to the political interventions that have ruined some or all parts of many similar previous efforts. Our proposed design is in
Done


Gary King - Home - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://dataverse.hmdc.harvard.edu/dvn/dv/king
Google
Opera Opera Community Opera Web Mail Support Desk Download.com Amazon.com Dealtime.com eBay
Gary King - Home
GARY KING
Bio & C.V.
Writings
Software
Data
Research Group
Class Materials
Links
Contact
Gary King Dataverse
IQSS Dataverse Network
powered by the DATAVER NETWOR
Home | About | Help | Site Map | Contact Us | Create Account | Log in
Search Studies within Gary King Dataverse
Advanced Search | Search Tips
Search: Cataloging Information For: Search
Browse
Gary King Datasets
10 Million International Dyadic Events by Gary King
Cause of Death Data by Frederico Girosi; Gary King
Replication Data Set for 'A Solution to the Ecological Inference Problem: Reconstructing Individual
Behavior from Aggregate Data' by Gary King
Replication Data Set for 'A Statistical Model of Multiparty Electoral Data' by Jonathan Katz; Gary King
Replication Data Set for 'A Unified Model of Cabinet Dissolution in Parliamentary Democracies' by Gary
King; James E. Alt ; Nancy Burns; and Michael Laver
Enter search text
This Site
Harvard University
Google Scholar
The Web
Google Search
Track changes
Done


Gary King - Study - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://dataverse.hmdc.harvard.edu/dvn/dv/king/faces/study/StudyPage.jsp?studyId=212
Opera Opera Community Opera Web Mail Support Desk Download.com Amazon.com Dealtime.com eBay
Gary King - Study
GARY KING
Bio & C.V.
Writings
Software
Data
Research Group
Class Materials
Links
Contact
Enter search text
This Site
Harvard University
Google Scholar
The Web
Google Search
Track changes
Gary King Dataverse
IQSS Dataverse Network
powered by the DATAVER NETWOR
Home | About | Help | Site Map | Contact Us | Create Account | Log in
Replication Data Set for 'Improving Quantitative Studies of International Conflict: A Conjecture'
Cataloging Information
Study Files
Citation Information
How to Cite
Nathaniel Beck; Gary King; and Langche Zeng, 2000, "Replication Data Set for 'Improving Quantitative Studies of International Conflict: A Conjecture'", hdl:1902.1/SZKONDGOMF http://id.thedata.org/hdl%3A1902.1%2FSZKONDGOMF UNF:3:rYRDzT8dCJ/BR7V9u8fObA== Murray Research Archive [Distributor(DDI)]
Study Global ID
hdl:1902.1/SZKONDGOMF
Authors
Nathaniel Beck; Gary King; and Langche Zeng
Production Date
2000
Distributor
Murray Research Archive
M R A
Date of Deposit
2006
Replication For
Beck, Nathaniel; King, Gary; and Zeng, Langche, 2000, "Improving Quantitative Studies of International Conflict: A Conjecture," American Political Science Review, Vol. 94, No. 1
Done


Gary King - Analysis and Subsetting - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://dataverse.hmdc.harvard.edu/dvn/faces/subsetting/SubsettingPage.jsp
Opera Opera Community Opera Web Mail Support Desk Download.com Amazon.com Dealtime.com eBay
Gary King - Analysis and Subsetti...
GARY KING
Bio & C.V.
Writings
Software
Data
Research Group
Class Materials
Links
Contact
Gary King Dataverse
IQSS Dataverse Network
powered by the DATAVERSE NETWORK PROJECT
Home | About | Help | Site Map | Contact Us | Create Account | Log in
Replication Data Set for 'Improving Quantitative Studies of International Conflict: A Conjecture'
>> apsr.tab
Download Subset
Recode
Descriptive Statistics
Advanced Statistical Analysis
Selected Variables
ally
aysm
ally
aysm
contig
dema
demb
disp
py
sq
(Select Variables from table below)
Choose File Format to download selected variables:
Text
R Data
S plus
Stata
Download
Enter search text
This Site
Harvard University
Google Scholar
The Web
Google Search
Track changes
Print page
Done


Gary King - Analysis and Subsetting - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://dataverse.hmdc.harvard.edu/dvn/faces/subsetting/SubsettingPage.jsp
Opera Opera Community Opera Web Mail Support Desk Download.com Amazon.com Dealtime.com eBay
Gary King - Analysis and Subsetti...
Data
Research Group
Class Materials
Links
Contact
Enter search text
This Site
Harvard University
Google Scholar
The Web
Google Search
Track changes
Print page
Email page
Bookmark page
Translate page t
Replication Data Set for 'Improving Quantitative Studies of International Conflict: A Conjecture'
>> apsr.tab
Download Subset
Recode
Descriptive Statistics
Advanced Statistical Analysis
Selected Variables
year
dema
demb
sq
disp
ally
contig
py
aysm
(Select Variables from table below)
Rare Events Logistic Regression for Dichotomous Dependent Variables
Bayesian Poisson Regression
Models for Continous Bounded Dependent Variables
Exponential Regression for Duration Dependent Variables
Gamma Regression for Continuous, Positive Dependent Variables
Log-Normal Regression for Duration Dependent Variables
Weibull Regression for Duration Dependent Variables
Models for Continous Dependent Variables
Bayesian Factor Analysis
Least Squares Regression for Continuous Dependent Variables
Linear regression for Left-Censored Dependet Variable
Bayesian Linear Regression for a Censored Dependent Variable
Models for Dichotomous Dependent Variables
Logistic Regression for Dichotomous Dependent Variables
Bayesian Logistic Regression for Dichotomous Dependent Variables
Probit Regression for Dichotomous Dependent Variables
Bayesian Probit Regression for Dichotomous Dependent Variables
Rare Events Logistic Regression for Dichotomous Dependent Variables
Show 20 V
Type
Variable ID
Variable Name
Variable Label
Quick Summary
D
202527
ally
Alliance lagged t-1
Done


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.

Your Own Dataverse

- Full service virtual archive, with numerous data services (citation, metadata, archiving, subsetting, conversion, translation, analysis, . . .)
- Hierarchically organized collection of data sets on chosen subjects; your view of the universe of data
- Branded as yours: with the look and feel of your web site
- Easy to setup: give the Dataverse Network your web style information, and include a link to your newly created dataverse
- Easy to manage: no software installation, no hardware, no backups, no need to worry about professional archiving standards, it still exists if you move or can be branded differently, etc.
- Reuse: a data set may appear on different dataverses if desired

Dataverse Uses

- Authors, for their own data
- Journals, for their replication data archives
- Future Researchers: browsing and searching for data, forward citation search, verification via UNFs, subsetting, analysis
- Teachers, a dataverse for class as informative 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
- Accessing data outside the Dataverse Network

The User Experience

- Find a dataverse of interest (from a Dataverse Network site)
- Browse or search for data of interest
- Find data, see abstract, read documentation
- (Or with a citation, go straight to its metadata)
- Check for new versions, or other data sets which incorporate this one
- If associated with an article, verify that the data hasn't changed
- Authenticate yourself and get access authorization
- Run descriptive statistics and graphics
- Run advanced, cutting-edge statistical analyses (with replication code)
- Subset data (only women 18-24 who voted for Clinton)
- Translate to a convenient format
- Download subset (with citation for the subset)

A Journal Dataverse for a Replication Data Archive

- Setup a journal dataverse: a few minutes of web work
- Dataverse is branded as the journal's web site
- Copyeditor ensures: data must be properly cited
- Author of accepted article given password to upload data and obtain citation
- No hardware or software installation; almost no extra work
- professional archiving standards automatically followed
- When the editor or publisher changes, the dataverse will continue to exist; branding can easily be changed
- Replication policies cause journals to be cited three times as frequently! (with dataverse, it should be more)

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

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
  - If you are at a university and put data on your web site, you are probably violating the law
  - DVN automates (aspects of) the IRB process

Possible Future Directions

- Campaign to sign up authors, journals, etc for dataverses (experiments indicate $> 90\%$ success)
- Install Dataverse Network software elsewhere
- Add more data services: makes including data more attractive
  - Spatially enabling data (planned)
  - Include more models in Zelig, and thus in Dataverse (underway)
  - Include more existing data providers (underway)
  - More data conversion options
  - Implement Unicode so Dataverse can operate in many languages
  - Include in other products, like GenePattern (underway)
  - Many back end technical developments (continuing)
- Result:
  - Empower existing archives to give back (dataverses) to their data contributors
  - Massive increase in publicly available, preserved data
  - Greatly increase ease of use and access for data
  - Extend data model to new data types (field notes, audio, video, etc.)

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

For more information

http://TheData.org