Access to Microdata: Developing Virtual Collaboratories

Chet Bowie,
Tim Mulcahy
NORC – University of Chicago
Overview

- Motivation
- Current Modalities
- What are the Barriers?
- Practical Approach
  - Enclave Structure
  - Portfolio Approach
  - Documentation
- Setting up Virtual Collaboratory
  - Structure
  - Developing metrics
- Summary
Micro data important for variety of public policies, e.g. understanding
- Mobility of STEM workforce
- Impact of academic institutions and discipline choice on earnings outcomes
- Career paths and job ladders
- Limited access => underutilization AND little input into survey creation, design and implementation

Huge loss for policy community
Current Modalities

- Public Use Files
  - Quality
  - Timeliness
  - Typically not available for linked microdata

- Licensing
  - Security
  - Typically not available for linked data

- RDC’s
  - Cost
  - Accessibility
Secure
Flexible
Low Cost
Meet Replication standard
• The only way to understand and evaluate an empirical analysis fully is to know the exact process by which the data were generated
• Replication dataset include all information necessary to replicate empirical results
• Metadata crucial to meet the standard
  • Composed of documentation and structured metadata
  • Undocumented data are useless
Create foundation for metadata documentation and extend data lifecycle
Data collection is not a static process - it’s a lifecycle
It dynamically evolved across time and involves many players
It extends to aggregate data to reach decision makers
Metadata are crucial to capture knowledge
What are the barriers?

- Security Concerns
- Perceptions
- Value Added for agency?
1. Security Concerns: Physical Data Protection
   a) Already collect data for multiple statistical agencies (BLS, Federal Reserve (IRS data), EIA, NSF/SRS etc.) => safeguards in place
   b) NIST approved IT security plan
2. Security Concerns: Provision of access – a portfolio approach
   a) Statistical protection (statistical)
   b) Researcher training (Educational)
   c) Dissemination to researcher community (Operational)
   d) Agency-specific data protection requirements (Legal)
3. Perceptions: Data archiving, indexing, and curation
4. Perceptions: Researcher engagement through virtual collaboratory
   a) High quality analysis leverages federal investment
   b) Metadata documentation improves scientific quality
security
Physical Data Protection

- Encrypted connection with the data enclave using virtual private network (VPN) technology. VPN technology enables the data enclave to prevent an outsider from reading the data transmitted between the researcher’s computer and NORC’s network.
- Users access the data enclave from specific, pre-defined IP addresses.
- Citrix’s Web-based technology.
  - All applications and data run on the server at the data enclave.
  - Data enclave can prevent the user from transferring any data from data enclave to a local computer.
  - Data files cannot be downloaded from the remote server to the user’s local PC.
  - User cannot use the “cut and paste” feature in Windows to move data from the Citrix session.
  - User is prevented from printing the data on a local computer.
- Audit logs and audit trails
### Menu Options for Agency X (and Study Y)

<table>
<thead>
<tr>
<th>Sample Modalities</th>
<th>Legal Options (1,2,3,4)</th>
<th>Statistical Options (1,2,3,4,5)</th>
<th>Operational Options (1,2,3,4,5)</th>
<th>Educational Options (1,2,3,4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote Access</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Onsite Access</td>
<td>3 with customization</td>
<td>3,5</td>
<td>1</td>
<td>None</td>
</tr>
<tr>
<td>Licensing</td>
<td>2</td>
<td>1</td>
<td>2,3</td>
<td>1,4</td>
</tr>
<tr>
<td>(different levels of anonymization)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Legal and Statistical Protections

- **Legal**
  - MOU between the agency and the institution
  - Data User Agreement signed by researcher & institution
  - Nondisclosure agreement
  - Access Agreement signed by institutional and individual researcher
  - Approved institutions
  - Access limited to data requested and authorized

- **Statistical**
  - Remove obvious identifiers and replace with unique identifiers
  - Statistical techniques chosen by agency (recognizing data quality issues)
  - Full disclosure review before results leave the enclave

Note: All are at discretion of agency and can go above and beyond the minimum level of protection.
Subjects
• Basic confidentiality
• Agency specific (joint with agency)
• Dataset specific (joint with agency)

Locations
• Onsite
• Researcher locations (AAEA, JSM, AOM, ASA, ASSA, NBER summer institute)

Note: The training is designed to go above and beyond current practice in terms of both frequency and coverage
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30-9:00</td>
<td>Welcome (NASS/ERS/NORC)</td>
</tr>
<tr>
<td>9:00-10:30</td>
<td>Data enclave navigation (NORC)</td>
</tr>
<tr>
<td>10:30-10:45</td>
<td>Break</td>
</tr>
<tr>
<td>10:45-12:15</td>
<td>Metadata documentation (NORC)</td>
</tr>
<tr>
<td>12:15-1:15</td>
<td>Lunch</td>
</tr>
<tr>
<td>1:15-2:45</td>
<td>Confidentiality and data disclosure (NORC)</td>
</tr>
<tr>
<td>2:45-3:00</td>
<td>Break</td>
</tr>
<tr>
<td>3:00-4:00</td>
<td>ARMS survey overview (ERS) – ERS Staff</td>
</tr>
<tr>
<td>4:00-4:10</td>
<td>Confidentiality agreement signing</td>
</tr>
</tbody>
</table>
Day 2

8:30-9:00  ARMS data files and documentation (ERS) – ERS Staff
9:00-10:00 Sampling and weights (NASS and ERS) – ERS Staff
10:00-10:15 Break
10:15-11:15 Item quality control and treatments for non-response ERS/NASS Staff
11:15-12:15 Statistical testing (NASS and ERS)
12:15-12:30 Closing and adjournment
Getting Value Added and Addressing Perceptions: Setting Up the Virtual Collaboratory
Goals

- **Value Added**
  - Serve Agency Mission
  - Metadata documentation
    - Code
    - Information about variables

- **Policy Relevance**
  - Research output
    - Cite sources
    - Evaluation and feedback
Reduce cost of analysis
- Create collaborative environment, with wiki and FAQ’s
- Capture research process through blogs (replication standard)
- Engage data producers

Facilitate publication of results and increase visibility of work
- Facilitate reporting, citations, etc.
- Facilitate reusability / extend the research
- Compare results

Create value for future research
- Integrate research results in the survey knowledge
Accessing the enclave

Enter your user name and password. The first time you access, you will need to change your password.

The message center will inform you on browser related technical issues. Note that you will first need to install the Citrix Client on your system (a download link will be provided).
Launch the desktop

The browser downloads the .ica file and launches the Citrix Client.

The first time you login, you will a message to exit full screen mode. For security reasons, this functionality has been disabled in the Enclave. Check “do not show again”
Enclave Desktop View

Shortcut provides access to the portal, the My Documents and the Research Group file system

The Producer collaborative portal page is displayed when you login

A tab provides access to you a private collaborative portal for your research group

Note that the desktop shortcuts are also accessible through the Start menu under the Data Enclave program group
Enhancing Communication
Use Startup Button to Access Applications
Software available in the enclave

- Stata/SE 9.2
- StatTransfer 9
- SAS v9.1
- Microsoft office 2007
- Adobe PDF Reader
- IHSN Microdata Management Toolkit / Nesstar Publisher
  (upon request, selected users only)
Collaborating: Research Group Portal

- Collaborate
  - Announcements
  - Calendar
  - Team Discussion
  - Tasks
- Share knowledge
  - Shared Documents
  - BYU Blog
  - BYU Wiki
- Get Support
  - NORC Enclave
  - NIST ATP Documentation
  - Technical Support

- Organize & Exchange
- Share Files
  - Capture Events & Knowledge
- Access to Global Information and Support
Editing content

No need to know HTML. Most of the content in SharePoint can be edited using this rich text editor. Basic functionalities include changing text font, colors or alignment and creating tables.
Use the discussion groups to exchange ideas, submit questions, etc.

Use the enclave announcement, tasks / todo, and calendar to distribute and organize the research work.

Organizing work and exchanging ideas
Organizing work and exchanging ideas

- Research is an iterative, evolving process
- Capturing ideas and milestone is crucial
- Personal logs have often been used in the past
Multiple authors can add, remove, and edit content (mass authoring).
Knowledge grows across time based on community contributions.
Pages automatically link to each other page on “topics”
Using document libraries

Click on documents and folders to open or navigate in the structure.

Use the menu to create folders or upload documents.

**ARMS Shared Library**

- Share a document with the team by adding it to this document library.
  - **New**
  - **Upload**
  - **Actions**
  - **Settings**

**Existing Documents**

- FBH2004 All Respondent Booklet Phase3
- FBH2004 Core Questionnaire Phase3
- FBH2004 CRR Contract Supplement Phase3
- FBH2004 CRR Questionnaire Phase3
- FBH2004 Hogs Questionnaire Phase3
- FBH2004 Peanut Questionnaire Phase3
- FBH2004 Wheat Questionnaire Phase3

**New Document**

- Create a new document in this library.

**New Folder**

- Add a new folder to this document library.

**Upload Document**

- Upload a document from your computer to this library.

**Upload Multiple Documents**

- Upload multiple documents from your computer to this library.
Researcher Collaboration:
Sharing Ideas, Code, Macros

ARMS Discussions

Shared SAS Files (Macros)

Brian Briggsman

All:

1. I have posted two SAS files in the ARMS Shared Library. The first SAS file contains two macros that pcalcute weighted descriptive statistics using the survey weights and the jackknife variance estimator. The first macro is for the entire dataset and the second macro will calculate weighted descriptive statistics by a group variable. The second SAS file contains a macro that runs a weighted OLS regression using the survey weights and the jackknife variance estimator. Each SAS file contains a discussion as to what each macro does and where to input your variables, data, weights, etc.

2. I use these SAS macros to develop results in SAS. If you have any questions, please contact me.

Regards,

Brian C. Briggsman
Assistant Professor - Agricultural Economics
Oklahoma State University
317 Agricultural Hall
Office - (405) 744-6171
brian.briggsman@okstate.edu
Engaging Data Producers

ARMS Discussions

Use the Team Discussion list to hold newsgroup-style discussions on topics relevant to your team.

Actions | Settings | View
---|---|---

Started: 09/12/2007 7:43 AM
Posted By: Linda Allinson

**Saving subsets**

I was asked about saving subsets of data. You can do this easily in SAS by saving to your SASUSER directory. For example,

```
libname arms 'C:\surveys\arms\sas';

data sasuser.ac00hogs;
set arms.ac00fhbh;
if hogs>0;
run;
```
Engaging the Data Custodian

ARMS Discussions

What are issues in adding R software package to the Enclave?

1. I enjoyed the discussion of the Enclave meeting the NIST standard for software use. It's absolutely necessary. Could we discuss what the specific issues are that are currently preventing the R package from being included? We can rule out cost, since R is a OSS (free and open-source software).

2. I'm likewise an R user and can see the benefits of having it available in the Enclave environment, in particular when it comes to data visualization. For those of you not familiar with the product, visit http://www.r-project.org (it is an open source version of the S-Plus software).

The main issues we'll have to consider are:
1. Does R meet the security standards of the Enclave?
2. How many users would be interested in using R?
3. If we move forward and deploy the software, what's the impact on the security plan and how long it will take for it to be reviewed by NIST?

I think we should then also identify an R "champion" who can assist less experienced users and seed the wiki with getting started information, tools, and tips.

Show Quoted Messages
Use the technical support facility for:
- Technical issue regarding the enclave
- Problems accessing or using data
- Report data quality issues
- Request a document export
- Any other question you might have!
Next steps

- Developing metrics
  - Number of interactions
  - Additions to the wiki
  - Research output (how to quantify)
- Developing incentives
  - Establish Leaders
  - External communications
Goal of enclave is to promote access to sensitive micro data while protecting confidentiality.

Clear challenges that can be addressed by collaboratory approach:
- Security
- Perceptions
- Value added

NORC data enclave essentially a “test-bed”

We welcome your ideas and comments!