

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



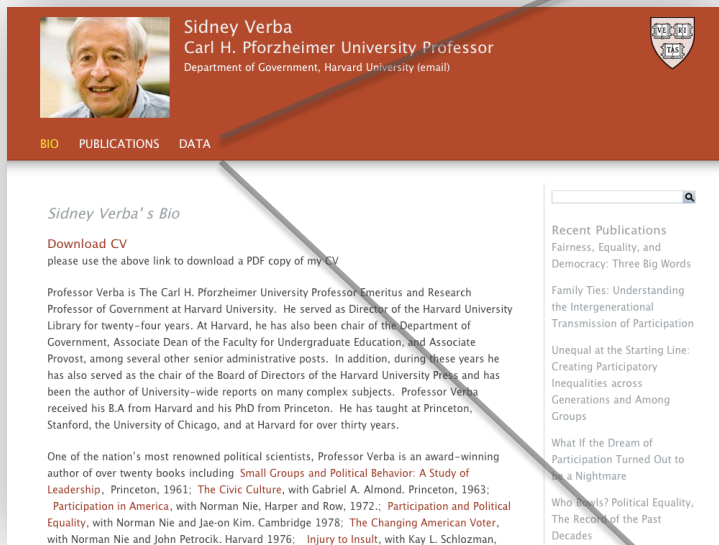
Distributed Recognition and Control

- 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

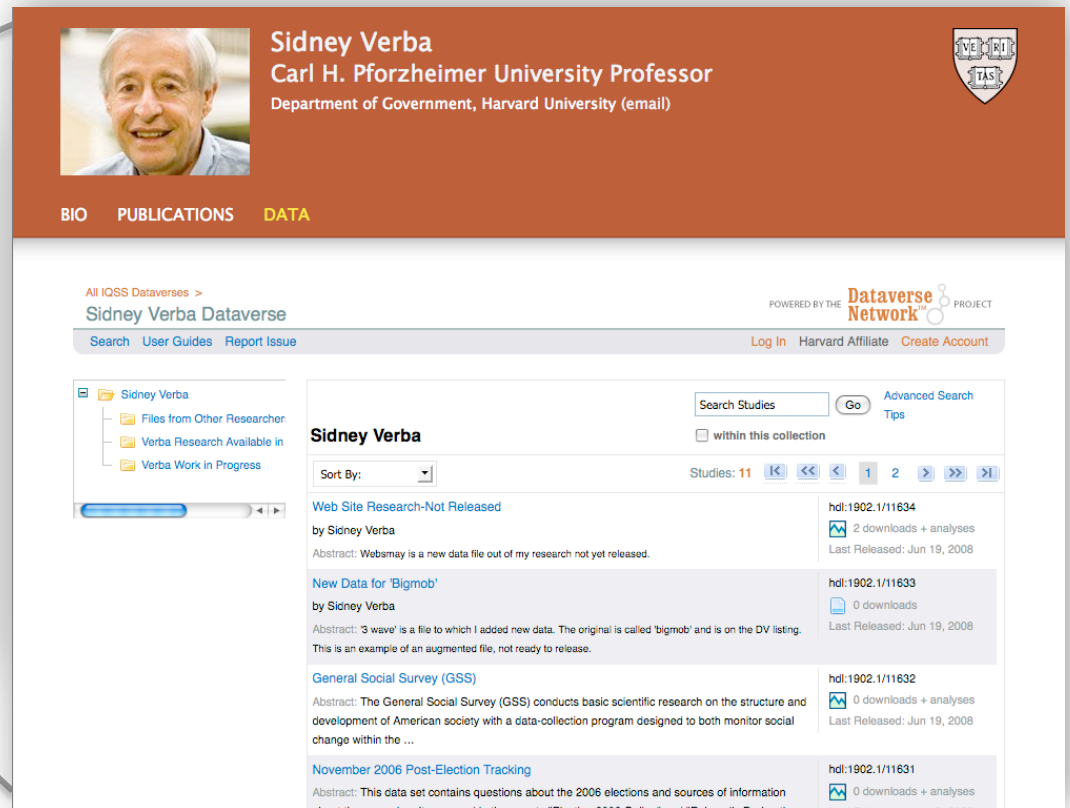
Family Ties: Understanding the Intergenerational Transmission of Participation

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 Votes? Political Equality, The Records 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

Search Studies Go Advanced Search Tips

within this collection

Sort By: Studies: 11

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 about the campaign. It was used in the reports "Election 2006: Critical" and "Debate's Duels the ...
hdl:1902.1/11631
0 downloads + analyses
Last Released: Jun 19, 2008

Your dataverse

Use Case 1: Dataverse for Researchers – Recognition through Citation

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

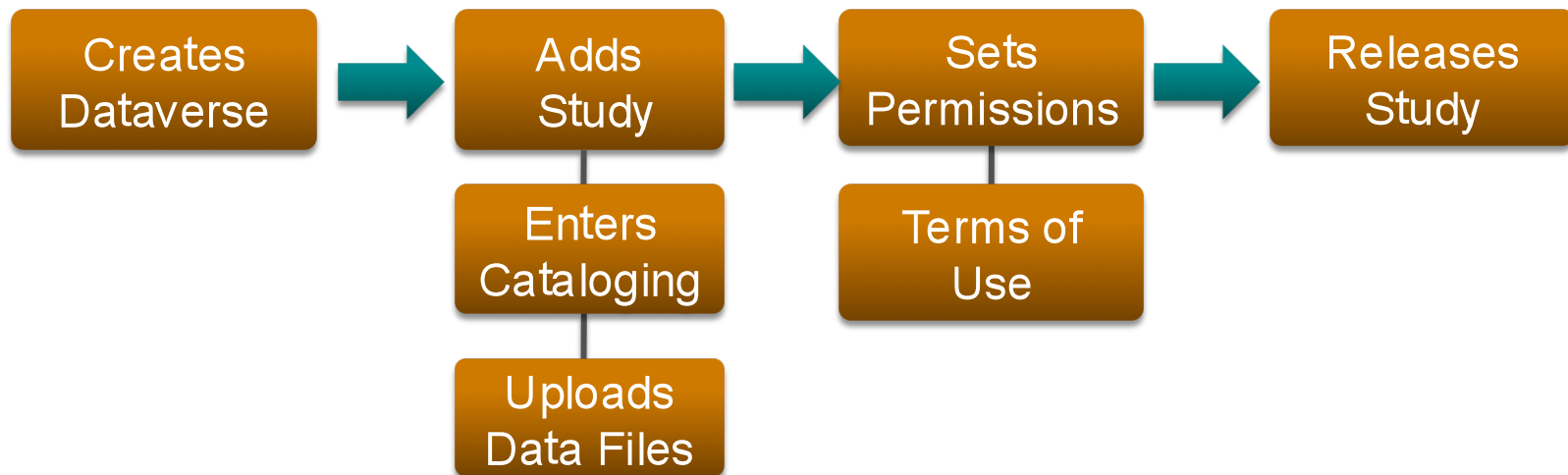
Study Global Id	hdl:1902.2/7768
Other ID	ICPSR: 7768
Authors	Verba, Sidney (Unknown); Nie, Norman H. (Unknown); Kim, Jae-On (Unknown)
Production Date	Please see full citation.
Funding Agency	National Science Foundation, Ford Foundation, and Carnegie Corporation.
Distributor	Inter-university Consortium 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

The screenshot displays the NYU Law Review Dataverse interface. On the left is a navigation menu for NYU LAW, including links for ABOUT, ADMISSIONS, ACADEMICS, and FACULTY, as well as a sidebar with links like Law Review, Masthead, Membership, Issues, Data Repository, Submissions, Copyright Policy, and Contact Us.

The main content area is titled "NYU Law Review Dataverse" and includes a search bar and navigation links. It features a list of articles with abstracts, such as "Paying-to-Play in Securities Class Actions: A..." by Drew Johnson-Skinner and "The Constitutional Life of Legislative Instruct..." by Christopher Terranova.

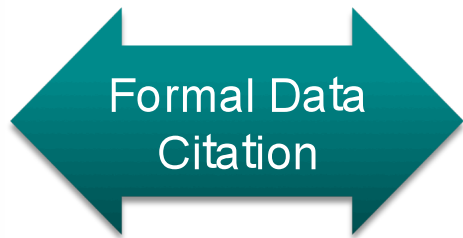
On the right, the "Journal of Information Technology & Politics Dataverse" is shown, powered by the Dataverse Network Project. It includes a search bar, a "Log In" button, and a section for "OPEN DATVERSE" with a link to "Create an account".

Below the search bar, there is a "JITP" section with a "Sort By:" dropdown and a "Search Studies" button. A list of studies is displayed, including "Replication data for: Political Science Computing: A Review of Trends in Computer Evolution and Political Science Research" by Euel W. Elliott, Karl Ho, and Jennifer S. Holmes, and "Replication data for: SES-0322662 'Democracy and E-Rulemaking: Comparing Traditional vs. Electronic Comment from a Discursive Democratic Framework'" by David Schlosberg, Stephen Zavestoski, and Stuart Shulman.

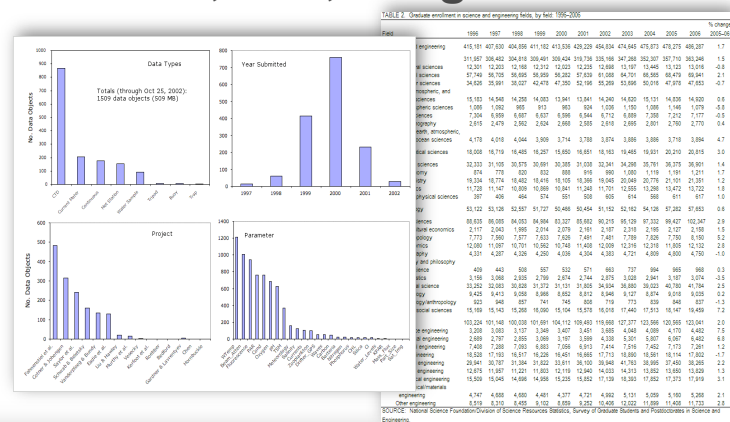
Each study entry includes its title, authors, abstract, and metadata such as the hdl number, number of downloads, and the last release date.

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

Journal, Book



Datasets, text, images



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:3:0sJRWertkBT9wlgcuP4Fg== Murray Research Archive [Distributor] V1 [Version]

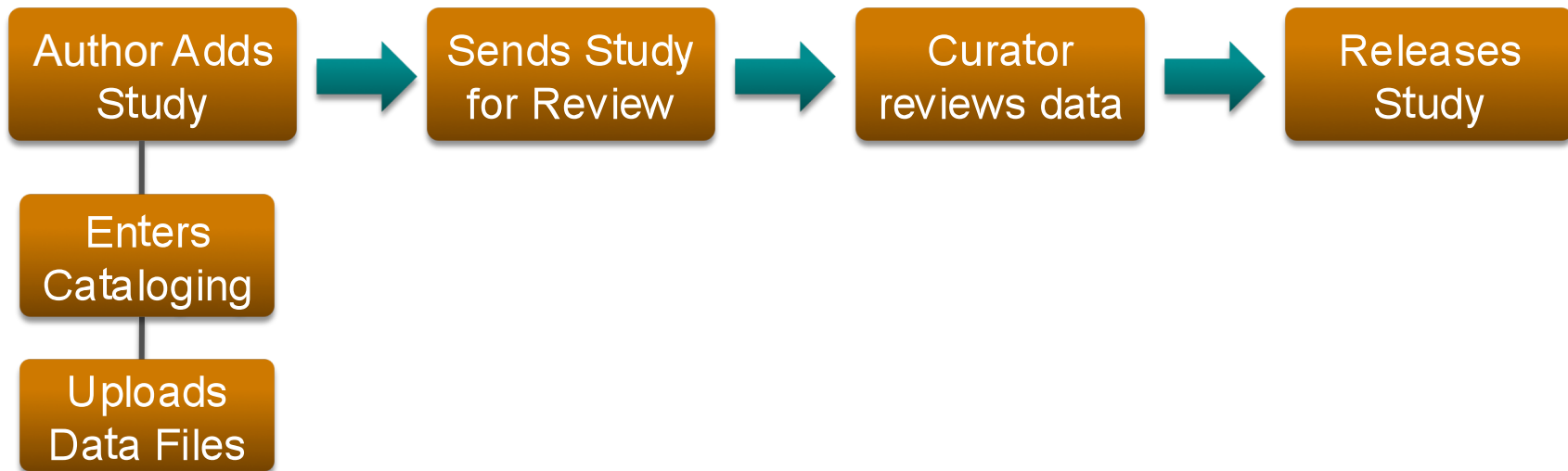
Citation For

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

The screenshot shows the Henry A. Murray Research Archive website. At the top, it features the name 'Henry A. Murray Research Archive at Harvard University' and navigation links for 'HOME', 'POLICIES', and 'CONTACT US'. Below this, there's a banner for 'Murray Research Archive Dataverse' powered by the 'Dataverse Network PROJECT'. The main content area includes a search bar, a 'Log In' button, and a 'Create Account' button. A paragraph describes the archive as an endowed, permanent repository for quantitative and qualitative research data. A sidebar on the left lists various research categories like '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', and 'Correlates of War'. The main content area displays a list of studies, including 'Udaipur Health Study' and 'Replication data for: Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment', each with its author, abstract, and download statistics.

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

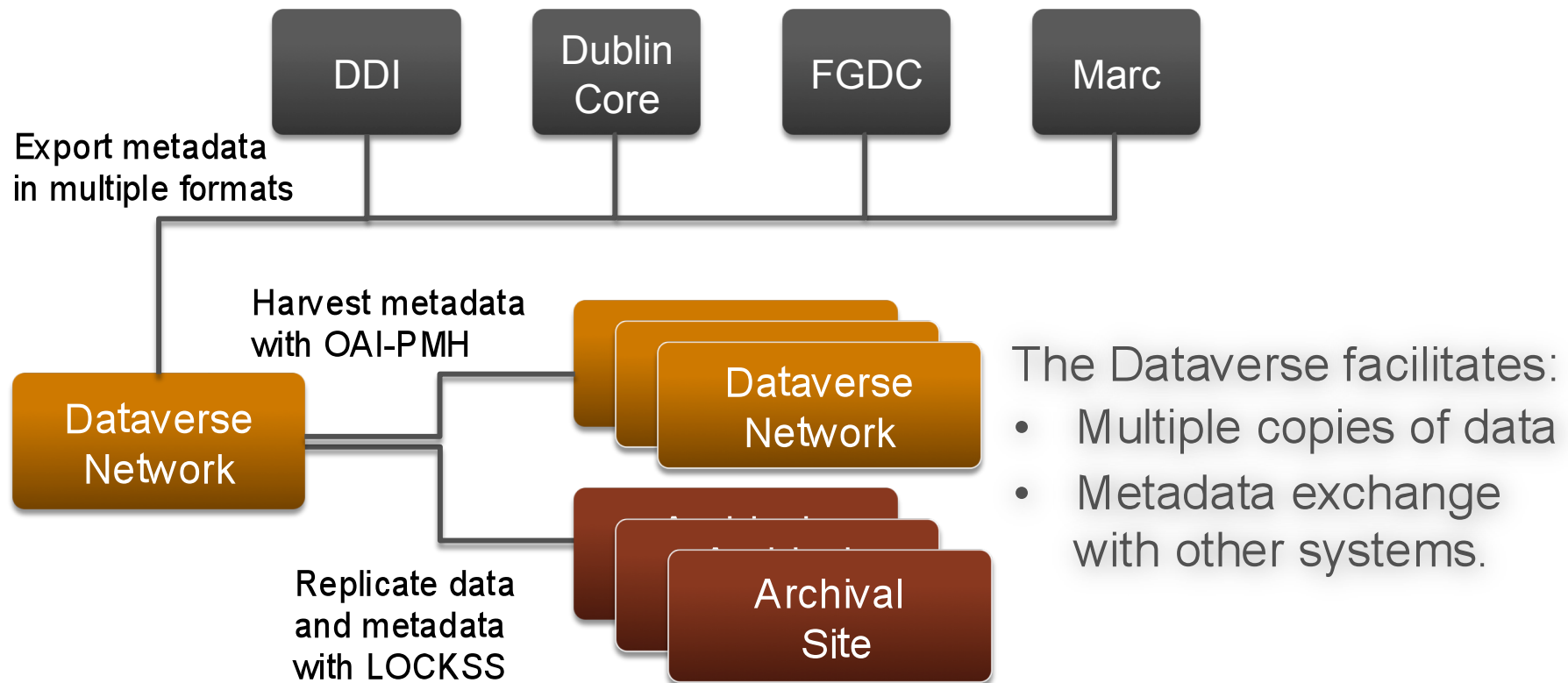
Variable and Format Information					
1	4	4	21	...	121
1	2	2	91	...	212
1	9	2	72	...	104
0	2	2	2	...	321
1	6	2	12	...	204
1	9	4	52	...	311
0	3	2	23	...	92
0	2	5	91	...	212
0	5	8	91	...	91
1	9	1	72	...	104
⋮	⋮	⋮	⋮	⋮	⋮
1	2	2	91	...	212

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:EKgHvTNfkkS86dNzABlhNw==

Use Case 3: Dataverse for Archives – Standards and Backups



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)

The screenshot shows the web interface for the Election Data Archive. At the top, there is a dark red header with the text "Election Data Archive" and "A space to share and improve election data" on the left, and a crest logo on the right. Below the header, the page title "Election Data Archive Dataverse" is displayed, along with navigation links for "Search", "User Guides", and "Report Issue". On the right side, there is a "Log In" and "Create Account" link. A central text block explains the purpose of the archive: "The Election Data Archive is a space to share and 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." To the right of this text is a "WIKI DATAVERSE" section with the text: "Create an account to add your own study or contribute to existing studies in this dataverse. Already have an account? Log In". At the bottom, there is a search bar with the text "Search Studies", a "Go" button, and links for "Advanced Search" and "Tips". Below the search bar, a message states: "There are no studies in this dataverse."

Election Data Archive
A space to share and improve election data

All IQSS Dataverses >
Election Data Archive Dataverse

POWERED BY THE **Dataverse Network**™ PROJECT

Search User Guides Report Issue Log In Create Account

The Election Data Archive is a space to share and 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.

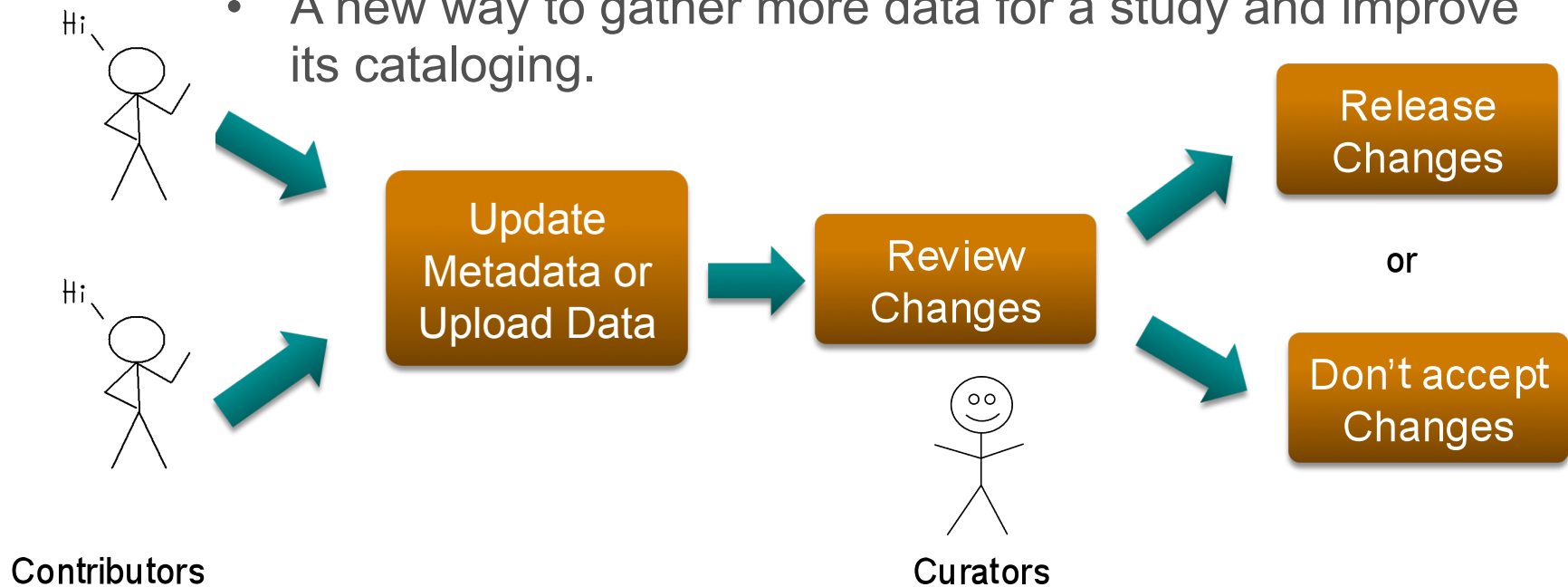
WIKI DATAVERSE
Create an account to add your own study or contribute to existing studies in this dataverse. Already have an account? Log In

Election Data Archive Search Studies Go Advanced Search Tips

There are no studies in this dataverse.

Use Case 4: Wiki Dataverses – Multiple 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.



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!