Remote Access to Microdata
Challenges and Practical Approaches

Julia Lane and Tim Mulcahy,
NORC/Uof Chicago
Pascal Heus, Open Data Foundation
Overview

► Challenges

► Practical Approach
  ▪ Enclave Structure
  ▪ Portfolio Approach
  ▪ Documentation

► Metadata Access
Challenges

► Provide access to micro data
► Protect Confidentiality
  ▪ Technical
  ▪ Legal
  ▪ Organizational
  ▪ Statistical
► Archive, Index and Curate Micro-data
Challenges to Providing Access

► Public Use Files
  - Quality
  - Timeliness

► Licensing
  - Security

► RDC’s
  - Cost
  - Accessibility
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.
Practical Approaches: NORC Data Enclave

➤ Consortium
- Founding Member NIST (ATP)
- Additional members: USDA (ERS), Kauffman Foundation

➤ Ongoing discussions with Federal Agencies
NORC Data Enclave

1. 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. Data archiving, indexing, and curation
   a) Microdata Dissemination Toolkit (DDI)
   b) Curation plan

3. 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)

4. Advisory Boards
### 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 (1,2,3,4,5)</th>
<th>Operational (1,2,3,4,5)</th>
<th>Educational (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>
The Data Enclave is fully compliant with DOC IT Security Program Policy, Section 6.5.2, the Federal Information Security Management Act, provisions of mandatory Federal Information Processing Standards (FIPS) and all other applicable NIST Data IT system and physical security requirements. This includes

- Employee Security
- Rules of Behavior
- Nondisclosure Agreements
- IT System Security
- Applicable Laws and Regulations
- Network Connectivity
- Remote Access
- Physical Access

See

IT Security

- 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
Legal and Statistical Protections

- **Legal**
  - 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 (recognising data quality issues)
    
  Note: Both are at discretion of agency and can go above and beyond the minimum level of protection
Researcher Training

- **Subjects**
  - Basic confidentiality (CDAC)
  - Agency specific (joint with agency)
  - Dataset specific (joint with agency)

- **Locations**
  - Onsite
  - Webbased
  - Researcher locations (AAEA, J SM, 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.*
ARMS Data Enclave Training Agenda

NORC.
1350 Connecticut Avenue (between N and O), NW, Suite 500, Washington D.C. 20036

Monday, September 10

8:30-9:00 Welcome (NASS/ERS/NORC)
9:00-10:30 Data enclave navigation (NORC)
10:30-10:45 Break
10:45-12:15 Metadata documentation (NORC)
12:15-1:15 Lunch
1:15-2:45 Confidentiality and data disclosure (NORC)
2:45-3:00 Break
3:00-4:00 ARMS survey overview (ERS) – Jim Johnson, ERS
4:00-4:10 Confidentiality agreement signing
Tuesday, September 11

8:30-9:00  ARMS data files and documentation (ERS) – Bob Dubman, ERS

9:00-10:00  Sampling and weights (NASS and ERS) – Nick Schauer (NASS) and Dave Banker (ERS)

10:00-10:15  Break

10:15-11:15  Item quality control and treatments for non-response (NASS and ERS) - Richard Barton (NASS) and Mitch Morehart (ERS)

11:15-12:15  Statistical testing (NASS and ERS) –Bob Dubman (ERS) and Phil Kott (NASS)

12:15-12:30  Closing and adjournment
Researcher Responsibilities

- Serve Agency Mission
- Metadata documentation
  - Code
  - Information about variables
- Post research output
- Cite sources
- Evaluation and feedback
Accessing the enclave and promoting metadata documentation
Accessing the enclave

[Image: Web Interface Log In - Windows Internet Explorer]

https://enclave.norc.org
*** don’t forget it’s https://, not http:// ***

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)
Once logged in, the system provides access to the data enclave through a virtual desktop. The full desktop will
The Research Group portal page is displayed when the user logs in.

This shortcut provides access to the research group data and documentation folders prepared by the enclave along with the group shared directories.

These two shortcuts provide access to the research group and NORC portals.

Note that the desktop shortcuts are also accessible through the Start menu under the Data Enclave program group.
Getting Work Done
You can access your application like you normally do on a regular computer using the Start menu. Note that administrative functionalities have been disabled for security reasons.
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)
## Announcements

**Company Size Assymetry vs. Motivation to Form R&D Alliances**
**5/14/2007 1:56 PM**

by Tim Mukohy

BYU researchers are investigating whether motivations for forming R&D alliances are different for large vs. small firms and whether the performance outcomes are different. They expect that small firms will be much more likely to commercialize technology.

- Add new announcement

## Team Discussion

**Subject**

There are no items to show in this view of the "Team Discussion" discussion board. To create a new item, click "Add new discussion" below.

- Add new discussion

## Shared Documents

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Modified By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Files for export</td>
<td>Pascal Haus</td>
<td></td>
</tr>
<tr>
<td>Scripts</td>
<td>Pascal Haus</td>
<td></td>
</tr>
<tr>
<td>References</td>
<td>Pascal Haus</td>
<td></td>
</tr>
</tbody>
</table>

- Add new document
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
Overview

- What is metadata and why is it important
- What can metadata do
- Guide to creating metadata
- Creating metadata in the enclave
Creating incentives for metadata documentation for researchers

► Reduce cost of analysis
  - Create collaborative environment, with wiki and FAQ’s
  - Capture research process through blogs (replication standard!)

► 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
Using the SharePoint based portal

But be aware that
(1) not all documented functionalities are available
(2) Some functions require administrative access

Use the HELP!
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.
Organizing work and exchanging ideas

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.
Research is an iterative, evolving process

Capturing ideas and milestone is crucial

Personal logs have often been used in the past

Blogs is today’s version of it
Using the wikiwiki to capture research knowledge

Familiar with Wikipedia?

A wiki is a shared web site but does not require programming skills to maintain

Multiple authors can add, remove, and edit content (mass authoring).

Knowledge grows across time based in community contributions

Pages automatically link to each other page on
Summary

► Goal: To promote access to sensitive micro data while protecting confidentiality

► Benefits:

- Secure, low-cost approach to microdata access
- Archiving, Indexing, and Curation of Micro-data
- Applicable and Customizable to agency needs and requirements