

Creating a Data Enclave for Sensitive Microdata

Stephanie Shipp

Science and Technology Policy Institute

202-419-5498

sshipp@ida.org

American Evaluation Association Annual Meeting

Denver, Colorado

November 6, 2008

Researcher Access to Microdata

Data Enclave

- Why the Data Enclave was created
- Applying to use the Data Enclave
- Data Enclave Innovations
 - Meta Data Documentation
 - Collaboratory

Keeping information private is important



Conceptual Framework

- Ethical and legal obligation to respondents to protect confidentiality
 - Increasingly sophisticated data matching technology magnify risks
 - Maximize protection (some type of noise/masking)
- Data producers want data examined and analyzed
 - To inform and improve program
 - To replicate scientific results
- Goal : protect respondents while allowing researchers to access data for analysis

Why Create a Data Enclave for the Advanced Technology Program (ATP)?

- ATP has a unique source of innovation data which researchers can use to study:
 - Entrepreneurship & innovation
 - Early stage technology development
 - Commercialization of high-risk R&D
- But not the resources
 - To conduct extensive analysis on own
 - To monitor access by researchers

ATP-NORC Collaboration to Create Data Enclave

- ATP contracted with NORC in 2006
- NORC created Data Enclave in 2007
- Currently a few dozen academic researchers are using ATP data, USDA data, or Kaufmann data

Provision of Research Access

Two Approaches

➤ Remote access

- external researchers access data via an encrypted connection with the data enclave using VPN
- Restrict user access from specific, pre-defined IP addresses
- Citrix technology to access applications – configured so no downloads, cut and paste or print possible

➤ Onsite access

- Secure room at NORC site (DC, Chicago)
- Secure machines
- Video camera
- Audit logs and trails
- Workspaces (that could be shared)

NORC Data Enclave: How to Access

Applying to use the Data Enclave:

<http://dataenclave.norc.org>

Eligibility: Researcher from recognized research institution

Cannot be an individual consultant

Cannot be a for-profit organization

Both researcher and research organization must complete application, promising to maintain confidentiality of data

Research output must be in aggregate format that does not identify individual companies.

Working papers (to be shared with ATP (TIP))

Journal Publications

Conference Presentations

Both NORC and ATP (TIP) approve application

Research Proposal

<http://dataenclave.norc.org>

Study ATP surveys and existing analysis:

What research questions do you want to answer?

Complete application:

- Identify research question and data needed to answer the question(s)
- Obtain department signoff
- Attend Data Enclave Training (at NORC, via web, or at conferences)

Conducting the Research

- Access the data remotely through the NORC data enclave.
 - Microsoft suite (WORD, EXCEL, ACCESS, etc.)
 - SAS, STATA, other statistical software packages
- Conduct research within Data Enclave.
 - Output must be reviewed by NORC staff and then emailed to you (to ensure that data cannot identify a company)

Sharing Research

- Save improvements to data within Data Enclave
 - Editing the data
 - Creating new variables
 - Matching to other data sources
- Document work (create metadata)
- Present results at ATP (TIP) and NORC organized conferences as well as at other conferences
- Acknowledge Data Enclave and ATP (TIP) in work.

Data Enclave Innovations:

- 1. Creating Meta-data**
- 2. Sharing Research Inputs and Outputs
(Collaboratory)**

Unlabeled cans

Labeled cans

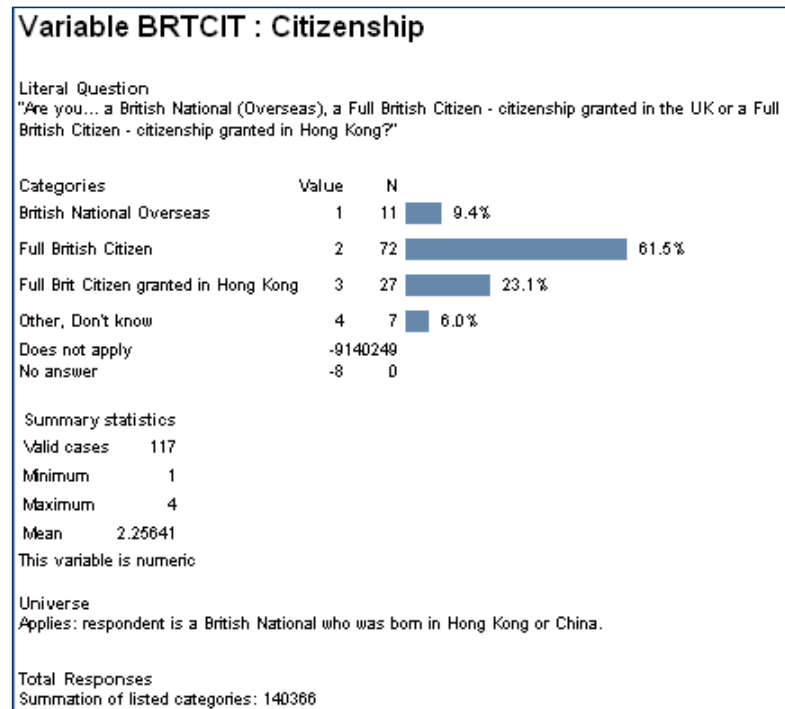


What is metadata?

Data about data

1	1	4	5	13
1	1	4	5	7
1	1	4	5	4
1	1	4	5	21
1	1	4	2	7
1	1	3	4	4
1	1	4	5	6
1	1	1	5	4
1	1	2	5	1
3	1	1	3	1
3	1	9	3	16
3	1	9	2	4
3	1	9	9	19
3	3	2	9	4
3	1	9	3	99

Unlabeled data



Labeled data

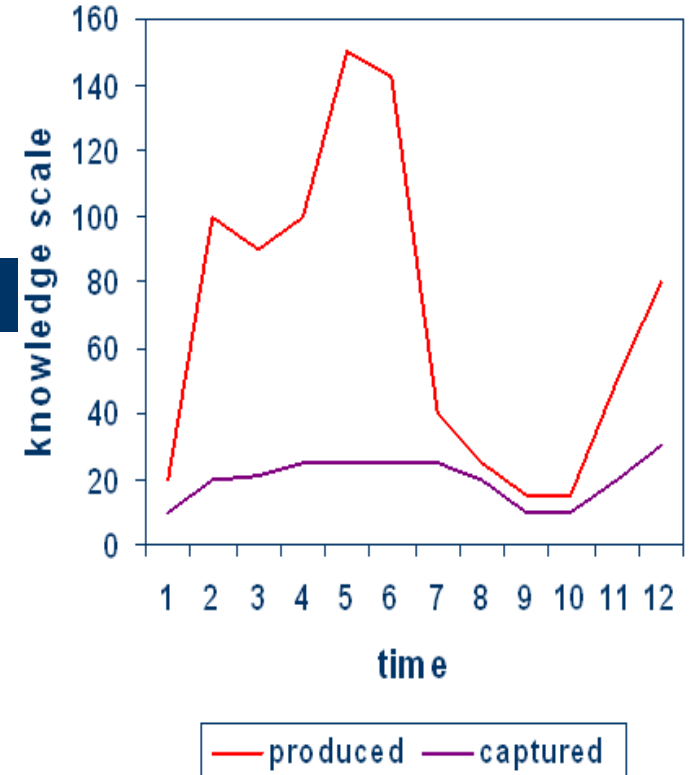
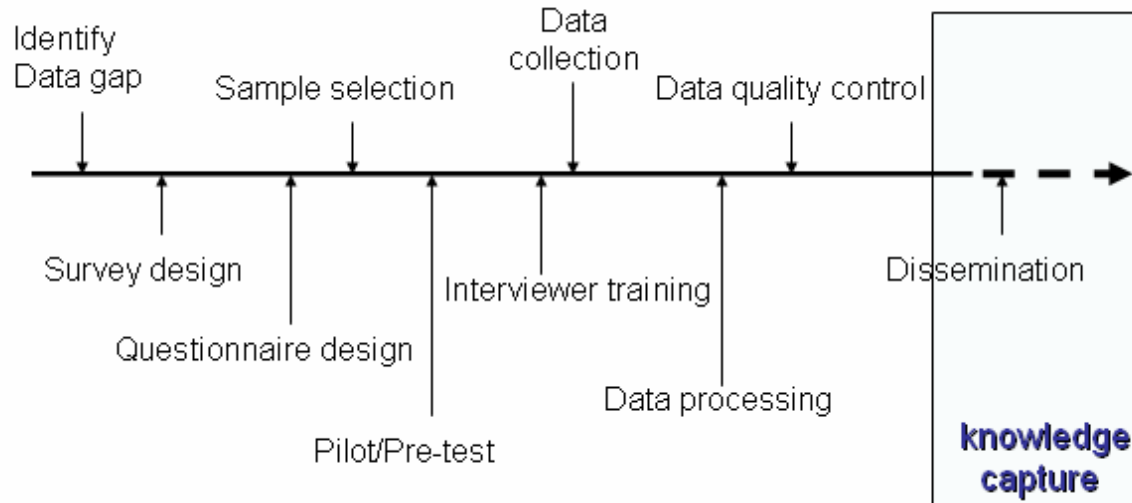
Why are meta-data important?

- To understand the data
- To track comments and improvements made to the data

Imagine a world without metadata....

- Users would say:
 - I can't find the right data! How do I get access?
 - Where is the report / questionnaire / methodology?
 - I don't understand this survey / file / variable
 - I can't merge the files
 - How do I weight the data?
 - My results don't match the report, I can't reproduce the same results
 - Are these things comparable?
 - I didn't know someone did this research before?
- Sounds familiar?
 - Metadata is an answer to a researcher's frustrations
- Producers and archivists are making efforts to improve metadata but similarly, metadata must also be captured by researchers (Life Cycle!)

When to capture metadata?



- Metadata must be captured at the time the event occurs!
- Documenting after the facts leads to considerable loss of information
- This is true for producers and researchers

Metadata and the Replication Standard

- Replication standard
 - Gary King, Harvard, 1995
 - The only way to understand and evaluate an empirical analysis fully is to know the exact process by which the data were generate
 - Replication dataset include all information necessary to replicate empirical results
- Metadata crucial to meet the standard
 - Composed of documentation and structured metadata
 - Undocumented data is useless

Collaboratory: Multiple ways to Share Data

- Improve the codebook by entering metadata
- Sharepointe
- Organizing and exchanging ideas
- Blogs
- Wikis
- Sharing and tagging files

Entering Metadata

Variables

Variables:

Number	Name	Label	Width	Start Col	End Col	Record	Dec
v1	APPTYPE	YEAR 2002 APPLICANT TYPE	1	*	*	*	
v2	WESID		4	*	*	*	
v3	Q1A	IMPORTANCE OF NO INTERI	1	*	*	*	
v4	Q1B		1	*	*	*	
v5	Q1C		1	*	*	*	
v6	Q1D		1	*	*	*	
v7	Q1E		1	*	*	*	
v8	Q1F		1	*	*	*	
v9	Q1G		1	*	*	*	

Variable Description:

Categories:

- Category Hierarchy
 - 1 - Extremely Important
 - 2 - Very important
 - 3 - Somewhat important
 - 4 - Not too important

Documentation

Statistics | Weights | Documentation

Fields:

- Description
 - Definition
 - Universe
 - Source of Information
- Question
 - Pre-Question Text
 - Literal Question
 - Post-Question Text
 - Interviewer Instructions
- Imputation and Derivation
 - Imputation
 - Recoding and Derivation
- Others
 - Security
 - Notes

Pre-Question Text

Below are several reasons why a company might apply to ATP for funding.
Please tell us how important each reason was in your decision to apply to ATP:

Literal Question

Internal company funding is not available.

Value: Label:

Category Text:

Level Name:

GeoMap URI:

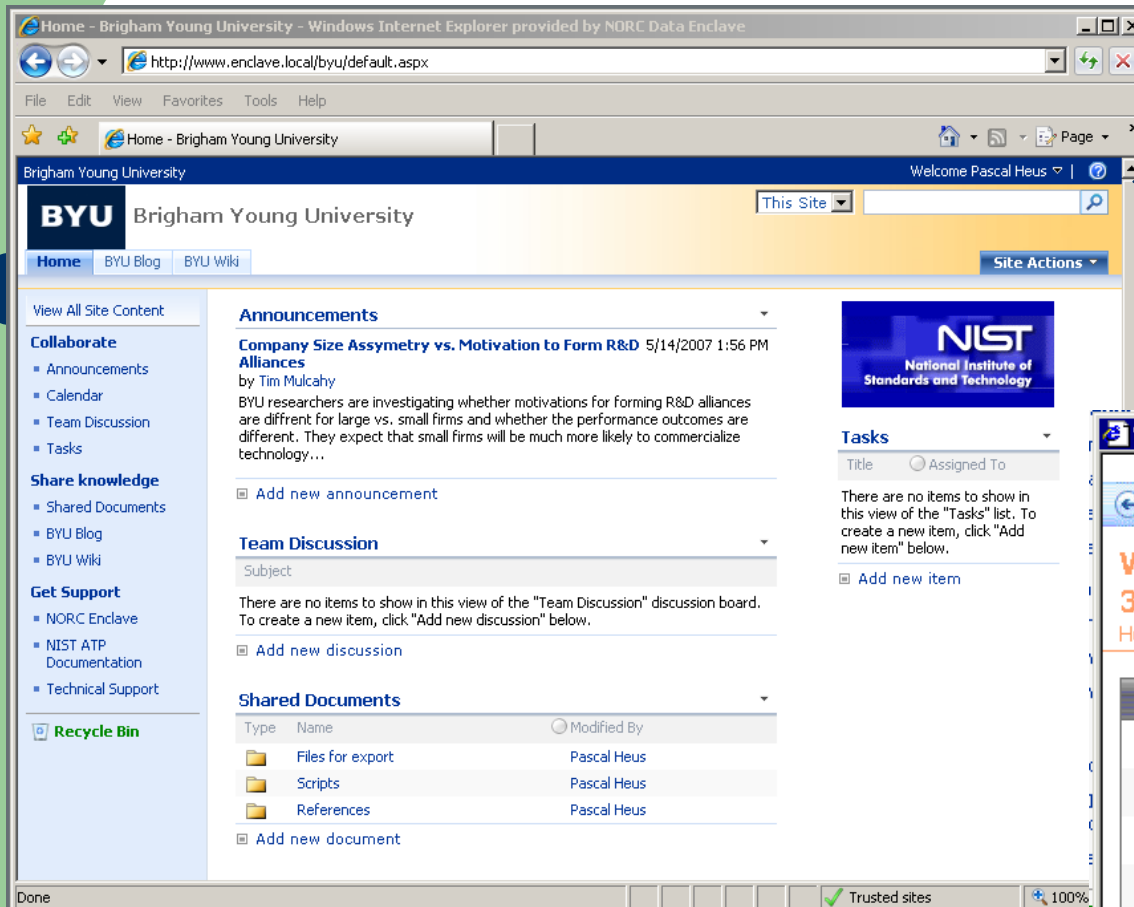
Variable information

Data Type:

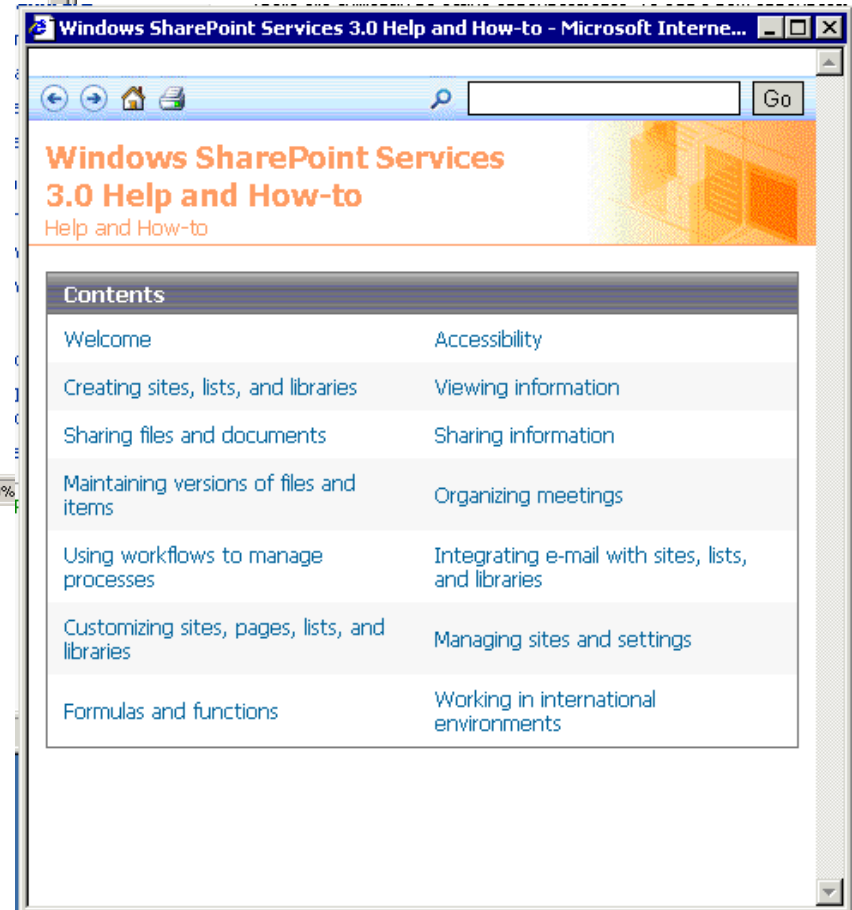
Measure:

Is Time Variable

Using the SharePoint based portal



Use the HELP!



But be aware that

- (1) not all documented functionalities are available
- (2) Some functions require administrative access

Organizing work and exchanging ideas

Announcements

New data available from USDA ! NEW 6/22/2007 12:02 AM
by Heus-Pascal

The 2004 ARMS survey is now available in the NORC data enclave. Contact your representative for further information.

[Add new announcement](#)

Calendar: New Item

Attach File * indicates a required field

enclave training

NORC Washington DC

6/27/2007 9 AM 00

6/27/2007 1 PM 00

enclave training for new enclave users. Introduction to data enclave and metadata.

Make this an all-day activity that doesn't start or end at a specific hour.

Recurrence: Make this a repeating event.

Workspace: Use a Meeting Workspace to organize attendees, agendas, documents, minutes, and other details for this event.

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

Tasks: Run regression of enterprise data

New Item | Edit Item | Delete Item

Title	Run regression of enterprise data
Priority	(2) Normal
Status	In Progress
% Complete	45%
Assigned To	Heus-Pascal
Description	Subset the data to small enterprise and run regression on funding amounts
Start Date	6/22/2007
Due Date	

Created at 6/21/2007 11:59 PM by Heus-Pascal
Last modified at 6/21/2007 11:59 PM by Heus-Pascal

Close

Use the discussion groups to exchange ideas, submit questions, etc

Brigham Young University > Team Discussion

Team Discussion

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

New Actions Settings

Discussion
Create a new discussion topic.



Team Discussion

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

Actions Settings View: Flat

Posted By Post

Started: 6/21/2007 11:48 PM View Properties Reply

How do I use SharePoint?
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.

Posted: 6/21/2007 11:48 PM View Properties Reply

Yes, I love this little editor!

Show Quoted Messages

Using the blog to capture research events

View All Site Content

Categories

- Category 1
- Category 2
- Category 3
- Add new category

Other Blogs

There are no items in this list.

Add new link

Links

- Photos
- Archive
- Archive (Calendar)
- Add new link

RSS Feed

Brigham Young University > BYU Blog

5/8/2007

Welcome to your Blog!

To begin using your site, click **Create a Post** under Admin Links to the right.

What is a Blog?

A Blog is a site designed to help you share information. Blogs can be used as news sites, journals, diaries, team sites, and more. It is your place on the World Wide Web.

Blogs are typically displayed in reverse chronological order (newest entries first), and consist of frequent short postings. With this Blog, it is also possible for your site visitors to comment on your postings.

In business, Blogs can be used as a team communication tool. Keep team members in touch by providing a central place for links, relevant news, and even gossip.

Posted at 1:19 PM by [Heus-Pascal](#) | [Permalink](#) | [Email this Post](#) | [Comments \(0\)](#)

Admin Links

- Create a post
- Manage posts
- Manage comments
- All content
- Set blog permissions
- Launch blog program to post

- 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 wiki to capture research knowledge



Brigham Young University > BYU Wiki > Wiki Pages > Home

Home

[Edit](#) | [History](#) | [Incoming Links](#)

Welcome to your wiki site!
You can get started and add content to this page by clicking **Edit** at the top of this page, or you can learn more about wiki sites by clicking **How to use this wiki site** in the Quick Launch.

What is a wiki site?

Wikiwiki means quick in Hawaiian. A wiki site is a Web site in which users can easily edit any page. The site grows organically by linking existing pages together or by creating links to new pages. If a user finds a link to an uncreated page, he or she can follow the link and create the page.

In business environments, a wiki site provides a low-maintenance way to record knowledge. Information that is usually traded in e-mail messages, gleaned from hallway conversations, or written on paper can instead be recorded in a wiki site, in context with similar knowledge.

Other example uses of wiki sites include brainstorming ideas, collaborating on designs, creating an instruction guide, gathering data from the field, tracking call center knowledge, and building an encyclopedia of knowledge.

View All Site Content

Wiki Pages

- Home
- How To Use This Wiki Site

Recycle Bin

Recent Changes

- Home
- How To Use This Wiki Site

[View All Pages](#)

- 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 “topics”

Sharing and tagging files

- Shared documents facilitate making information available to others
 - Documents, paper, etc
 - Scripts, programs
 - Tables
- Documents are organized by keyword/topics and can be used with the search function

Report data quality issues!

- A survey is not perfect, problems are always detected during research
- Data issues
 - Invalid code, missing values, file that cannot be merged, missing files or variables, inconsistent results, bad distribution, etc
- Metadata / Documentation issues
 - Undocumented variables or codes, discrepancies between docs and data, the post-processing/cleaning/quality assurance black box, etc
- Reporting this is crucial for other researchers and for the producer

Summary

- ▶ Goal: To promote access to sensitive micro data while protecting confidentiality of respondents (in this case ATP companies)
- ▶ Benefits:
 - Secure, low-cost approach to microdata access
 - ATP (TIP) program enhances knowledge about their program as well as improving their surveys
 - Researchers have access to new source of business microdata
 - Expands community of researchers who study innovation and entrepreneurship

Contact Information

- Website
 - <http://dataenclave.norc.org>
- Tim Mulcahy:
 - mulcahy-tim@norc.uchicago.edu
- Stephen Campbell:
 - stephen.campbell@nist.gov

Data Enclave Members

➤ Consortium

- Founding Member NIST (ATP)
- Additional member: USDA (ERS)
- Non-federal members: Kauffman Foundation

➤ Ongoing discussions with Federal Agencies

References

Lane, Julia, and Stephanie Shipp, 2007, “Using a Remote Access Data Enclave for Data Dissemination,” *The International Journal of Digital Curation*, Issue 1, Volume 2.

Lane, Julia, 2005, “Optimizing the Use of Micro-data: An Overview of the Issues,” paper presented at the American Statistical Association annual meetings, Minneapolis, MN.

NORC Data Enclave Newsletters:

○ <http://www.norc.org/NR/rdonlyres/81CDE8EB-438E-4689-A2BB-D2D7211C8E49/0/Newsletter34.pdf>

○ http://www.norc.org/NR/rdonlyres/F3BF266C-44F8-4105-A0F1-63011ABA8308/0/dataenclavenewsletter_vol1_issue2.pdf

○ <http://www.norc.org/NR/rdonlyres/127F0255-3DE1-4D4D-84FC-14EA32B7E90A/0/dataenclavenewsletter1.pdf>

Website: <http://dataenclave.norc.org>