Finding, Analyzing, Disseminating, and Preserving Numeric Data

Gary King
Harvard University

Joint work with Micah Altman and Sidney Verba
What is Numeric Data For?

- Ready reference: What is the percent of women 18-24 who voted for Clinton in Massachusetts?
- Secondary analysis: Using data for purposes not originally envisioned
- Replication: validation & extension of scientific results
- Dissemination and Preservation: important for science, often a requirement of grants and journals
What is Numeric Data For?

- **Ready reference:** What is the percent of women 18-24 who voted for Clinton in Massachusetts?
What is Numeric Data For?

- **Ready reference:** What is the percent of women 18-24 who voted for Clinton in Massachusetts?
- **Secondary analysis:** Using data for purposes not originally envisioned
What is Numeric Data For?

- **Ready reference:** What is the percent of women 18-24 who voted for Clinton in Massachusetts?
- **Secondary analysis:** Using data for purposes not originally envisioned
- **Replication:** validation & extension of scientific results
What is Numeric Data For?

- **Ready reference**: What is the percent of women 18-24 who voted for Clinton in Massachusetts?
- **Secondary analysis**: Using data for purposes not originally envisioned
- **Replication**: validation & extension of scientific results
- **Dissemination and Preservation**: important for science, often a requirement of grants and journals
What this talk is about

Protocols for citing numeric data

Protocols for sharing, finding, and preserving data

Easy ways to query and analyze data

Automation of some library tasks

Virtual Data Center software that implements these protocols and runs at Harvard and MIT

Making the same software available for others

Federating with at other sites
What this talk is about

- Protocols for citing numeric data
What this talk is about

- Protocols for citing numeric data
- Protocols for sharing, finding, and preserving data
What this talk is about

- Protocols for citing numeric data
- Protocols for sharing, finding, and preserving data
- Easy ways to query and analyze data
What this talk is about

- Protocols for citing numeric data
- Protocols for sharing, finding, and preserving data
- Easy ways to query and analyze data
- Automation of some library tasks
What this talk is about

- Protocols for citing numeric data
- Protocols for sharing, finding, and preserving data
- Easy ways to query and analyze data
- Automation of some library tasks
- Virtual Data Center software that implements these protocols and runs at Harvard and MIT
What this talk is about

- Protocols for citing numeric data
- Protocols for sharing, finding, and preserving data
- Easy ways to query and analyze data
- Automation of some library tasks
- Virtual Data Center software that implements these protocols and runs at Harvard and MIT
- Making the same software available for others
What this talk is about

- Protocols for citing numeric data
- Protocols for sharing, finding, and preserving data
- Easy ways to query and analyze data
- Automation of some library tasks
- **Virtual Data Center** software that implements these protocols and runs at Harvard and MIT
- Making the same software available for others
- Federating with at other sites
Rules for Citing Printed Matter


Gary King ()
Numeric Data

World’s most important social scientist

Journal (no longer exists)

Special formatting codes

Citations: rule-based, precise, redundant
Lack of Rules for Citing Numeric Data

No fixed rules for copyeditors
No consistency in practice
Sometimes in the list of references; sometimes just a casual mention in the text
Sometimes the archive is noted
Sometimes a version number exists
Sometimes the version number is listed (if it exists)
Archive numbers are sometimes given, if they exist
Sometimes the author is noted
Date of creation is sometimes given
URLs often given, rarely persist
Dates of access: protect the researcher, do not help find the data
The data may not be available publicly
The data may no longer exist
The data may not have ever been held by anyone but the investigator
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists

- URLs often given, rarely persist
- Dates of access: protect the researcher, do not help find the data
- The data may not be available publicly
- The data may no longer exist
- The data may not have ever been held by anyone but the investigator
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
- Sometimes the author is noted
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
- Sometimes the author is noted
- Date of creation is sometimes given
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
- Sometimes the author is noted
- Date of creation is sometimes given
- URLs often given, rarely persist
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
- Sometimes the author is noted
- Date of creation is sometimes given
- URLs often given, rarely persist
- Dates of access: protect the researcher, do not help find the data
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
- Sometimes the author is noted
- Date of creation is sometimes given
- URLs often given, rarely persist
- Dates of access: protect the researcher, do not help find the data
- The data may not be available publicly
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
- Sometimes the author is noted
- Date of creation is sometimes given
- URLs often given, rarely persist
- Dates of access: protect the researcher, do not help find the data
- The data may not be available publicly
- The data may no longer exist
Lack of Rules for Citing Numeric Data

- No fixed rules for copyeditors
- No consistency in practice
- Sometimes in the list of references; sometimes just a casual mention in the text
- Sometimes the archive is noted
- Sometimes a version number exists
- Sometimes the version number is listed (if it exists)
- Archive numbers are sometimes given, if they exist
- Sometimes the author is noted
- Date of creation is sometimes given
- URLs often given, rarely persist
- Dates of access: protect the researcher, do not help find the data
- The data may not be available publicly
- The data may no longer exist
- The data may not have ever been held by anyone but the investigator
Lack of Rules for Preserving Data

A major archive renumbered all its acquisitions, but the same data distributed by different archives have different identifiers. Publishers sometimes withdraw data from some archives, but it remains in others, making study numbers rendered invalid or ambiguous. When a dataset is expanded, the old study number is sometimes "deaccessioned" and a new one assigned. (Data remains available, but citation is invalid.) Researchers sometimes distribute modified (or corrected) versions of data as in archives, using the same identifiers. Changes to datasets are made and the existing identifier is "reused," with old data being lost.
A major archive renumbered all its acquisitions.
Lack of Rules for Preserving Data

- A major archive renumbered all its acquisitions
- The same data distributed by different archives have different identifiers

Publishers sometimes withdraw data from some archives, but it remains in others. Study numbers rendered invalid or ambiguous. When a dataset is expanded, the old study number is sometimes "deaccessioned" and a new one assigned. (Data remains available, but citation is invalid.) Researchers sometimes distribute modified (or corrected) versions of data as in archives, using the same identifiers. Changes to datasets are made and existing identifier is "reused"; old data lost.
Lack of Rules for Preserving Data

- A major archive renumbered all its acquisitions
- The same data distributed by different archives have different identifiers
- Publishers sometimes withdraw data from some archives, but it remains in others. Study numbers rendered invalid or ambiguous.

Researchers sometimes distribute modified (or corrected) versions of data as in archives, using the same identifiers.

Changes to datasets are made and existing identifier is "reused"; old data lost.
Lack of Rules for Preserving Data

- A major archive renumbered all its acquisitions
- The same data distributed by different archives have different identifiers
- Publishers sometimes withdraw data from some archives, but it remains in others. Study numbers rendered invalid or ambiguous.
- When a dataset is expanded, the old study number is sometimes “deaccessioned” and a new one assigned. (Data remains available, but citation is invalid.)
A major archive renumbered all its acquisitions

The same data distributed by different archives have different identifiers

Publishers sometimes withdraw data from some archives, but it remains in others. Study numbers rendered invalid or ambiguous.

When a dataset is expanded, the old study number is sometimes “deaccessioned” and a new one assigned. (Data remains available, but citation is invalid.)

Researchers sometimes distribute modified (or corrected) versions of data as in archives, using the same identifiers.
Lack of Rules for Preserving Data

- A major archive renumbered all its acquisitions
- The same data distributed by different archives have different identifiers
- Publishers sometimes withdraw data from some archives, but it remains in others. Study numbers rendered invalid or ambiguous.
- When a dataset is expanded, the old study number is sometimes “deaccessioned” and a new one assigned. (Data remains available, but citation is invalid.)
- Researchers sometimes distribute modified (or corrected) versions of data as in archives, using the same identifiers.
- Changes to datasets are made and existing identifier is “reused”; old data lost.

1. Author
2. Year
3. Title
4. VDC Unique Global Identifier (handle)

1. Author
2. Year
3. Title
4. VDC Unique Global Identifier (handle)
5. Permanent Universal Resource Locator (PURL)

1. Author
2. Year
3. Title
4. VDC Unique Global Identifier (handle)
5. *Permanent* Universal Resource Locator (PURL)
6. Universal Numeric Fingerprint (UNF)
Same UNF regardless of hardware, operating system, statistical software, database, or spreadsheet software.
Same UNF regardless of hardware, operating system, statistical software, database, or spreadsheet software.

![Matrix](image)

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

\[\Rightarrow \]

Gary King ()

Numeric Data
Data to Universal Numeric Fingerprints

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

\[\Rightarrow \quad \text{ZNQRI14053UZq389x0Bffg?==}\]
Data to Universal Numeric Fingerprints

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

\[\Rightarrow \text{ZNQRI14053UZq389x0Bffg?==} \]

Same UNF regardless of hardware,
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{ZNQRI14053UZq389x0Bffg?==}

Same UNF regardless of hardware, operating system, statistical software, database, or spreadsheet software.
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}
\Rightarrow \text{ZNQRI14053UZq389xBffg?==}

Same UNF regardless of hardware, operating system, statistical software,
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}
\rightarrow \quad ZNQRI14053UZq389xBffg?==

Same UNF regardless of hardware, operating system, statistical software, database,
Data to Universal Numeric Fingerprints

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

Same UNF regardless of hardware, operating system, statistical software, database, or spreadsheet software.
Replication Problems: Solved by Our Citation Standard

Science is not (only) about being scientific. Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals. Of what value is an article with claims that cannot be replicated? Scholarly articles are summaries, not the actual research results. The real research is the data and methods used. But: replication data often not available. More journals now requiring data submission with article. Finding the data is still hard. Hard for journal editors to verify. Even if you find it, how do you know it is the same? Class replication projects: most published articles cannot be replicated.
• Science is not (only) about being scientific
Science is not (only) about being scientific

Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals
Science is not (only) about being scientific

Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals

Of what value is an article with claims that cannot be replicated?
Science is not (only) about being scientific

Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals

Of what value is an article with claims that cannot be replicated?

Scholarly articles are summaries, not the actual research results.
Science is not (only) about being scientific

Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals

Of what value is an article with claims that cannot be replicated?

Scholarly articles are summaries, not the actual research results.

The real research is the data and methods used
Science is not (only) about being scientific

Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals

Of what value is an article with claims that cannot be replicated?

Scholarly articles are summaries, not the actual research results.

The real research is the data and methods used

But: replication data often not available
Science is not (only) about being scientific

Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals

Of what value is an article with claims that cannot be replicated?

Scholarly articles are summaries, not the actual research results.

The real research is the data and methods used

But: replication data often not available

More journals now requiring data submission with article
Science is not (only) about being scientific
Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals
Of what value is an article with claims that cannot be replicated?
Scholarly articles are summaries, not the actual research results.
The real research is the data and methods used
But: replication data often not available
More journals now requiring data submission with article
Finding the data is still hard
Science is not (only) about being scientific
Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals
Of what value is an article with claims that cannot be replicated?
Scholarly articles are summaries, not the actual research results.
The real research is the data and methods used
But: replication data often not available
More journals now requiring data submission with article
Finding the data is still hard
Hard for journal editors to verify
Science is not (only) about being scientific

Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals

Of what value is an article with claims that cannot be replicated?

Scholarly articles are summaries, not the actual research results.

The real research is the data and methods used

But: replication data often not available

More journals now requiring data submission with article

Finding the data is still hard

Hard for journal editors to verify

Even if you find it, how do you know it is the same?
Science is not (only) about being scientific
Scientific progress requires community: scholars competing and cooperating in the pursuit of the same goals
Of what value is an article with claims that cannot be replicated?
Scholarly articles are summaries, not the actual research results.
The real research is the data and methods used
But: replication data often not available
More journals now requiring data submission with article
Finding the data is still hard
Hard for journal editors to verify
Even if you find it, how do you know it is the same?
Class replication projects: most published articles cannot be replicated
The Data Center When We Came to Harvard

Give me my data!!!!
The Data Center When We Came to Harvard

Give me my data!!!!
The Data Center Today

The VDC has automated most previously uninteresting activities. It's more fun to work at HMDC. We're now a research organization, in part the R&D arm of the Harvard libraries.
The VDC has automated most previously uninteresting activities
The Data Center Today

- The VDC has automated most previously uninteresting activities
- Its more fun to work at HMDC
The VDC has automated most previously uninteresting activities

- It's more fun to work at HMDC

- We’re now a research organization, in part the R&D arm of the Harvard libraries
Who the VDC Serves

- Used in production for data delivery to Harvard and MIT
- Provides virtual access to local and remote data collections
- Disseminates Murray Research Archive collection
- Can now be installed at other sites at Harvard and around the world
used in production for data delivery to Harvard and MIT
Who the VDC Serves

- used in production for data delivery to Harvard and MIT
- 1000s of users annually, from every Harvard school
Who the VDC Serves

- used in production for data delivery to Harvard and MIT
- 1000s of users annually, from every Harvard school
- 10,000s of quantitative studies available through system
Who the VDC Serves

- used in production for data delivery to Harvard and MIT
- 1000s of users annually, from every Harvard school
- 10,000s of quantitative studies available through system
- Provides virtual access to local and remote data collections
Who the VDC Serves

- used in production for data delivery to Harvard and MIT
- 1000s of users annually, from every Harvard school
- 10,000s of quantitative studies available through system
- Provides virtual access to local and remote data collections
- Disseminates Murray Research Archive collection
Who the VDC Serves

- used in production for data delivery to Harvard and MIT
- 1000s of users annually, from every Harvard school
- 10,000s of quantitative studies available through system
- Provides virtual access to local and remote data collections
- Disseminates Murray Research Archive collection
- Can now be installed at other sites at Harvard and around the world
What the VDC Does

- Study Preparation (ingest; conversion of data and documentation formats; catalog record creation)
- User Interfaces (data users, data producers, data archive administrators, data curators, librarians)
- Study Management (file-format independent storage, archival formatting, cataloging)
- Metadata Search and Harvesting (DC, MARC and DDI metadata import and export; OpenArchives and Z39.50 protocol gateways)
- Dissemination (download packaging, format conversion, subset selection and generation)
What the VDC Does

- **Study Preparation** (ingest; conversion of data and documentation formats; catalog record creation)
What the VDC Does

- **Study Preparation** (ingest; conversion of data and documentation formats; catalog record creation)
- **User Interfaces** (data users, data producers, data archive administrators, data curators, librarians)
What the VDC Does

- **Study Preparation** (ingest; conversion of data and documentation formats; catalog record creation)
- **User Interfaces** (data users, data producers, data archive administrators, data curators, librarians)
- **Study Management** (file-format independent storage, archival formatting, cataloging)
What the VDC Does

- **Study Preparation** (ingest; conversion of data and documentation formats; catalog record creation)
- **User Interfaces** (data users, data producers, data archive administrators, data curators, librarians)
- **Study Management** (file-format independent storage, archival formatting, cataloging)
- **Metadata Search and Harvesting** (DC, MARC and DDI metadata import and export; OpenArchives and Z39.50 protocol gateways)
What the VDC Does

- **Study Preparation** (ingest; conversion of data and documentation formats; catalog record creation)
- **User Interfaces** (data users, data producers, data archive administrators, data curators, librarians)
- **Study Management** (file-format independent storage, archival formatting, cataloging)
- **Metadata Search and Harvesting** (DC, MARC and DDI metadata import and export; OpenArchives and Z39.50 protocol gateways)
- **Dissemination** (download packaging, format conversion, subset selection and generation).
What the VDC also Does

On-line analysis (Large array of statistical procedures available; interface with R and Zelig)

Distribution and Federation (federated searching and browsing, distributed virtual collections, metadata harvesting, repository caching, and federated authentication and authorization)

Replication and Citation (creation and management of persistent identifiers for datasets, UNF (universal numeric fingerprints) generation, replication code generation for analyses)

Preservation (preservation formatting, preservation metadata)
What the VDC also Does

- **On-line analysis** (Large array of statistical procedures available; interface with R and Zelig)
What the VDC also Does

- **On-line analysis** (Large array of statistical procedures available; interface with R and Zelig)
- **Distribution and Federation** (federated searching and browsing, distributed virtual collections, metadata harvesting, repository caching, and federated authentication and authorization)
What the VDC also Does

- **On-line analysis** (Large array of statistical procedures available; interface with R and Zelig)
- **Distribution and Federation** (federated searching and browsing, distributed virtual collections, metadata harvesting, repository caching, and federated authentication and authorization)
- **Replication and Citation** (creation and management of persistent identifiers for datasets, UNF (universal numeric fingerprints) generation, replication code generation for analyses)
What the VDC also Does

- **On-line analysis** (Large array of statistical procedures available; interface with R and Zelig)
- **Distribution and Federation** (federated searching and browsing, distributed virtual collections, metadata harvesting, repository caching, and federated authentication and authorization)
- **Replication and Citation** (creation and management of persistent identifiers for datasets, UNF (universal numeric fingerprints) generation, replication code generation for analyses)
- **Preservation** (preservation formatting, preservation metadata)
Development Principles

- Web-based, light client for users, administrators, curators
- Follows Open Source Standards
  - Search/Harvest: OAI, Z39.50
  - Metadata: DC, Marc, DDI
  - Identifiers: URN, Handles
- Built with off-the-shelf components
  - E.g.: Apache web server, OpenLDAP, R, Zelig, PostgresSQL
- Integration: Perl, Java Servlets, XSL/XML
- Open-Source
  - Source code is included
  - You own the program; if you don't like what we do, you can go in a different direction
- Modifiable & Redistributable
  - Does not restrict use of commercial data services
- Completely distributed
  - Simple components-based architecture
  - Any component can be on any computer hardware
- Distributed catalog: harvesting, distributed search
- Distributed data: proxying, caching, replication
Development Principles

- **Web-based**, light client for users, administrators, curators
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgreSQL Integration: Perl, Java Servlets, XSL/XML
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgreSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**

  Source code is included
  You own the program
  If you don't like what we do, you can go in a different direction

  Modifiable & Redistributable
  Does not restrict use of commercial data services

  Completely distributed
  Simple components-based architecture
  Any component can be on any computer hardware
  Distributed catalog: harvesting, distributed search
  Distributed data: proxying, caching, replication
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgreSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgreSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
  - You own the program; if you don’t like what we do, you can go in a different direction
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgresSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
  - **You own the program**; if you don’t like what we do, you can go in a different direction
  - Modifiable & Redistributable
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the-shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgreSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
  - **You own the program**; if you don’t like what we do, you can go in a different direction
  - Modifiable & Redistributable
  - Does not restrict use of commercial data services
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgresSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
  - **You own the program**; if you don’t like what we do, you can go in a different direction
  - Modifiable & Redistributable
  - Does not restrict use of commercial data services
- **Completely distributed**
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgresSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
  - **You own the program**; if you don’t like what we do, you can go in a different direction
  - Modifiable & Redistributable
  - Does not restrict use of commercial data services
- **Completely distributed**
  - Simple components-based architecture
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the-shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgresSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
  - **You own the program**; if you don’t like what we do, you can go in a different direction
  - Modifiable & Redistributable
  - Does not restrict use of commercial data services
- **Completely distributed**
  - Simple components-based architecture
  - Any component can be on any computer hardware
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the-shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgresSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
  - You own the program; if you don’t like what we do, you can go in a different direction
  - Modifiable & Redistributable
  - Does not restrict use of commercial data services
- **Completely distributed**
  - Simple components-based architecture
  - Any component can be on any computer hardware
  - Distributed catalog: harvesting, distributed search
Development Principles

- **Web-based**, light client for users, administrators, curators
- **Follows Open Source Standards** Search/Harvest: OAI, Z39.50; Metadata: DC, Marc, DDI; Identifiers: URN, Handles
- **Built with off-the-shelf components** E.g.: Apache web server, OpenLDAP, R, Zelig, PostgresSQL Integration: Perl, Java Servlets, XSL/XML
- **Open-Source**
  - Source code is included
  - *You own the program*; if you don’t like what we do, you can go in a different direction
  - Modifiable & Redistributable
  - Does not restrict use of commercial data services
- **Completely distributed**
  - Simple components-based architecture
  - Any component can be on any computer hardware
  - Distributed catalog: harvesting, distributed search
  - Distributed data: proxying, caching, replication
Next at the VDC
Next at the VDC

- First public version just released
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
- Agreement, under Library of Congress auspices among
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
- Agreement, under Library of Congress auspices among
  - ICPSR (U Michigan),
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
- Agreement, under Library of Congress auspices among
  - ICPSR (U Michigan), Roper Center (UConn),
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
- Agreement, under Library of Congress auspices among
  - ICPSR (U Michigan), Roper Center (UConn), Odum Institute (UNC),
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
- Agreement, under Library of Congress auspices among
  - ICPSR (U Michigan), Roper Center (UConn), Odum Institute (UNC), NARA,
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
- Agreement, under Library of Congress auspices among
  - ICPSR (U Michigan), Roper Center (UConn), Odum Institute (UNC), NARA, HMDC,
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
- Agreement, under Library of Congress auspices among
  - ICPSR (U Michigan), Roper Center (UConn), Odum Institute (UNC), NARA, HMDC, Murray
Next at the VDC

- First public version just released
- Being installed by University of North Carolina
- Agreement with U.S. Census Bureau
- Agreement, under Library of Congress auspices among
  - ICPSR (U Michigan), Roper Center (UConn), Odum Institute (UNC), NARA, HMDC, Murray
- Interest from many universities and other organizations
http://GKing.Harvard.edu