Microsoft® Research
Faculty Summit 2010

The Dataverse Network – Combining Preservation with Scholarly Recognition

Mercè Crosas, Ph.D.
Director of Product Development
Institute for Quantitative Social Science (IQSS)
Harvard University

Acknowledgements

This work is done with contribution from:

- The Dataverse Network software development team at IQSS: Gustavo Durand, Ellen Kraffmiller, Kevin Condon, Leonid Andrev, Bob Treacy, Michael Heppler, Steve Kraffmiller
- Gary King, Albert J. Weatherhead III University Professor, Harvard University
- Micah Altman, Senior Research Scientist, IQSS, Harvard University

The Problem with Data

Professional archives focus on long term access by the wider community

- Persistent identifiers
- Fixity
- Backups and recovery
- Metadata standards
- Conversion standards
- Preservation standards

… but do not attract maximum contribution from data owners

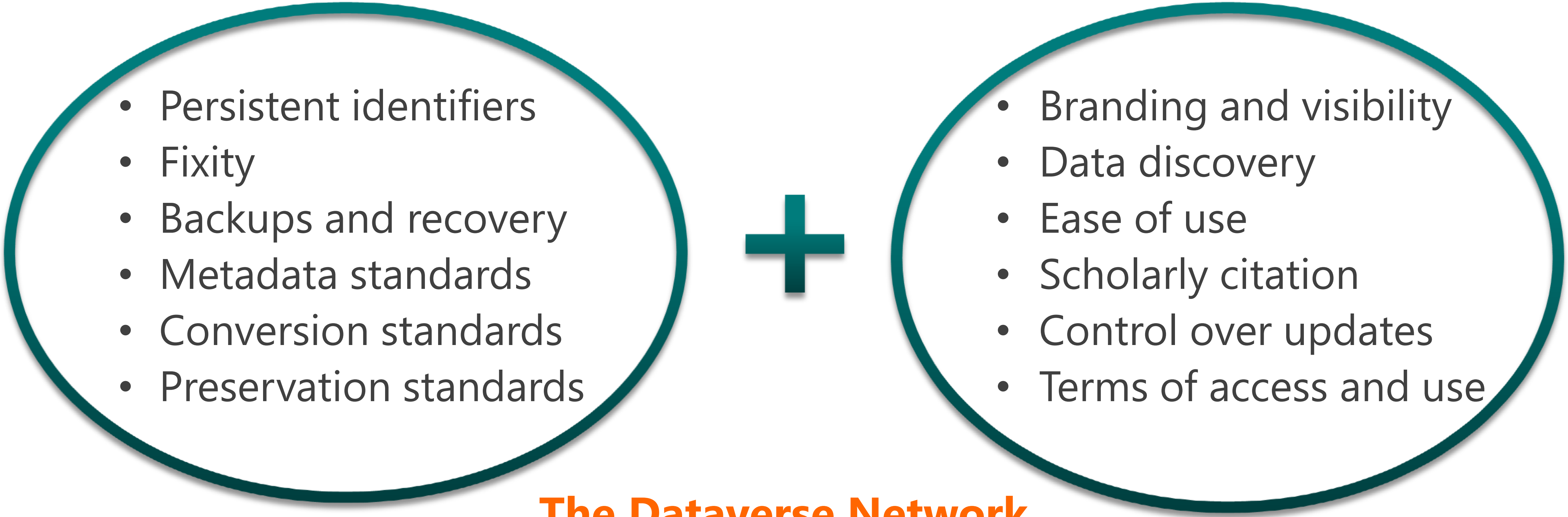
≠

Data owners focus on recognition and control

- Branding and visibility
- Data discovery
- Ease of use
- Scholarly citation
- Control over updates
- Terms of access and use

… but do not assure long-term use as would a professional archive

A Solution through technology



Centralized Archival Infrastructure

- Persistent identifiers
- Fixity
- Backups and recovery
- Metadata standards
- Conversion standards
- Preservation standards

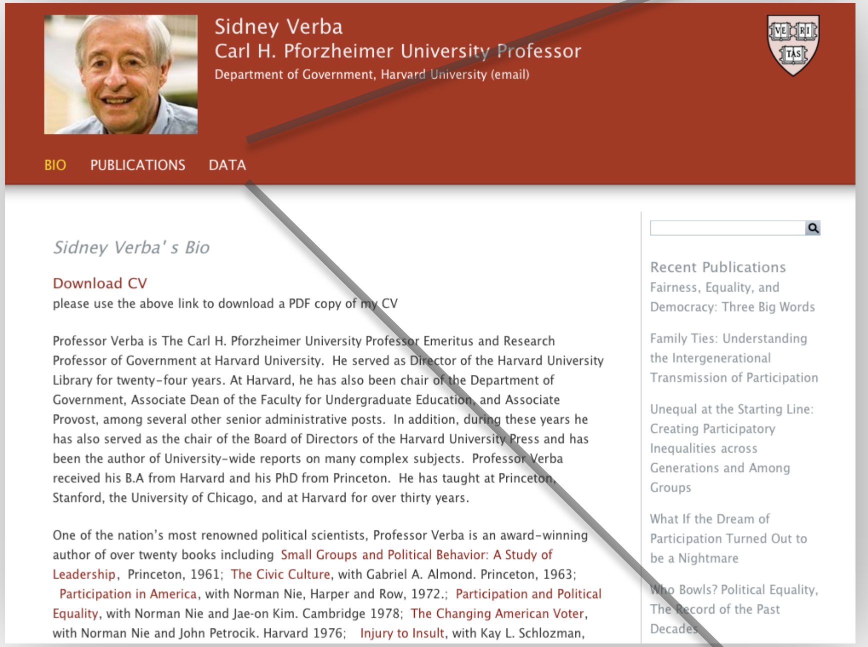
Distributed Recognition and Control

- Branding and visibility
- Data discovery
- Ease of use
- Scholarly citation
- Control over updates
- Terms of access and use

The Dataverse Network

Enables people to solve a political problem through technology – Combining centralized archiving with distributed ownership

Use Case 1: Dataverse for Researchers – Web Visibility and Branding



Sidney Verba
Carl H. Pforzheimer University Professor
Department of Government, Harvard University (email)

BIO PUBLICATIONS DATA

Sidney Verba' s Bio

Download CV
please use the above link to download a PDF copy of my CV

Professor Verba is The Carl H. Pforzheimer University Professor Emeritus and Research Professor of Government at Harvard University. He served as Director of the Harvard University Library for twenty-four years. At Harvard, he has also been chair of the Department of Government, Associate Dean of the Faculty for Undergraduate Education, and Associate Provost, among several other senior administrative posts. In addition, during these years he has also served as the chair of the Board of Directors of the Harvard University Press and has been the author of University-wide reports on many complex subjects. Professor Verba received his B.A from Harvard and his PhD from Princeton. He has taught at Princeton, Stanford, the University of Chicago, and at Harvard for over thirty years.

One of the nation's most renowned political scientists, Professor Verba is an award-winning author of over twenty books including Small Groups and Political Behavior: A Study of Leadership, Princeton, 1961; The Civic Culture, with Gabriel A. Almond. Princeton, 1963; Participation in America, with Norman Nie, Harper and Row, 1972.; Participation and Political Equality, with Norman Nie and Jae-on Kim. Cambridge 1978; The Changing American Voter, with Norman Nie and John Petrocik. Harvard 1976; Injury to Insult, with Kay L. Schlozman,

Recent Publications

Fairness, Equality, and Democracy: Three Big Words

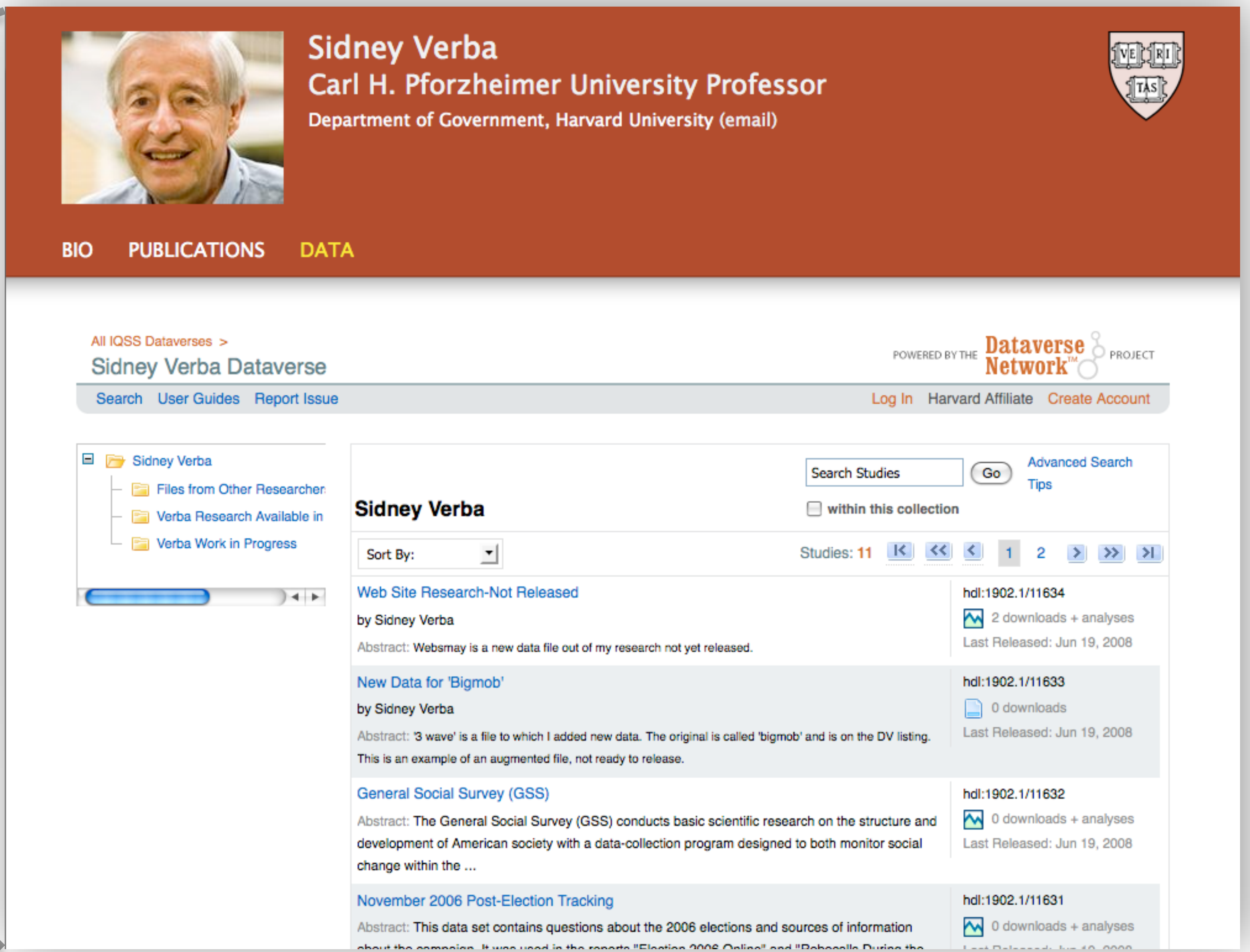
Family Ties: Understanding the Intergenerational Transmission of Participation

Unequal at the Starting Line: Creating Participatory Inequalities across Generations and Among Groups

What If the Dream of Participation Turned Out to be a Nightmare

Who Bowls? Political Equality, The Record of the Past Decades

Your website



Sidney Verba
Carl H. Pforzheimer University Professor
Department of Government, Harvard University (email)

BIO PUBLICATIONS DATA

All IQSS Dataverses >
Sidney Verba Dataverse

POWERED BY THE Dataverse Network™ PROJECT

Search User Guides Report Issue Log In Harvard Affiliate Create Account

Sidney Verba
- Files from Other Researcher
- Verba Research Available in
- Verba Work in Progress

Sidney Verba

Search Studies Go Advanced Search Tips

□ within this collection

Sort By: Studies: 11 1 2

Web Site Research-Not Released
by Sidney Verba
Abstract: Websmay is a new data file out of my research not yet released.
hdl:1902.1/11634
2 downloads + analyses
Last Released: Jun 19, 2008

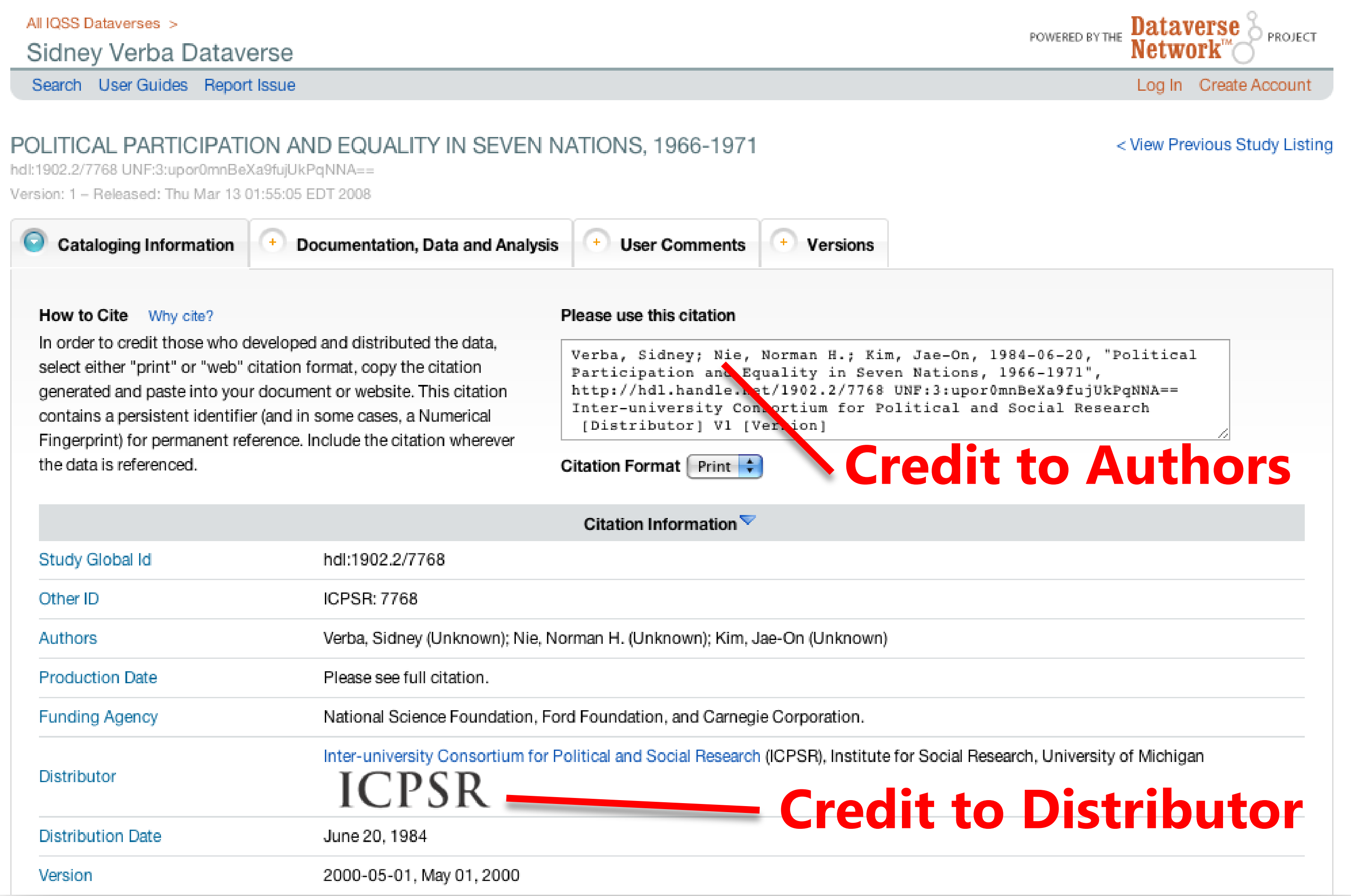
New Data for 'Bigmob'
by Sidney Verba
Abstract: '3 wave' is a file to which I added new data. The original is called 'bigmob' and is on the DV listing. This is an example of an augmented file, not ready to release.
hdl:1902.1/11633
0 downloads
Last Released: Jun 19, 2008

General Social Survey (GSS)
Abstract: The General Social Survey (GSS) conducts basic scientific research on the structure and development of American society with a data-collection program designed to both monitor social change within the ...
hdl:1902.1/11632
0 downloads + analyses
Last Released: Jun 19, 2008

November 2006 Post-Election Tracking
Abstract: This data set contains questions about the 2006 elections and sources of information
hdl:1902.1/11631
0 downloads + analyses

Your dataverse

Use Case 1: Dataverse for Researchers – Recognition through Citatition



All IQSS Dataverses >

Sidney Verba Dataverse

POWERED BY THE Dataverse Network™ PROJECT

Search   User Guides   Report Issue   Log In   Create Account

POLITICAL PARTICIPATION AND EQUALITY IN SEVEN NATIONS, 1966-1971

< View Previous Study Listing

hdl:1902.2/7768 UNF:3:upor0mnBeXa9fujUkPqNNA==

Version: 1 – Released: Thu Mar 13 01:55:05 EDT 2008

Cataloging Information | Documentation, Data and Analysis | User Comments | Versions

How to Cite Why cite?

In order to credit those who developed and distributed the data, select either "print" or "web" citation format, copy the citation generated and paste into your document or website. This citation contains a persistent identifier (and in some cases, a Numerical Fingerprint) for permanent reference. Include the citation wherever the data is referenced.

Please use this citation

```
Verba, Sidney; Nie, Norman H.; Kim, Jae-On, 1984-06-20, "Political
Participation and Equality in Seven Nations, 1966-1971",
http://hdl.handle.net/1902.2/7768 UNF:3:upor0mnBeXa9fujUkPqNNA==
Inter-university Consortium for Political and Social Research
 [Distributor] V1 [Version]
```

Citation Format: Print

Credit to Authors

Citation Information

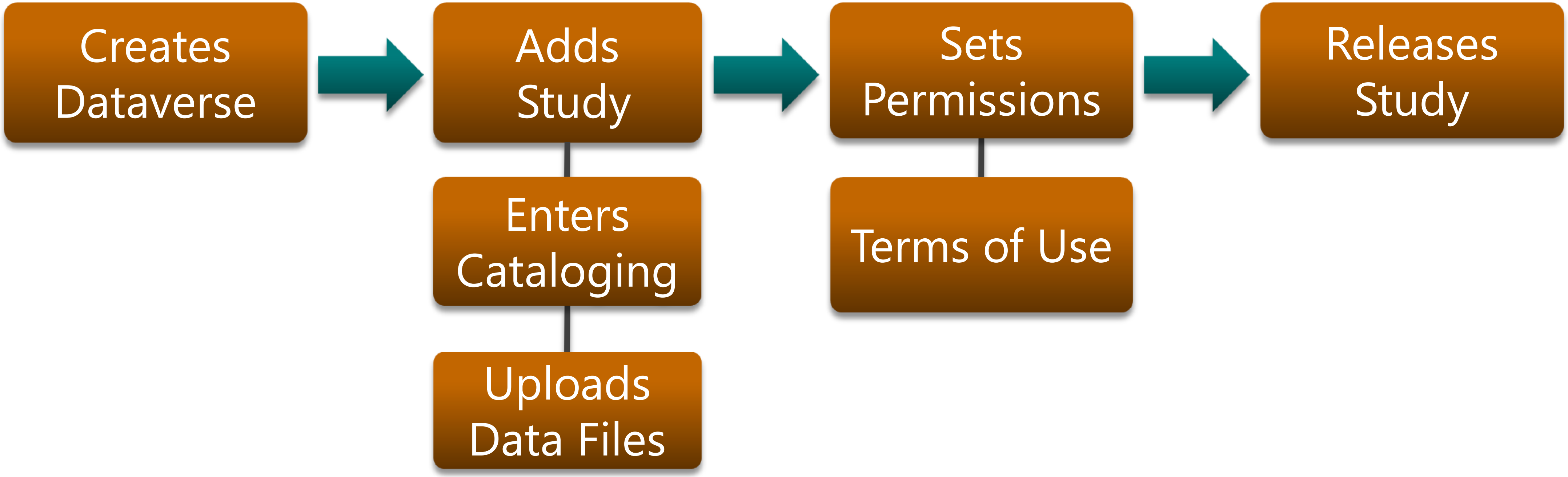
Study Global Id	hdl:1902.2/7768
Other ID	ICPSR: 7768
Authors	Verba, Sidney (Unknown); Nie, Norman H. (Unknown); Kim, Jae-On (Unknown)
Production Date	Please see full citation.
Funding Agency	National Science Foundation, Ford Foundation, and Carnegie Corporation.
Distributor	Inter-university Consortium for Political and Social Research (ICPSR), Institute for Social Research, University of Michigan ICPSR
Distribution Date	June 20, 1984
Version	2000-05-01, May 01, 2000

Credit to Distributor

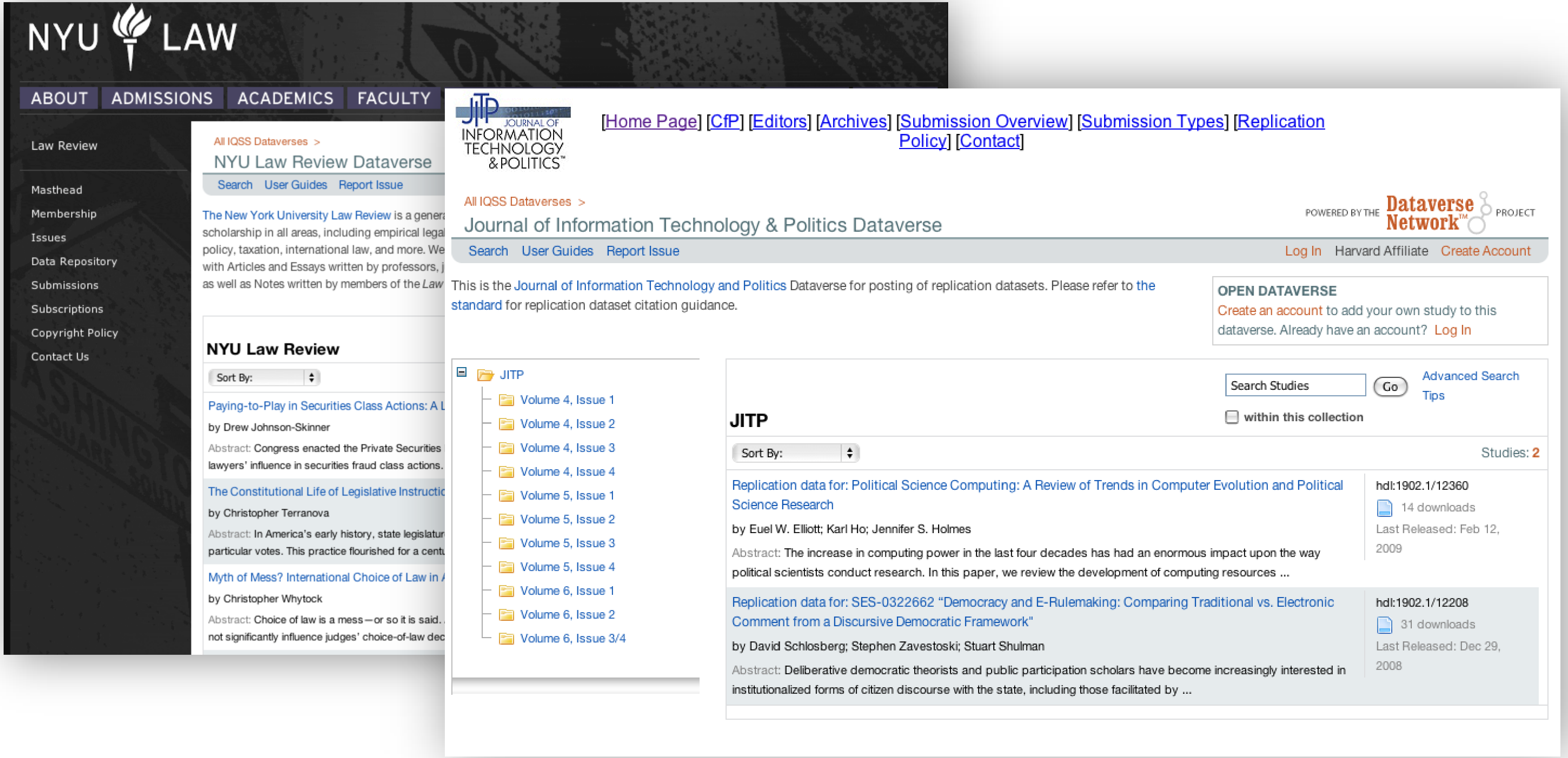
Use Case 1: Dataverse for Researchers – Control and Ownership

- The researcher is the administrator of his/her dataverse.
- Manages data curation workflow.
- Sets permissions and terms of use to datasets.

A Researcher …



Use Case 2: Dataverse for Journals – Replication of Published Work



NYU Law Review Dataverse

The New York University Law Review is a genera
scholarship in all areas, including empirical legal
policy, taxation, international law, and more. We
with Articles and Essays written by professors, j
as well as Notes written by members of the *Law*

NYU Law Review

Paying-to-Play in Securities Class Actions: A L
by Drew Johnson-Skinner
Abstract: Congress enacted the Private Securities
lawyers' influence in securities fraud class actions.

The Constitutional Life of Legislative Instructio
by Christopher Terranova
Abstract: In America's early history, state legislatur
particular votes. This practice flourished for a centu

Myth of Mess? International Choice of Law in A
by Christopher Whytock
Abstract: Choice of law is a mess—or so it is said.
not significantly influence judges' choice-of-law dec

Journal of Information Technology & Politics Dataverse

This is the Journal of Information Technology and Politics Dataverse for posting of replication datasets. Please refer to the standard for replication dataset citation guidance.

OPEN DATAVERSE
Create an account to add your own study to this dataverse. Already have an account? Log In

JITP

Studies: 2

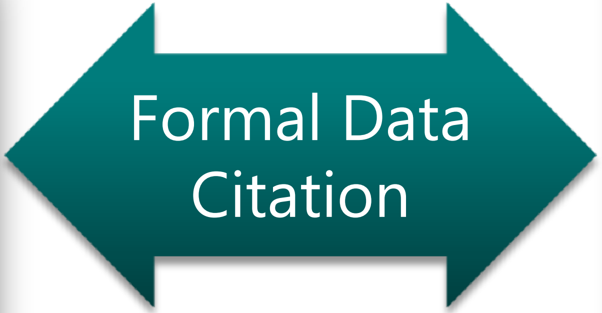
Replication data for: Political Science Computing: A Review of Trends in Computer Evolution and Political Science Research
by Euel W. Elliott; Karl Ho; Jennifer S. Holmes
Abstract: The increase in computing power in the last four decades has had an enormous impact upon the way political scientists conduct research. In this paper, we review the development of computing resources ...
hdl:1902.1/12360
14 downloads
Last Released: Feb 12, 2009

Replication data for: SES-0322662 "Democracy and E-Rulemaking: Comparing Traditional vs. Electronic Comment from a Discursive Democratic Framework"
by David Schlosberg; Stephen Zavestoski; Stuart Shulman
Abstract: Deliberative democratic theorists and public participation scholars have become increasingly interested in institutionalized forms of citizen discourse with the state, including those facilitated by ...
hdl:1902.1/12208
31 downloads
Last Released: Dec 29, 2008

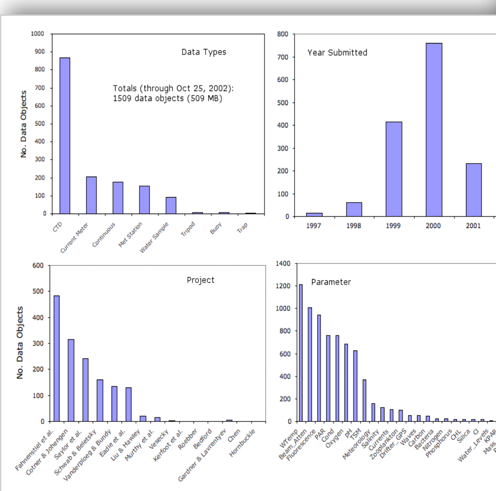
Use Case 2: Dataverse for Journals – From Publication to Data

Journal, Book

Datasets, text, images



Formal Data Citation



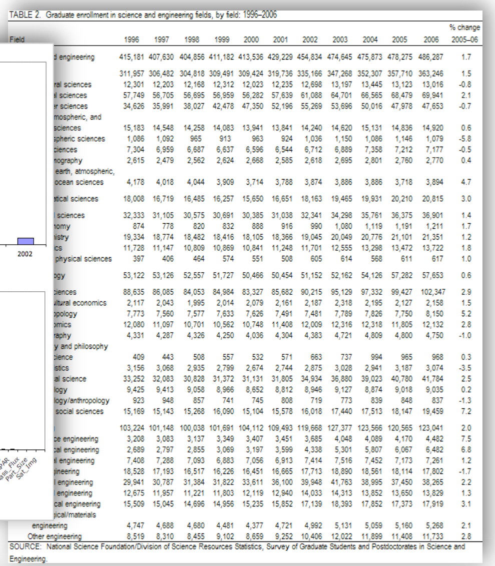
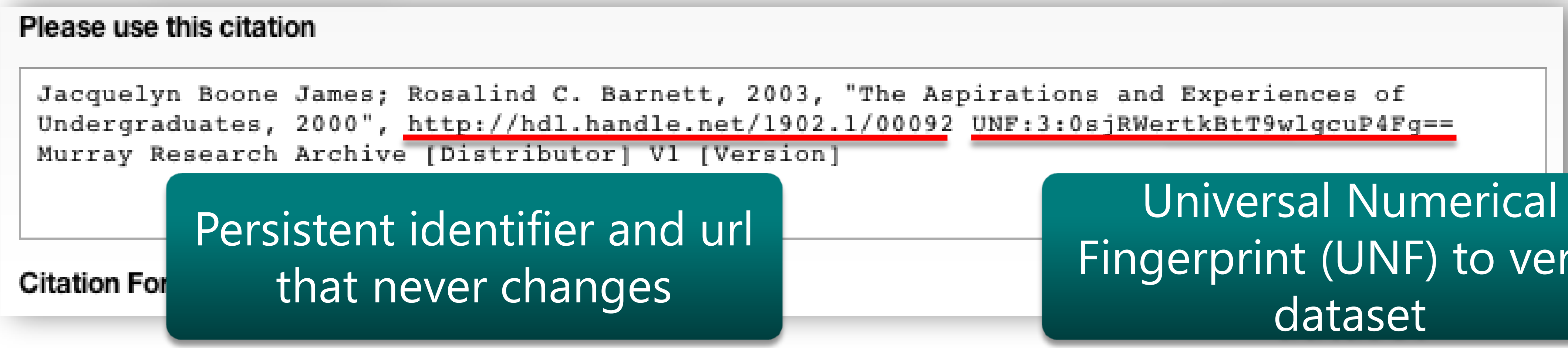


TABLE 2. Graduate enrollment in science and engineering fields, by field: 1996–2006

Field	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	% change 2005–06
d engineering	415,181	407,630	404,856	411,182	413,536	429,229	454,834	474,645	475,873	478,275	486,287	1.7
	311,957	306,482	304,818	309,491	309,424	319,736	335,166	347,268	352,307	357,710	363,246	1.5
ral sciences	12,301	12,203	12,168	12,312	12,023	12,235	12,698	13,197	13,445	13,123	13,016	-0.8
l sciences	57,749	56,705	56,695	56,959	56,282	57,639	61,088	64,701	66,565	68,479	69,941	2.1
r sciences	34,626	35,991	38,027	42,478	47,350	52,196	55,269	53,696	50,016	47,978	47,653	-0.7
mospheric, and												
sciences	15,183	14,548	14,258	14,083	13,941	13,841	14,240	14,620	15,131	14,836	14,920	0.6
spheric sciences	1,086	1,092	965	913	963	924	1,036	1,150	1,086	1,146	1,079	-5.8
ciences	7,304	6,959	6,687	6,637	6,596	6,544	6,712	6,889	7,358	7,212	7,177	-0.5
nography	2,615	2,479	2,562	2,624	2,668	2,585	2,618	2,695	2,801	2,760	2,770	0.4
earth, atmospheric,												
ocean sciences	4,178	4,018	4,044	3,909	3,714	3,788	3,874	3,886	3,886	3,718	3,894	4.7
atical sciences	18,008	16,719	16,485	16,257	15,650	16,651	18,163	19,465	19,931	20,210	20,815	3.0
l sciences	32,333	31,105	30,575	30,691	30,385	31,038	32,341	34,298	35,761	36,375	36,901	1.4
nomy	874	778	820	832	888	916	990	1,080	1,119	1,191	1,211	1.7
istry	19,334	18,774	18,482	18,416	18,105	18,366	19,045	20,049	20,776	21,101	21,351	1.2
ics	11,728	11,147	10,809	10,869	10,841	11,248	11,701	12,555	13,298	13,472	13,722	1.8
physical sciences	397	406	464	574	551	508	605	614	568	611	617	1.0
ogy	53,122	53,126	52,557	51,727	50,466	50,454	51,152	52,162	54,126	57,282	57,653	0.6
ciences	88,635	86,085	84,053	84,984	83,327	85,682	90,215	95,129	97,332	99,427	102,347	2.9
ultural economics	2,117	2,043	1,995	2,014	2,079	2,161	2,187	2,318	2,195	2,127	2,158	1.5
pology	7,773	7,560	7,577	7,633	7,626	7,491	7,481	7,789	7,826	7,750	8,150	5.2
omics	12,080	11,097	10,701	10,562	10,748	11,408	12,009	12,316	12,318	11,805	12,132	2.8
raphy	4,331	4,287	4,326	4,250	4,036	4,304	4,383	4,721	4,809	4,800	4,750	-1.0
y and philosophy												
cience	409	443	508	557	532	571	663	737	994	965	968	0.3
stics	3,156	3,068	2,935	2,799	2,674	2,744	2,875	3,028	2,941	3,187	3,074	-3.5
al science	33,252	32,083	30,828	31,372	31,131	31,805	34,934	36,880	39,023	40,780	41,784	2.5
logy	9,425	9,413	9,058	8,966	8,652	8,812	8,946	9,127	8,874	9,018	9,035	0.2
logy/anthropology	923	948	857	741	745	808	719	773	839	848	837	-1.3
social sciences	15,169	15,143	15,268	16,090	15,104	15,578	16,018	17,440	17,513	18,147	19,459	7.2
	103,224	101,148	100,038	101,691	104,112	109,493	119,668	127,377	123,566	120,565	123,041	2.0
ce engineering	3,208	3,083	3,137	3,349	3,407	3,451	3,685	4,048	4,089	4,170	4,482	7.5
cal engineering	2,689	2,797	2,855	3,069	3,197	3,599	4,338	5,301	5,807	6,067	6,482	6.8
al engineering	7,408	7,288	7,093	6,883	7,056	6,913	7,414	7,516	7,452	7,173	7,261	1.2
gineering	18,528	17,193	16,517	16,226	16,451	16,665	17,713	18,890	18,561	18,114	17,802	-1.7
l engineering	29,941	30,787	31,384	31,822	33,611	36,100	39,948	41,763	38,995	37,450	38,265	2.2
l engineering	12,675	11,957	11,221	11,803	12,119	12,940	14,033	14,313	13,852	13,650	13,829	1.3
cal engineering	15,509	15,045	14,696	14,956	15,235	15,852	17,139	18,393	17,852	17,373	17,919	3.1
gical/materials												
engineering	4,747	4,688	4,680	4,481	4,377	4,721	4,992	5,131	5,059	5,160	5,268	2.1
Other engineering	8,519	8,310	8,455	9,102	8,659	9,252	10,406	12,022	11,899	11,408	11,733	2.8

SOURCE: National Science Foundation/Division of Science Resources Statistics, Survey of Graduate Students and Postdoctorates in Science and Engineering.



Please use this citation

Jacquelyn Boone James; Rosalind C. Barnett, 2003, "The Aspirations and Experiences of Undergraduates, 2000", http://hdl.handle.net/1902.1/00092 UNF:3:0sjRWertkBtT9wlgcuP4Fg== Murray Research Archive [Distributor] V1 [Version]

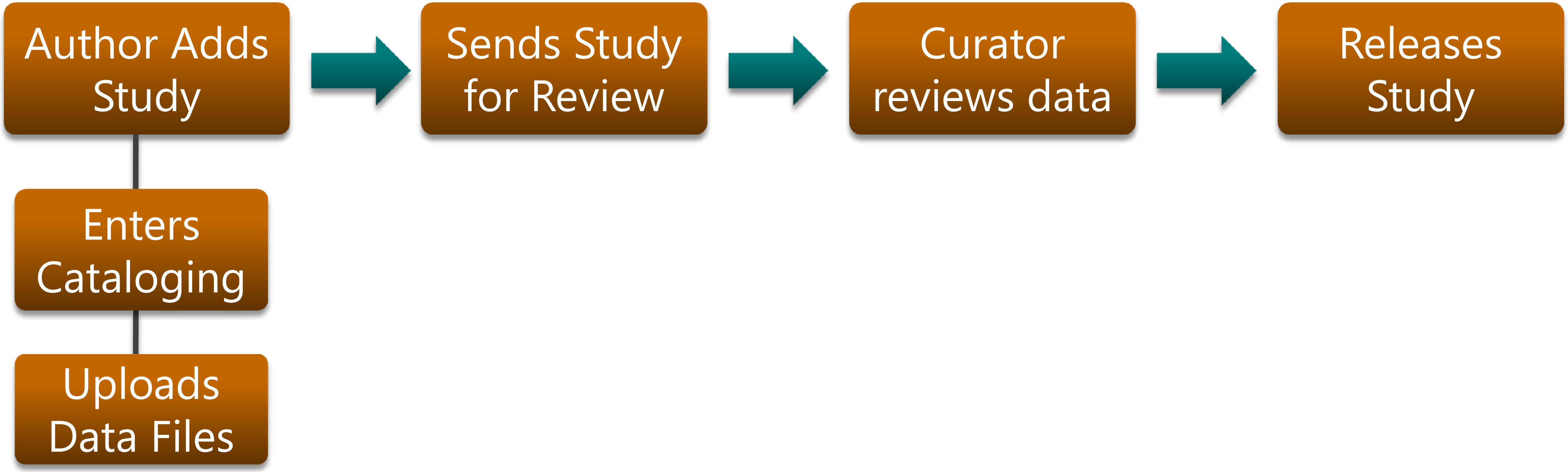
Citation For

Persistent identifier and url that never changes

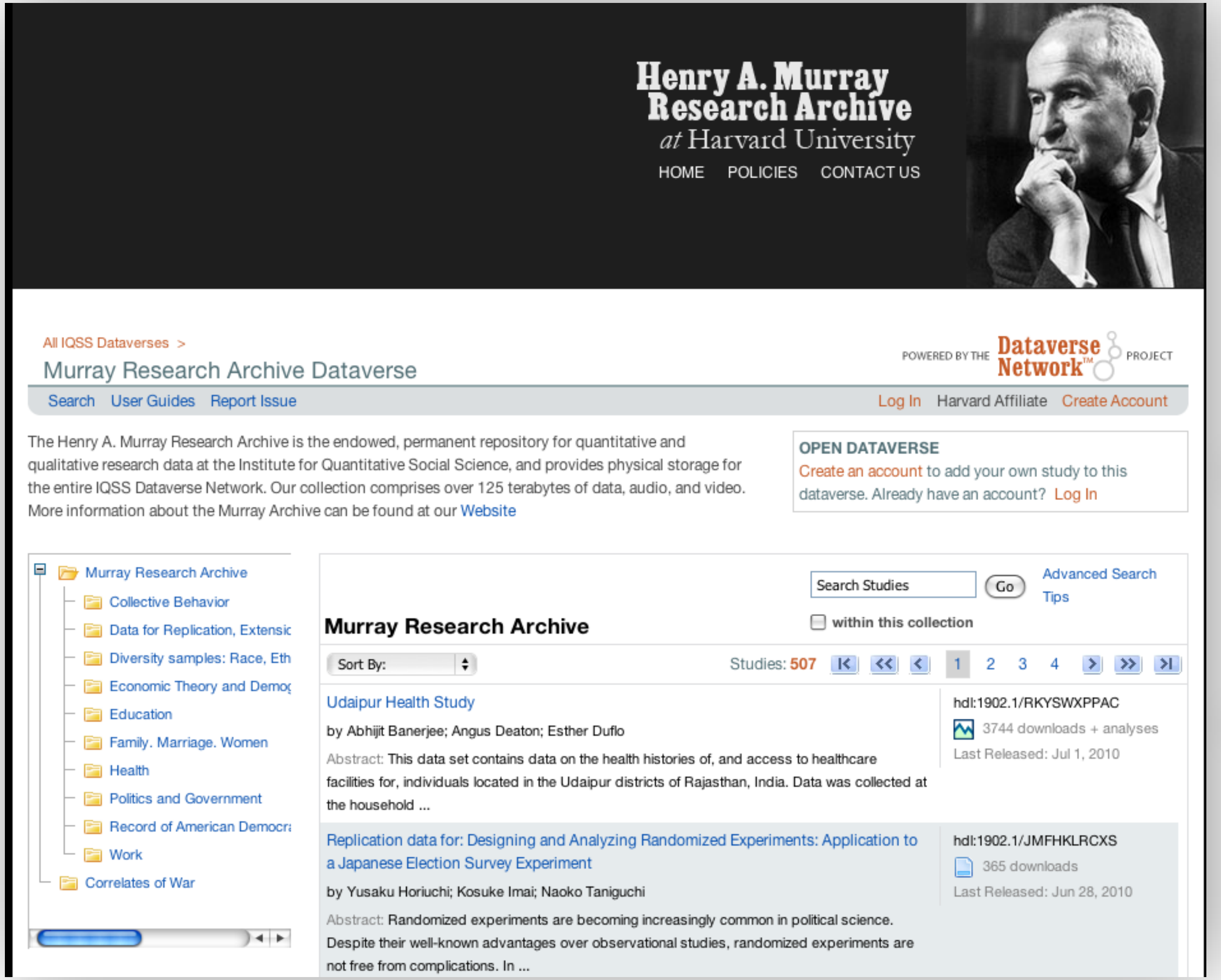
Universal Numerical Fingerprint (UNF) to verify dataset

Use Case 2: Dataverse for Journals – Submission and Review Process

- A Journal for replication data can be set as an open dataverse.
- Authors upload data after registering.
- Journal reviews data before releasing it.



Use Case 3: Dataverse for Archives – Evidence is Never Lost



Henry A. Murray Research Archive *at* Harvard University

HOME POLICIES CONTACT US

All IQSS Dataverses >

Murray Research Archive Dataverse

POWERED BY THE Dataverse Network™ PROJECT

Search User Guides Report Issue Log In Harvard Affiliate Create Account

The Henry A. Murray Research Archive is the endowed, permanent repository for quantitative and qualitative research data at the Institute for Quantitative Social Science, and provides physical storage for the entire IQSS Dataverse Network. Our collection comprises over 125 terabytes of data, audio, and video. More information about the Murray Archive can be found at our Website

OPEN DATAVERSE

Create an account to add your own study to this dataverse. Already have an account? Log In

- Murray Research Archive
  - Collective Behavior
  - Data for Replication, Extensic
  - Diversity samples: Race, Eth
  - Economic Theory and Demog
  - Education
  - Family. Marriage. Women
  - Health
  - Politics and Government
  - Record of American Democr:
  - Work
- Correlates of War

Search Studies Go Advanced Search Tips

□ within this collection

Murray Research Archive

Sort By:

Studies: 507 1 2 3 4

Udaipur Health Study

by Abhijit Banerjee; Angus Deaton; Esther Duflo

Abstract: This data set contains data on the health histories of, and access to healthcare facilities for, individuals located in the Udaipur districts of Rajasthan, India. Data was collected at the household ...

hdl:1902.1/RKYSWXPPAC

3744 downloads + analyses

Last Released: Jul 1, 2010

Replication data for: Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment

by Yusaku Horiuchi; Kosuke Imai; Naoko Taniguchi

Abstract: Randomized experiments are becoming increasingly common in political science. Despite their well-known advantages over observational studies, randomized experiments are not free from complications. In ...

hdl:1902.1/JMFHKLRCXS

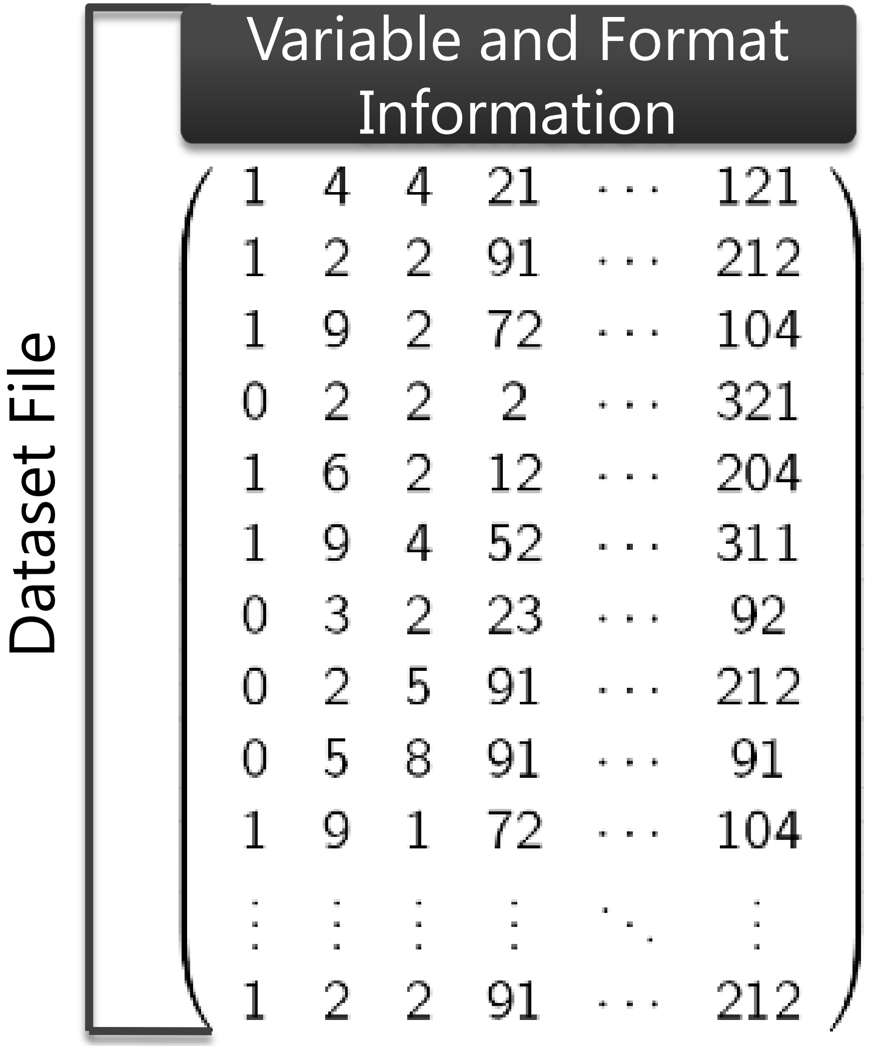
365 downloads

Last Released: Jun 28, 2010

The Dataverse provides:

- **Versioning** – Keeps all changes to data and metadata to allow citation of old versions.
- **Deaccession** – Doesn't delete studies permanently so citation is valid forever.
- **Universal Numeric Fingerprints** – Fixity that survives changes of format.

Use Case 3: Dataverse for Archives – Re-formatting for Preservation



Dataset File

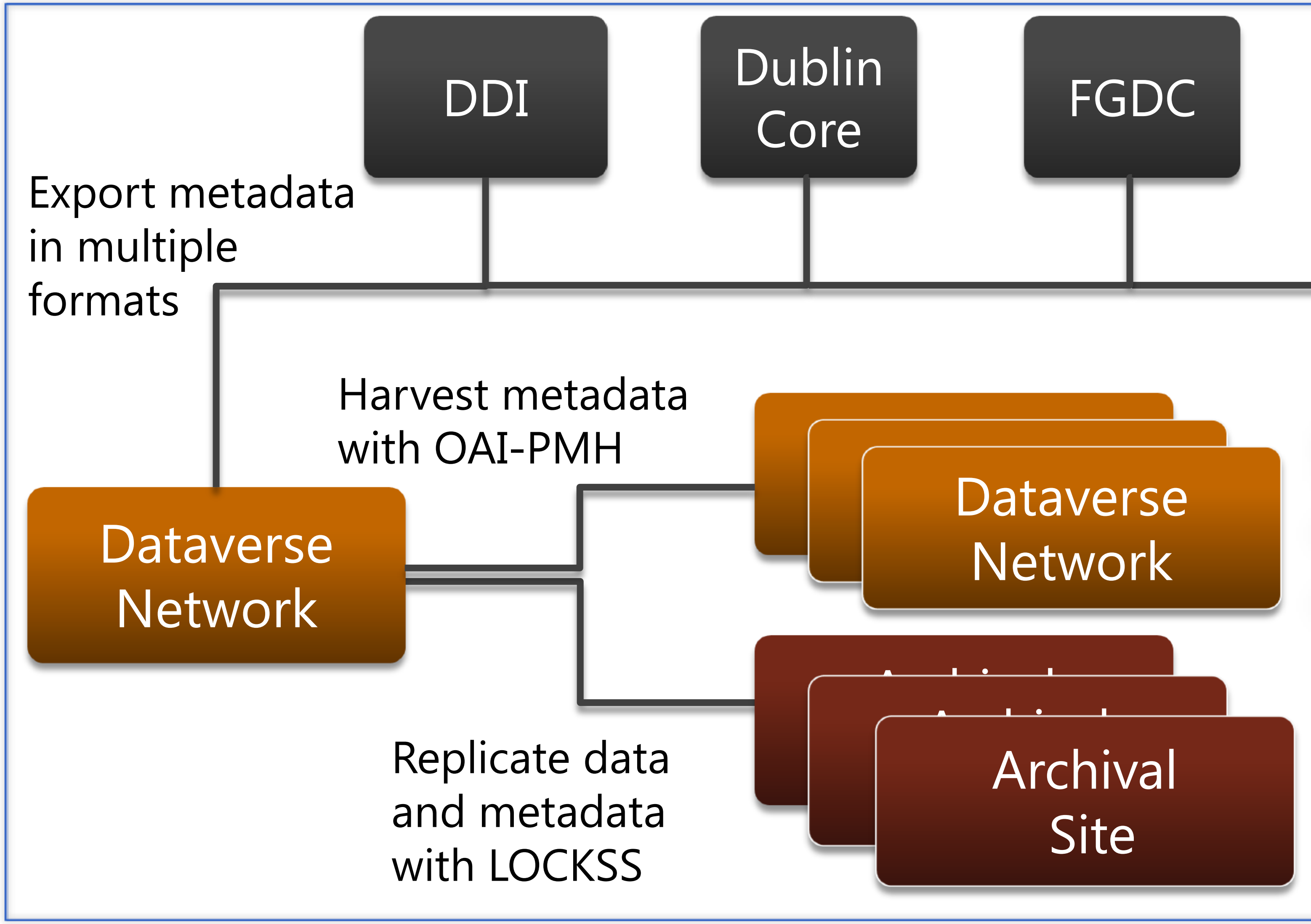
Variable and Format Information

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

Preservation and verifiable format by:

1. Extracting variable and format metadata.
2. Converting dataset to a format independent of software package.
3. Applying cryptographic algorithm to canonical format.
4. Obtaining an alphanumeric string based on semantic content: UNF5:EKgHvTNfkkS86dNzABIhNw==

Use Case 3: Dataverse for Archives – Standards and Backups

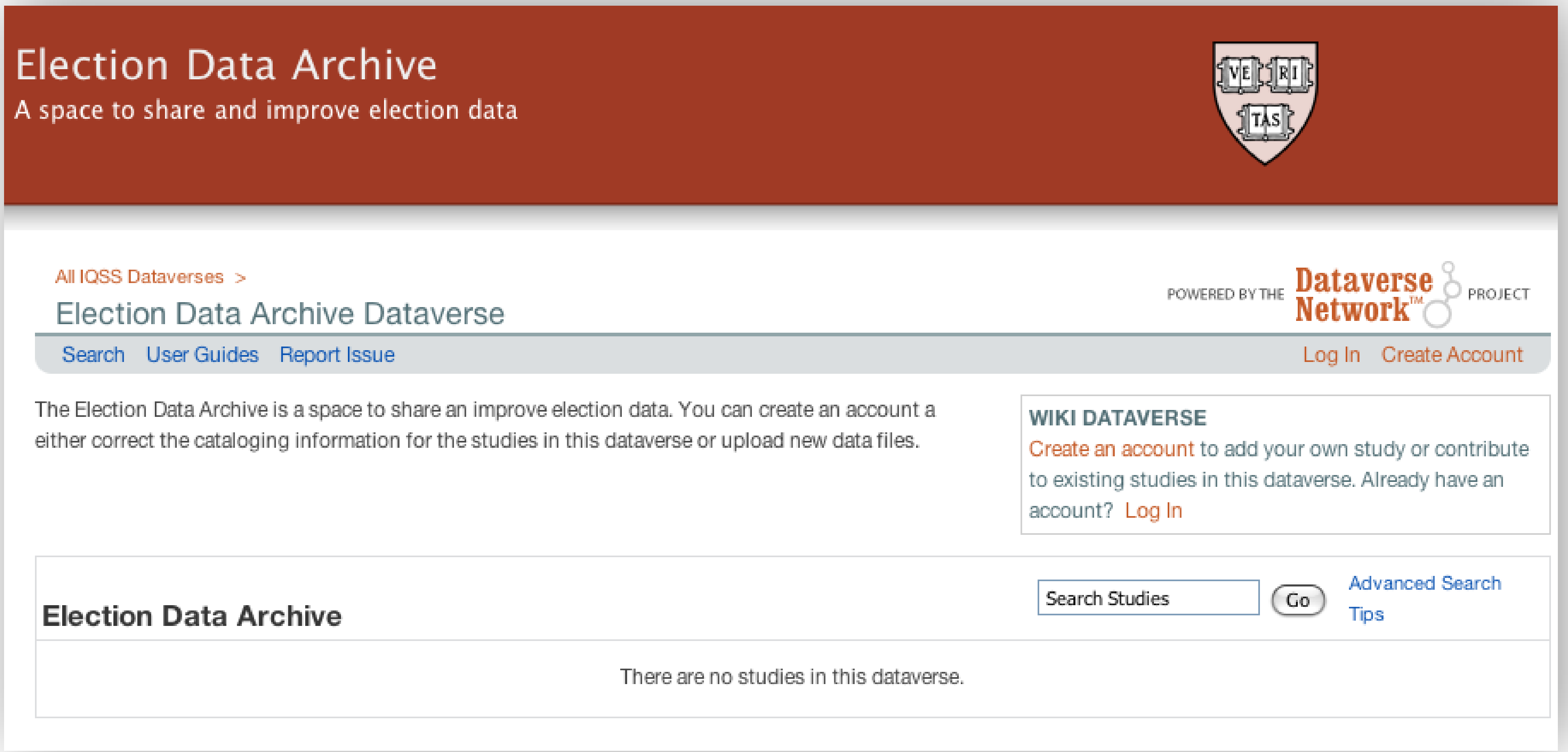

DDI
Dublin Core
FGDC
Marc
Export metadata in multiple formats
Harvest metadata with OAI-PMH
Dataverse Network
Dataverse Network
Archival Site
Replicate data and metadata with LOCKSS

The Dataverse facilitates:

- Multiple copies of data
- Metadata exchange with other systems.

Use Case 4: Wiki Dataverses – Collaborative Collection Development

New: A ‘wiki’ dataverse is an archive to share and improve data through contributions (with a moderator to review them before their release)



Election Data Archive

A space to share and improve election data

All IQSS Dataverses >

Election Data Archive Dataverse

The Election Data Archive is a space to share an improve election data. You can create an account a either correct the cataloging information for the studies in this dataverse or upload new data files.

WIKI DATAVERSE

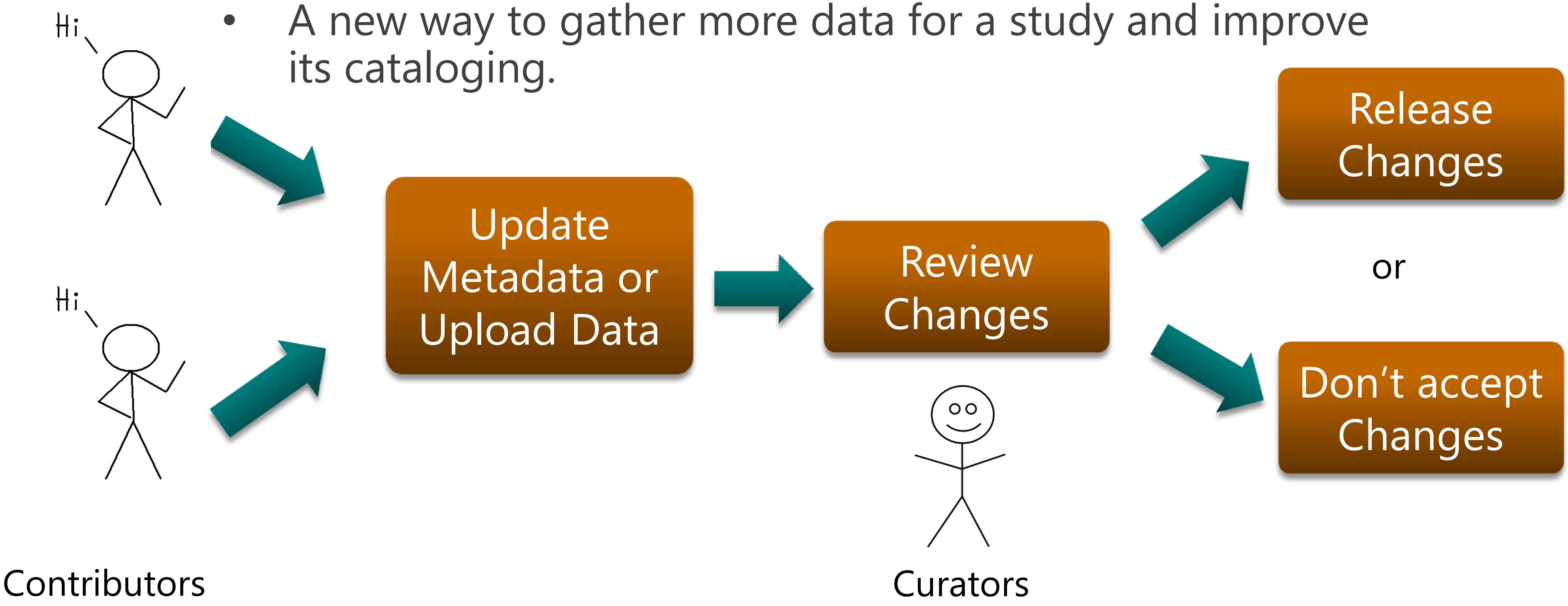
Create an account to add your own study or contribute to existing studies in this dataverse. Already have an account? Log In

Election Data Archive

There are no studies in this dataverse.

Use Case 4: Wiki Dataverses – Mulitple Roles and Workflows

- Multiple contributors can be updating the same study.
- Changes are only visible after being reviewed.
- A new way to gather more data for a study and improve its cataloging.



Update Metadata or Upload Data

Review Changes

Release Changes

or

Don’t accept Changes

Contributors

Curators

Current Usage

- The IQSS Dataverse Network at Harvard holds:
  - 257 public dataverses from around the world
  - 36,000 studies (mostly in Social Science)
  - 640,000 files
- … *and* there are over 10 more installations
  of the Dataverse Network at other Institutions.

How to Get Started …

- Create a dataverse at: http://dvn.iq.harvard.edu
- Or install the Dataverse Network software for your Institution.
- For more information, go to: http://thedata.org
- Contact me at: mcrosas@hmdc.harvard.edu

THANK YOU!

Microsoft®

The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.

MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Microsoft® Research

Faculty Summit 2010