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)

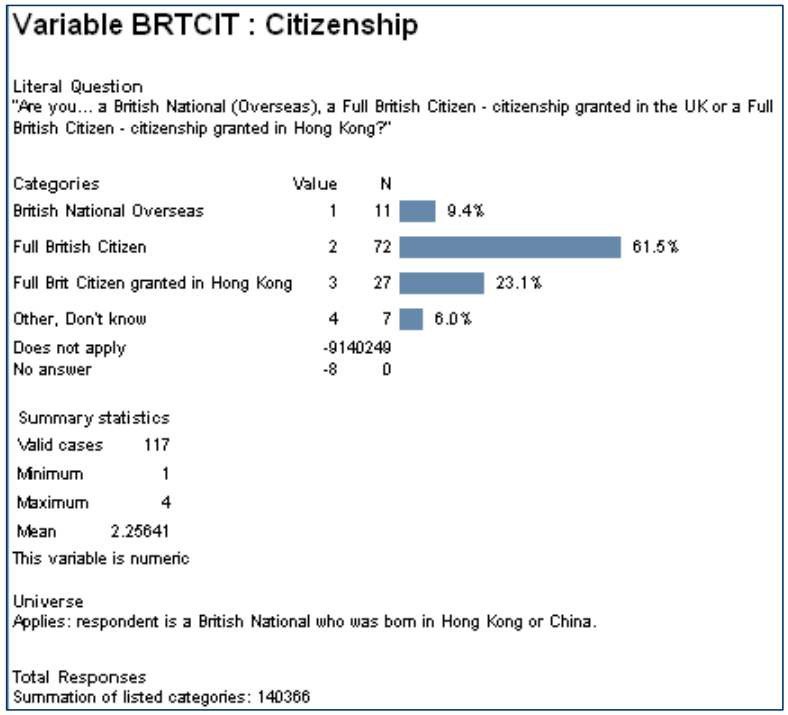
Unlabeled cans Labeled cans

This example is taken from: A Manager's Introduction to Adobe eXtensible Metadata Platform,http://www.adobe.com/products/xmp/pdfs/whitepaper.pdf

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



Variable BRTCIT : Citizenship

Literal Question
"Are you... a British National (Overseas), a Full British Citizen - citizenship granted in the UK or a Full British Citizen - citizenship granted in Hong Kong?"

Categories	Value	N	
British National Overseas	1	11	9.4%
Full British Citizen	2	72	61.5%
Full Brit Citizen granted in Hong Kong	3	27	23.1%
Other, Don't know	4	7	6.0%
Does not apply	-9	140249	
No answer	-8	0	

Summary statistics

Valid cases	117
Minimum	1
Maximum	4
Mean	2.25641

This variable is numeric

Universe
Applies: respondent is a British National who was born in Hong Kong or China.

Total Responses
Summation of listed categories: 140366

Labeled data

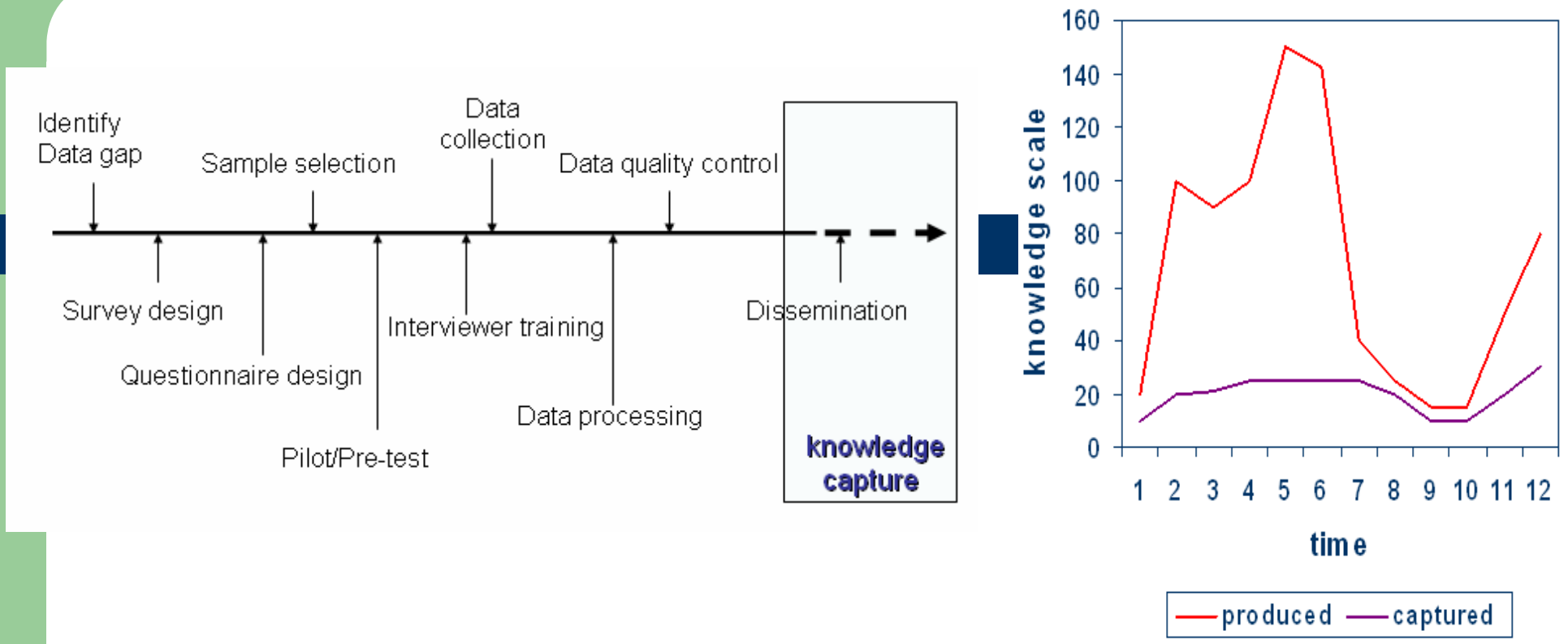
Why are meta-data important?

- 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

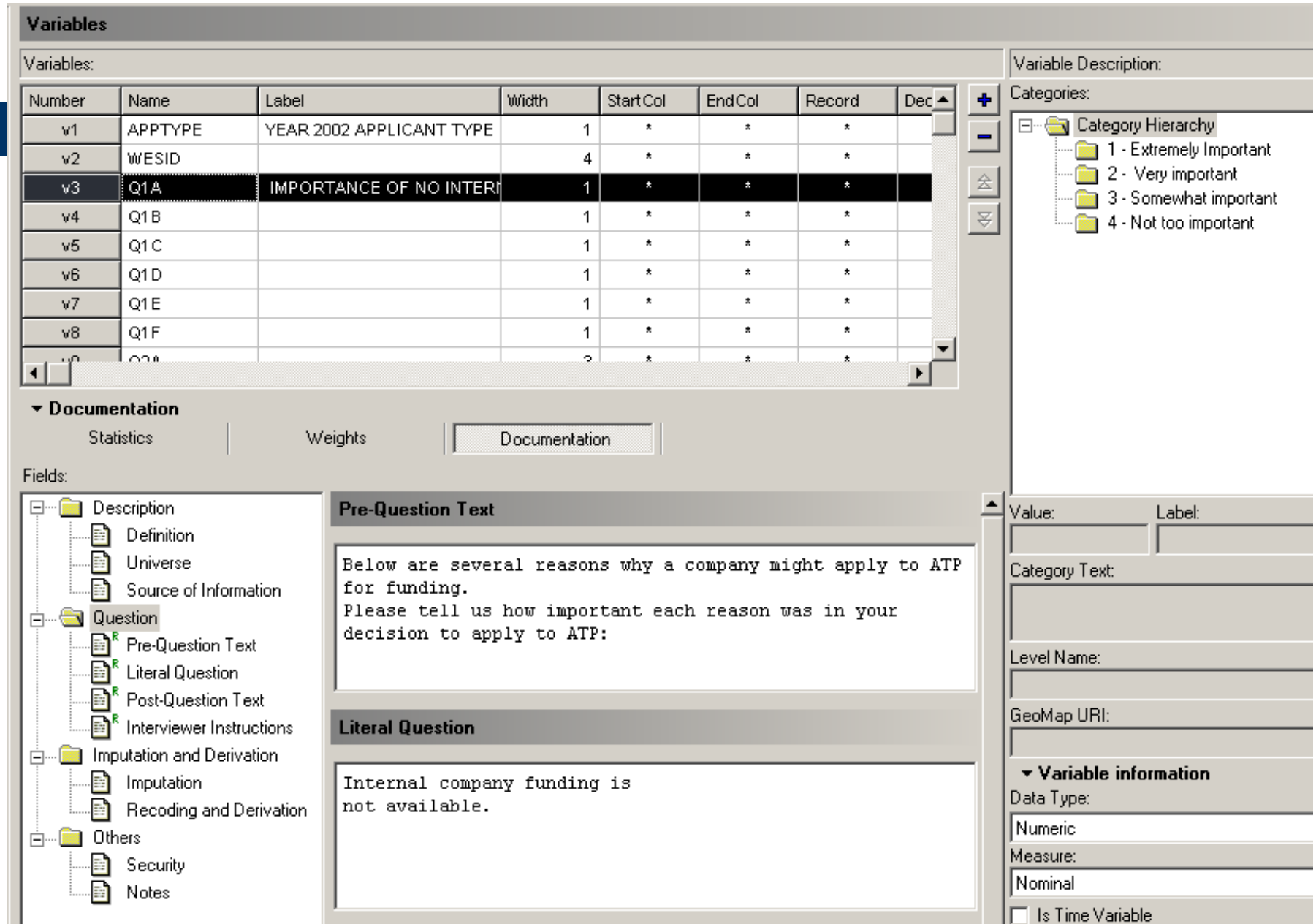
Metadata and the Replication Standard

- 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

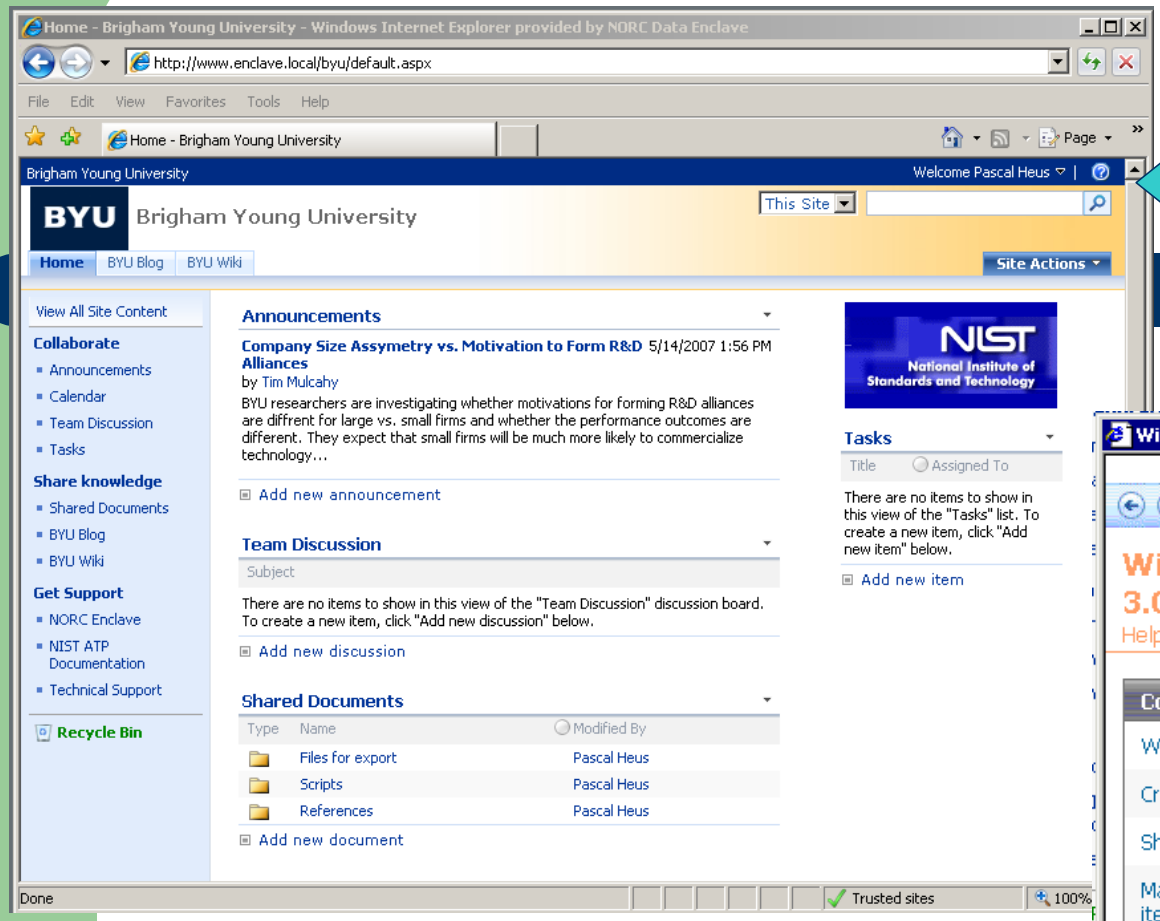
Collaboratory: Multiple ways to Share Data

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

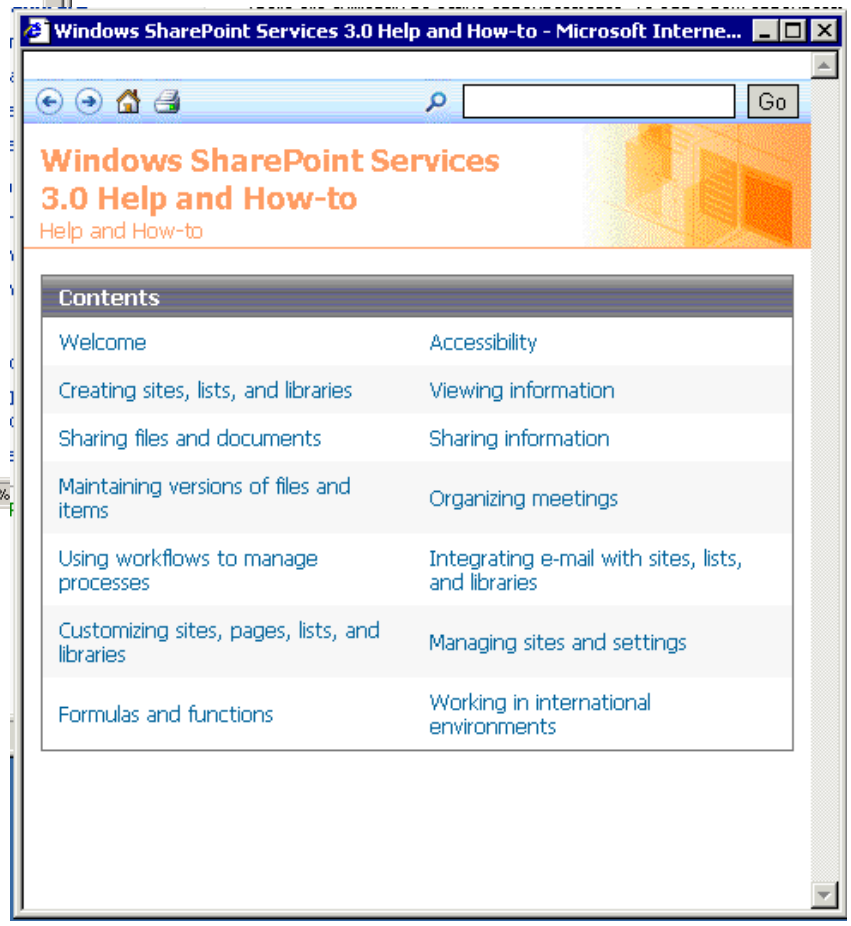
Entering Metadata



Using the SharePoint based portal

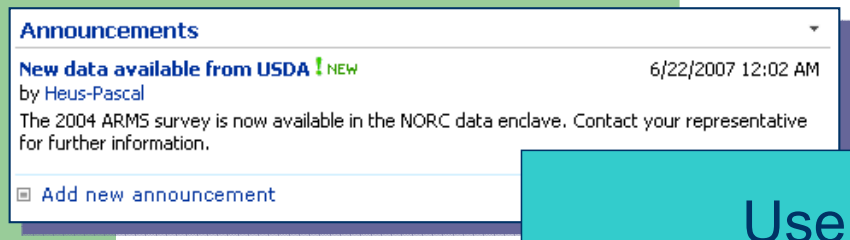


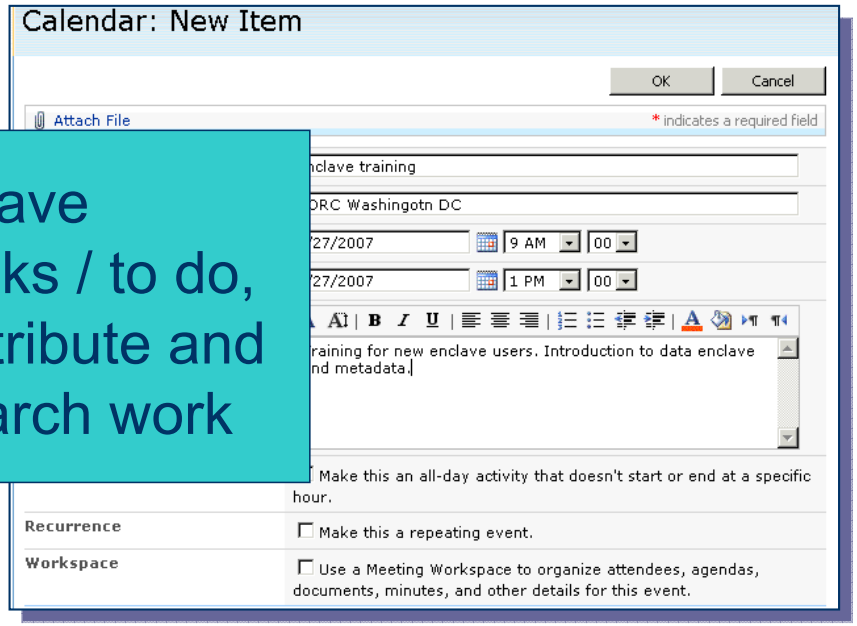
Use the HELP!



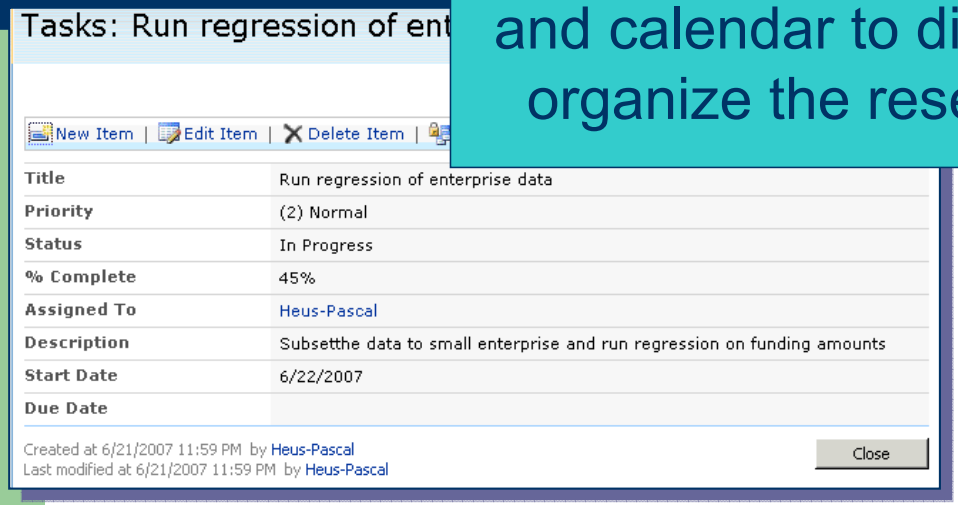
But be aware that
(1) not all documented functionalities are available
(2) Some functions require administrative access

Organizing work and exchanging ideas

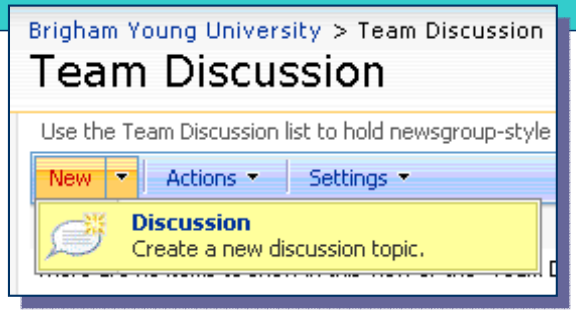


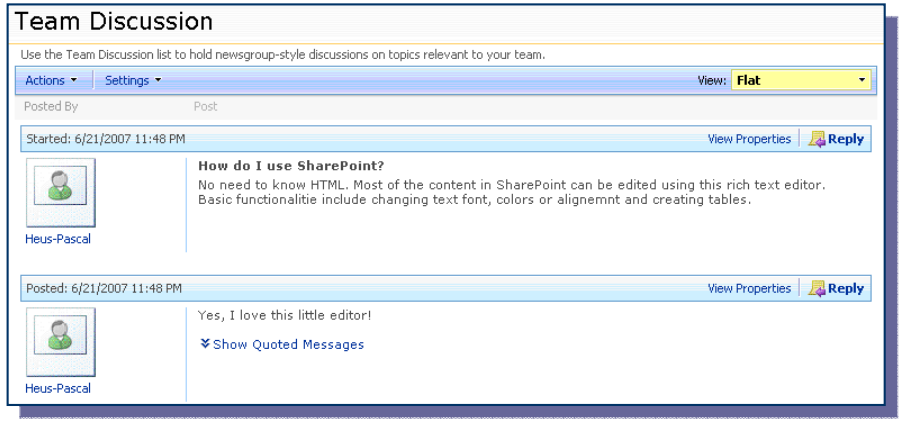


Use the enclave announcement, tasks / to do, and calendar to distribute and organize the research work



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

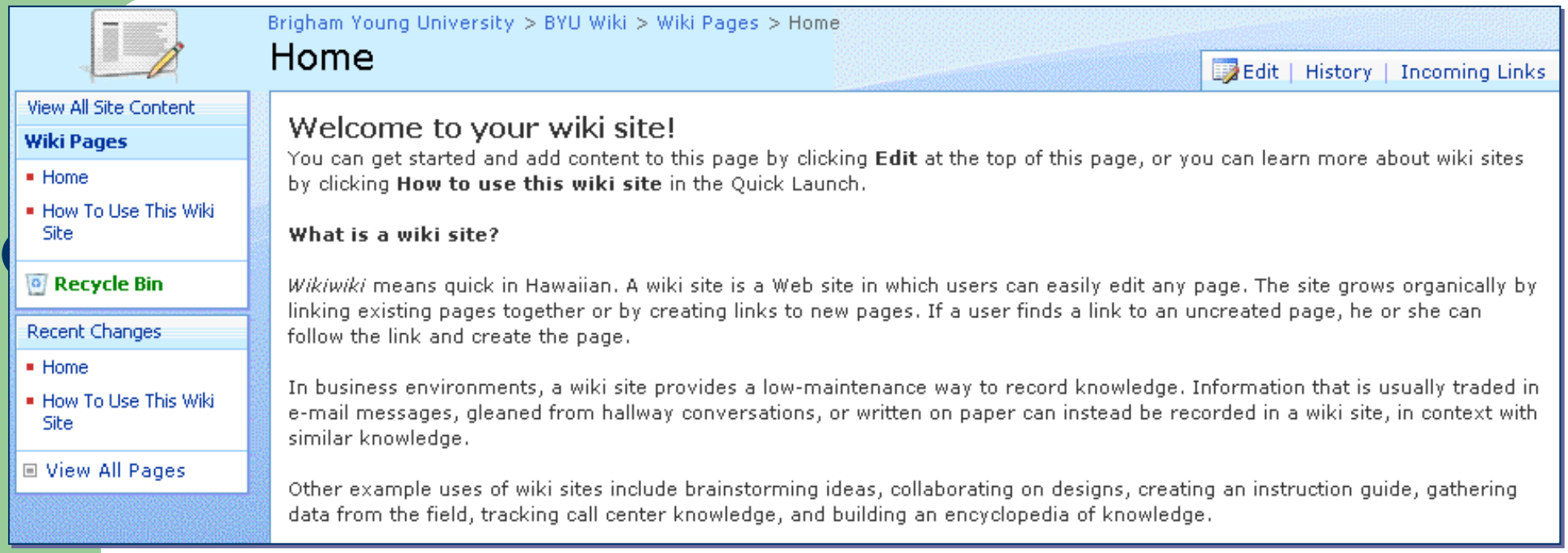




Using the blog to capture research events

- 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



- 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