# A MULTIDISCIPLINARY ANALYSIS OF DATA REUSE BEHAVIORS: AN INITIAL FRAMEWORK

GSLIS Graduate School of Library and Information Science

1 June 2011
IASSIST

Tiffany Chao
&
Nicholas Weber

# [OVERVIEW]

- Introduction
- Motivation
- Research questions
- Methods

- Data Reuse Framework
- Discussion
- Future directions

# [WHAT COMPLICATES STUDYING REUSE?]

- Lack of standards and practices for attribution of data producers
- Many disciplines lack infrastructure for archiving / publishing / sharing
- Social norms are informal
- Policies to influence proper data stewardship and management  are just beginning to go into place
- Research methods are rarely transparent

# [PREVIOUS QUALITATIVE STUDIES]

Most studies previously conducted establish fact that re-use is exceptionally difficult, regardless of domain affiliation or familiarity with conceptual data practice

(Cragin and Shankar 2006, Jirotka et al., 2005; Zimmerman 2008)

Re-use is often context dependent:

- Gathering and recording standards are important but rarely sufficient

  (Zimmerman 2007)

- Documentation can both help…

  (Carlson & Anderson, 2007; Faniel & Jacobsen, 2010)

- …or be insufficient for accurate interpretation

  (Birnholtz & Bietz, 2003; Shankar, 2007)

# [INFORMATION USE & USERS -> DATA USE & USERS]

Lack of immediately intelligible semantic content and the rawness of data as a research product inhibits us from treating this is a purely "information behavior" study.

Studying re-use might include

- Tracking re-use
- Measuring impact of re-use
- Behavior of re-use*

# [STUDYING VARIOUS REUSE BEHAVIORS]

## Comparative Approach

### Environmental science

- Small teams or single researchers conducting hypothesis driven research. Data is gathered in remote 'field' cites through observations or instruments like networked censors.

### Social Science

- Also small or individual research teams involved in field observation for data gathering, as well as purely statistical analysis of quantitative data.

# [METHOD]

- Snap-Shot / Baseline
  - Performed literature search through three portals:
    - ISI Web of Science
    - Scopus
    - Google Scholar
  - Temporal range: 2001-2010
  - Some search queries used:
    - *'Data sharing'*
    - *'Data reuse'*
    - *'Secondary data analysis'*
    - *'Reanalysis'*
- (Generic search- breadth not depth)

# [QUESTIONS]

Who re-uses data (in these domains), What products do they re-use (and privilege), Where do they find data to re-use

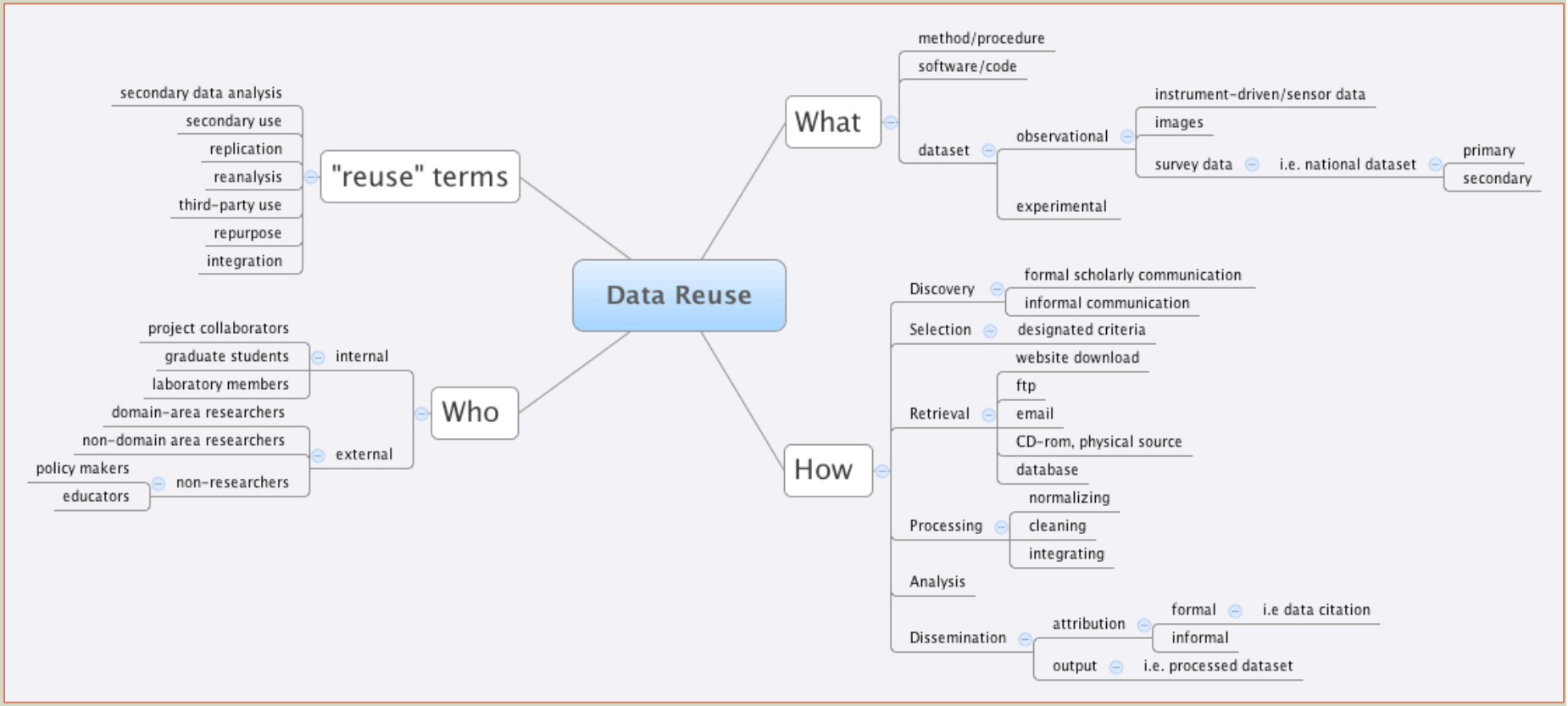
## Analysis Focused On

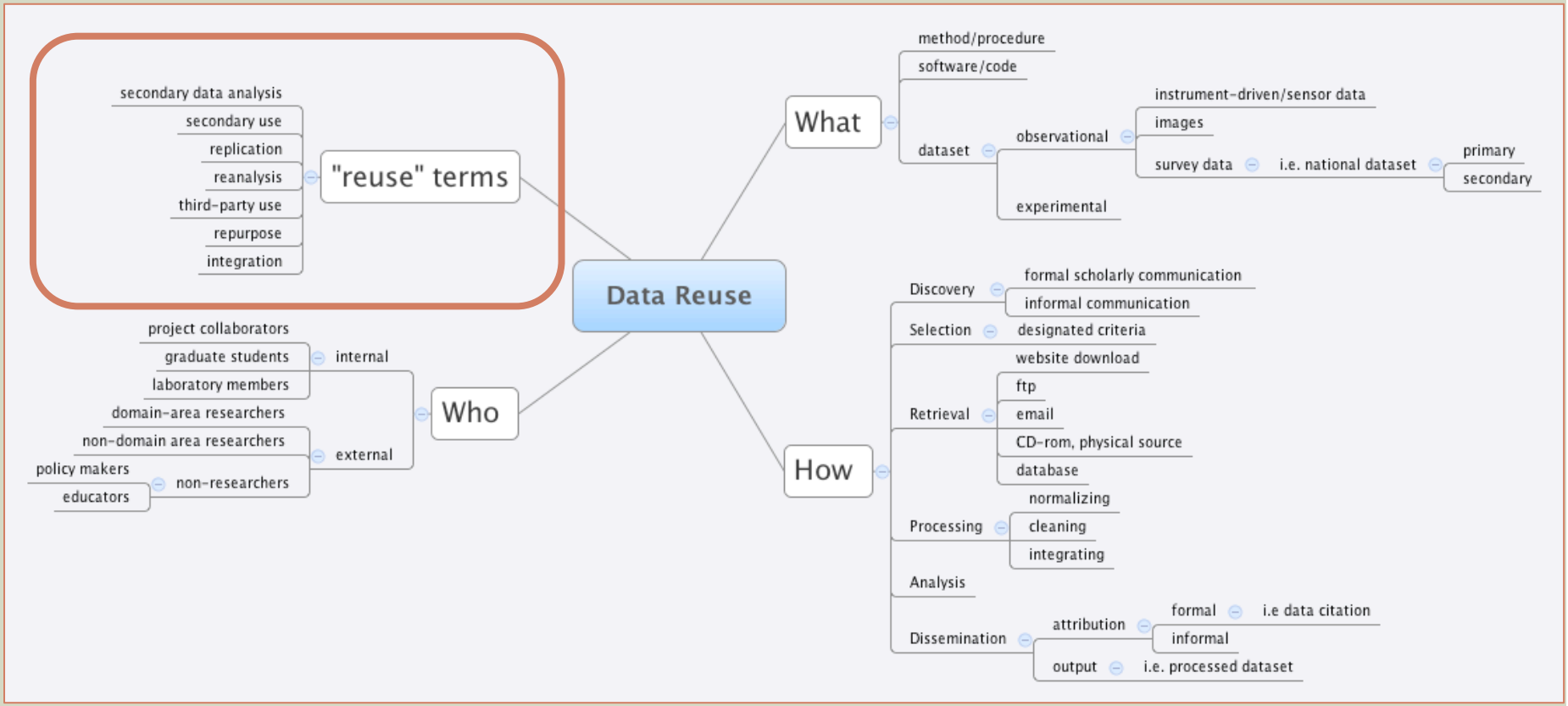
What are the similarities and differences in re-use activities for Environmental Science and Social Science?

What is the discovery, access and transformation necessary for 'successful' reuse?

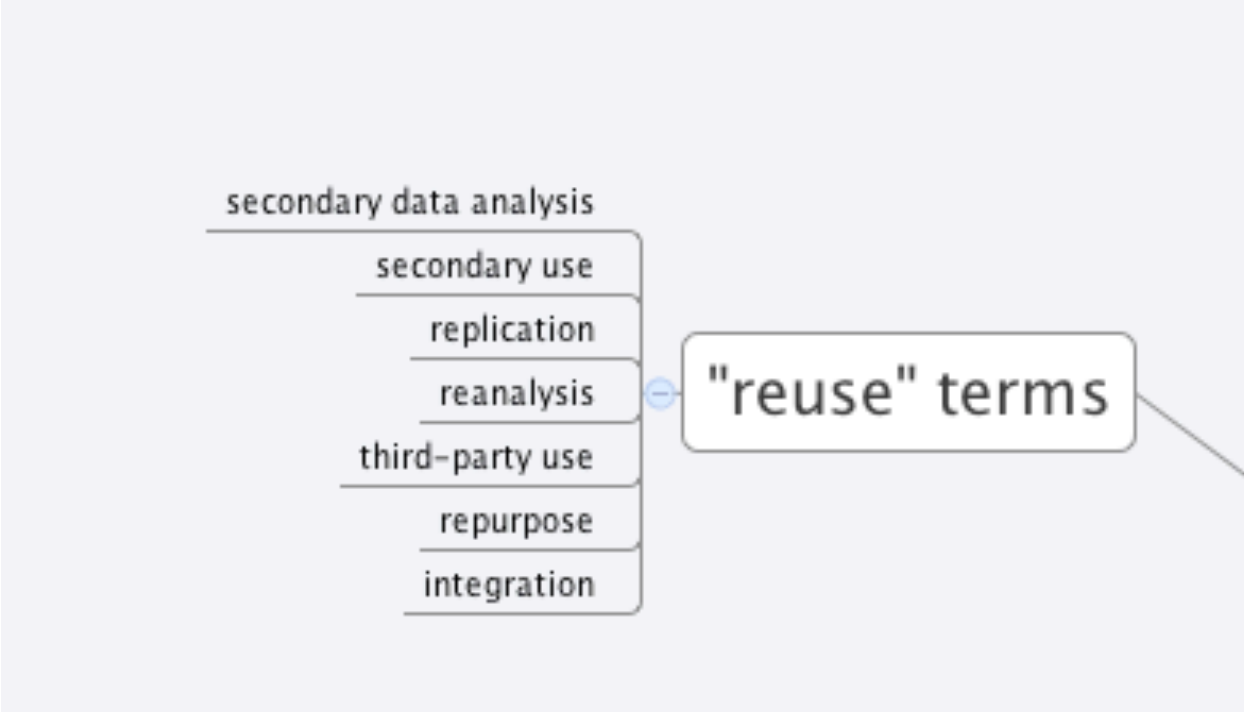
How can we recognize re-use (how do the research communities describe or acknowledge re-use?

# [Data Reuse Framework]




Data Reuse
"reuse" terms
secondary data analysis
secondary use
replication
reanalysis
third-party use
repurpose
integration
Who
internal
project collaborators
graduate students
laboratory members
external
domain-area researchers
non-domain area researchers
non-researchers
policy makers
educators
What
method/procedure
software/code
dataset
observational
instrument-driven/sensor data
images
survey data
i.e. national dataset
primary
secondary
experimental
How
Discovery
formal scholarly communication
informal communication
Selection
designated criteria
Retrieval
website download
ftp
email
CD-rom, physical source
database
Processing
normalizing
cleaning
integrating
Analysis
Dissemination
attribution
formal
i.e data citation
informal
output
i.e. processed dataset

# What's in a name?
## Data "reuse" in situ




## DATA REUSE FRAMEWORK

How can we recognize re-use?

How does the research community describe or acknowledge reuse?

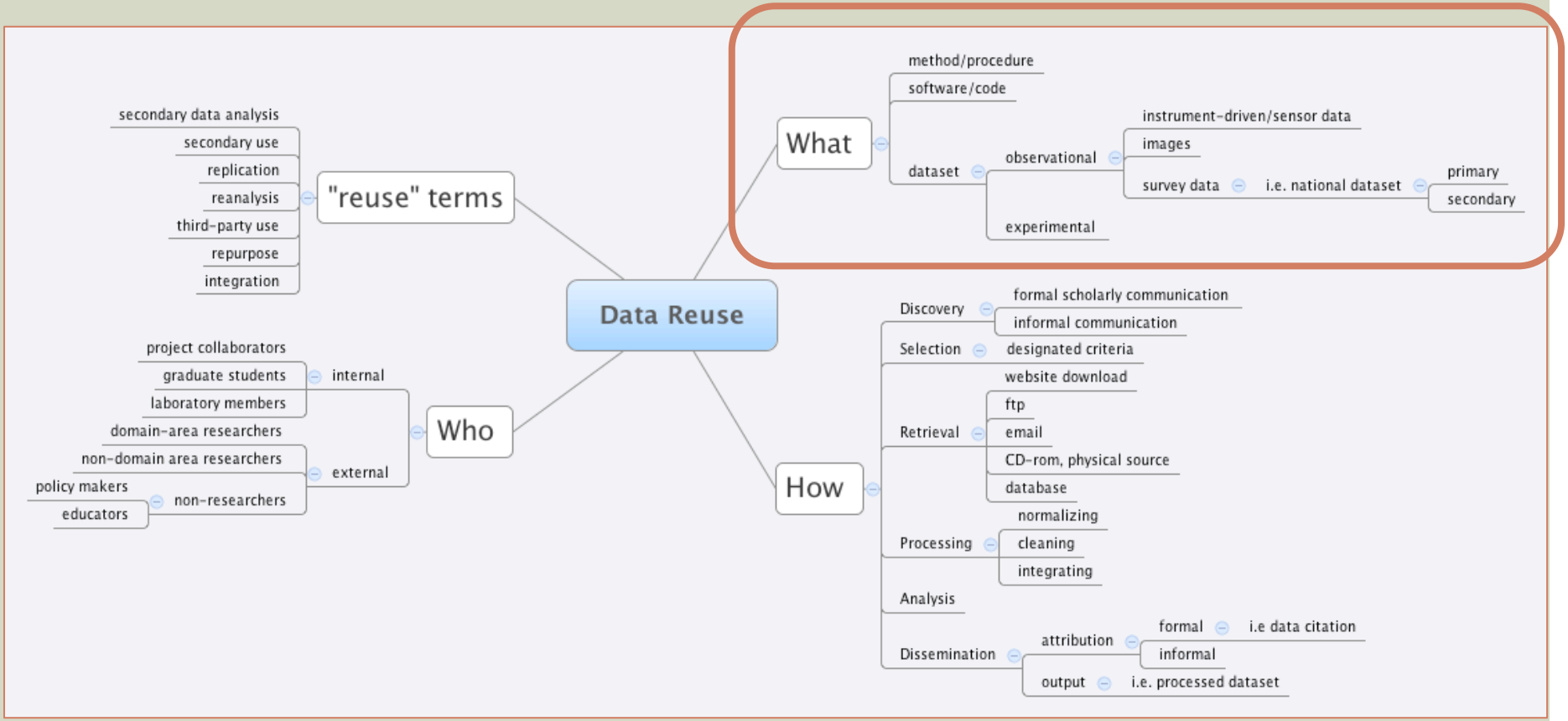
# [DEFINITIONS AND VARIATIONS ON DATA "REUSE"]

- No consistent term used across domains
  - Social science: 'secondary data analysis'

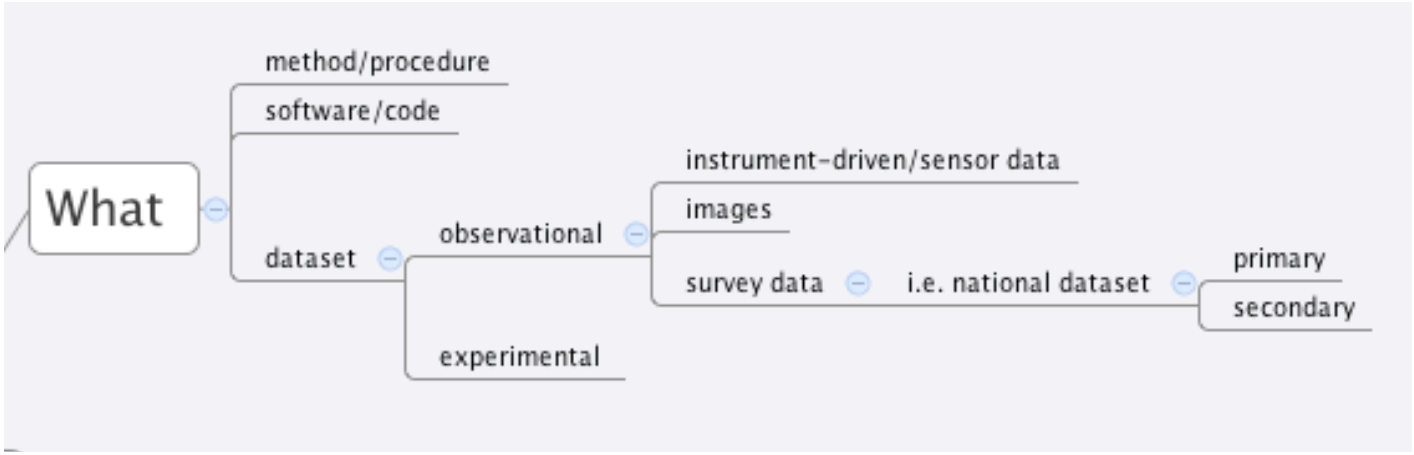
    *"the analysis of data for a different purpose than what the data were originally collected for, possibly by the original data producers themselves, or in collaboration with other people, or by entirely different people."* *(Niu, 2009, p.4)*
    - other terms used: 'reanalysis', 'replication'

  - Environmental science: 'secondary use'

    *"the use of data collected for one purpose to study a new problem"* (Zimmerman, 2003, p. 7)
    - Other terms used: 'integration', 'repurpose'


Data Reuse
"reuse" terms
secondary data analysis
secondary use
replication
reanalysis
third-party use
repurpose
integration
Who
internal
project collaborators
graduate students
laboratory members
external
domain-area researchers
non-domain area researchers
non-researchers
policy makers
educators
What
method/procedure
software/code
dataset
observational
instrument-driven/sensor data
images
survey data
i.e. national dataset
primary
secondary
experimental
How
Discovery
formal scholarly communication
informal communication
Selection
designated criteria
Retrieval
website download
ftp
email
CD-rom, physical source
database
Processing
normalizing
cleaning
integrating
Analysis
Dissemination
attribution
formal
i.e data citation
informal
output
i.e. processed dataset

# What data are reused?

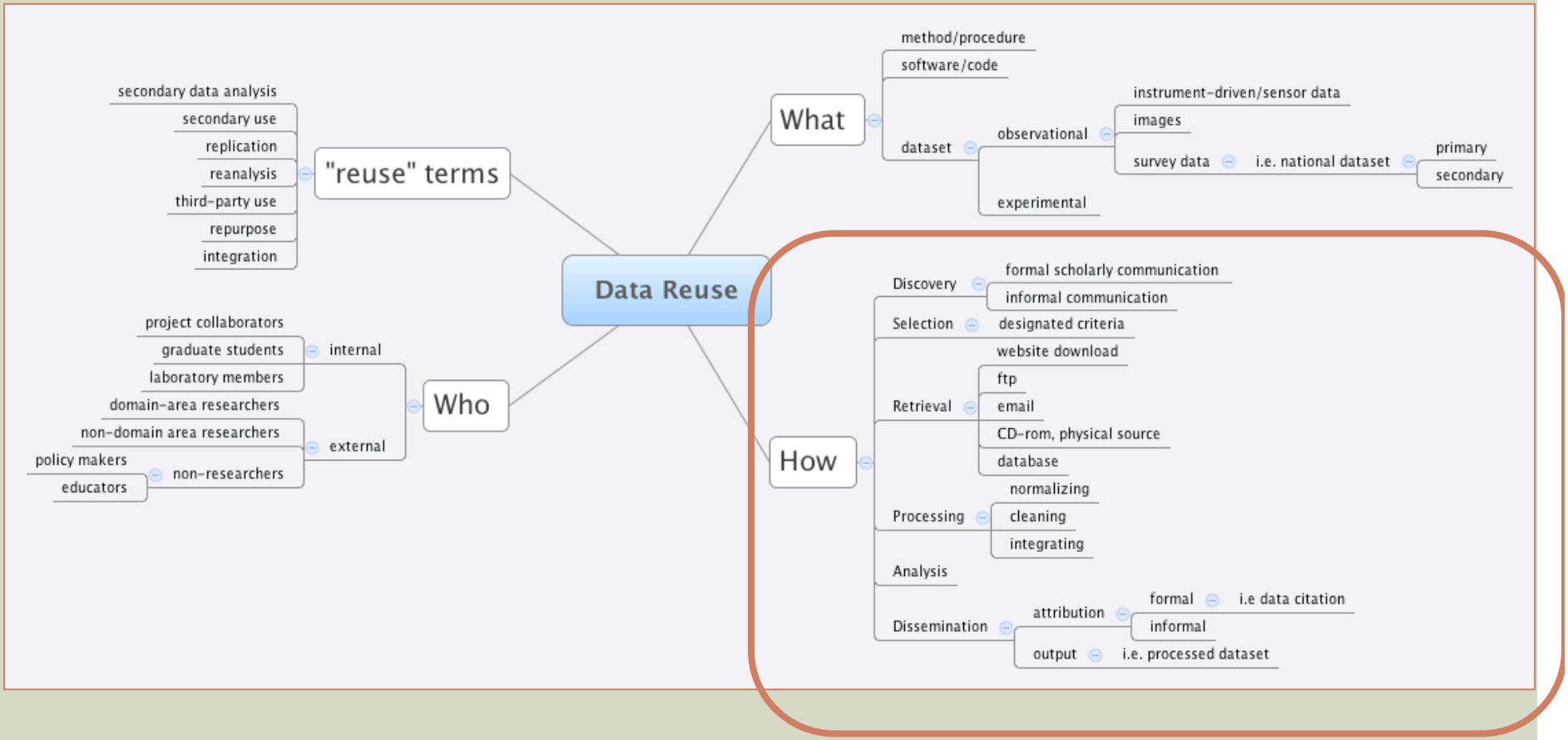



## DATA REUSE FRAMEWORK

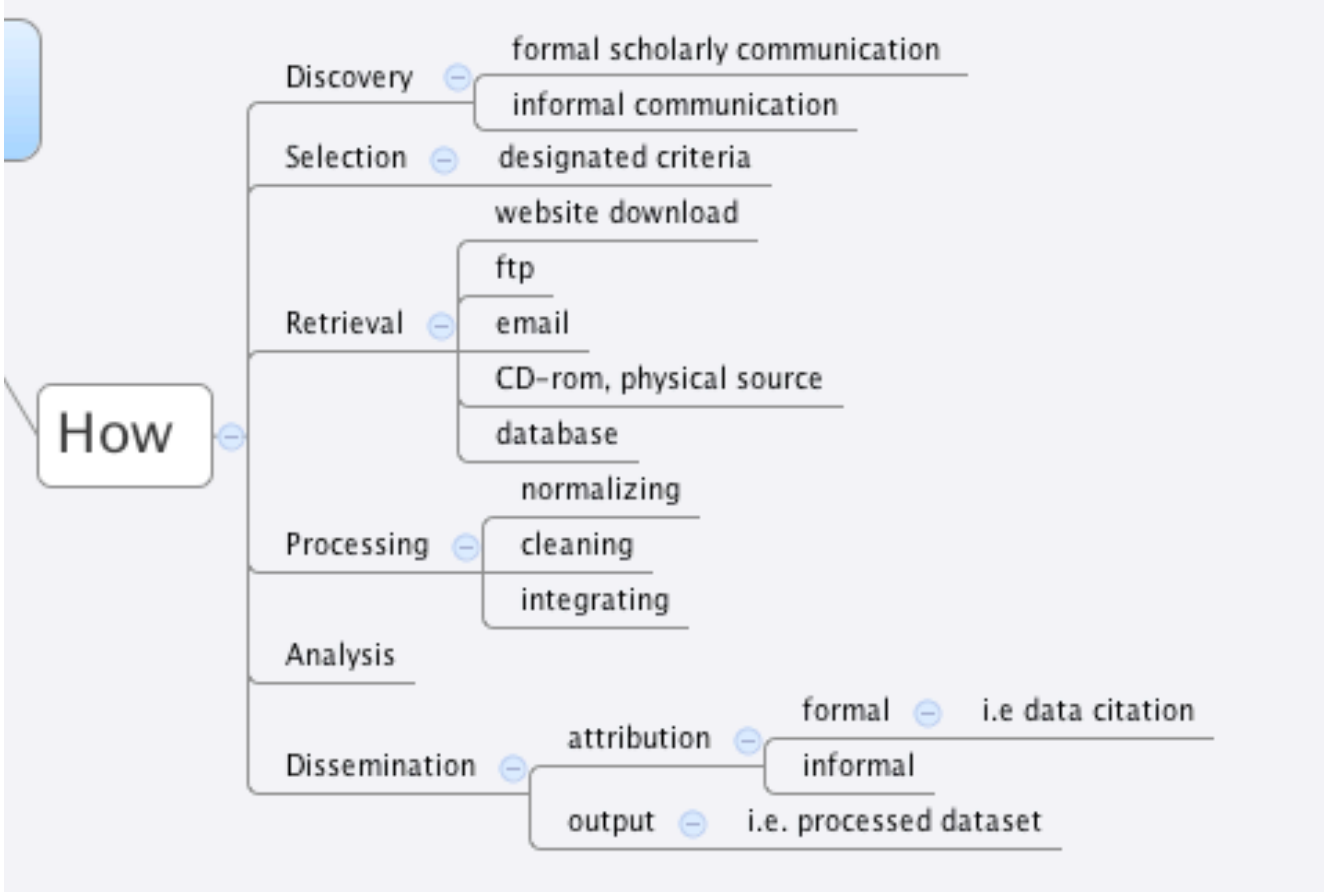
What products are reused (and privileged)?

# ['WHAT' IS REUSED]

- Patterns and characteristics of reused data:
  - Quantitative datasets
    - Types:
      - Census data
      - Longitudinal studies (i.e. Add Health)
      - Instrument-driven data (i.e. embedded sensors, satellite)
    - Sources acknowledged:
      - Data archives and Data centers (i.e. ICPSR, NSIDC)
- Growing interest in the reuse of qualitative data (social sciences)
- Other information sources integral to data reuse


Data Reuse
"reuse" terms
secondary data analysis
secondary use
replication
reanalysis
third-party use
repurpose
integration
Who
internal
project collaborators
graduate students
laboratory members
external
domain-area researchers
non-domain area researchers
non-researchers
policy makers
educators
What
method/procedure
software/code
dataset
observational
instrument-driven/sensor data
images
survey data
i.e. national dataset
primary
secondary
experimental
How
Discovery
formal scholarly communication
informal communication
Selection
designated criteria
Retrieval
website download
ftp
email
CD-rom, physical source
database
Processing
normalizing
cleaning
integrating
Analysis
Dissemination
attribution
formal
i.e data citation
informal
output
i.e. processed dataset

# How does reuse happen?



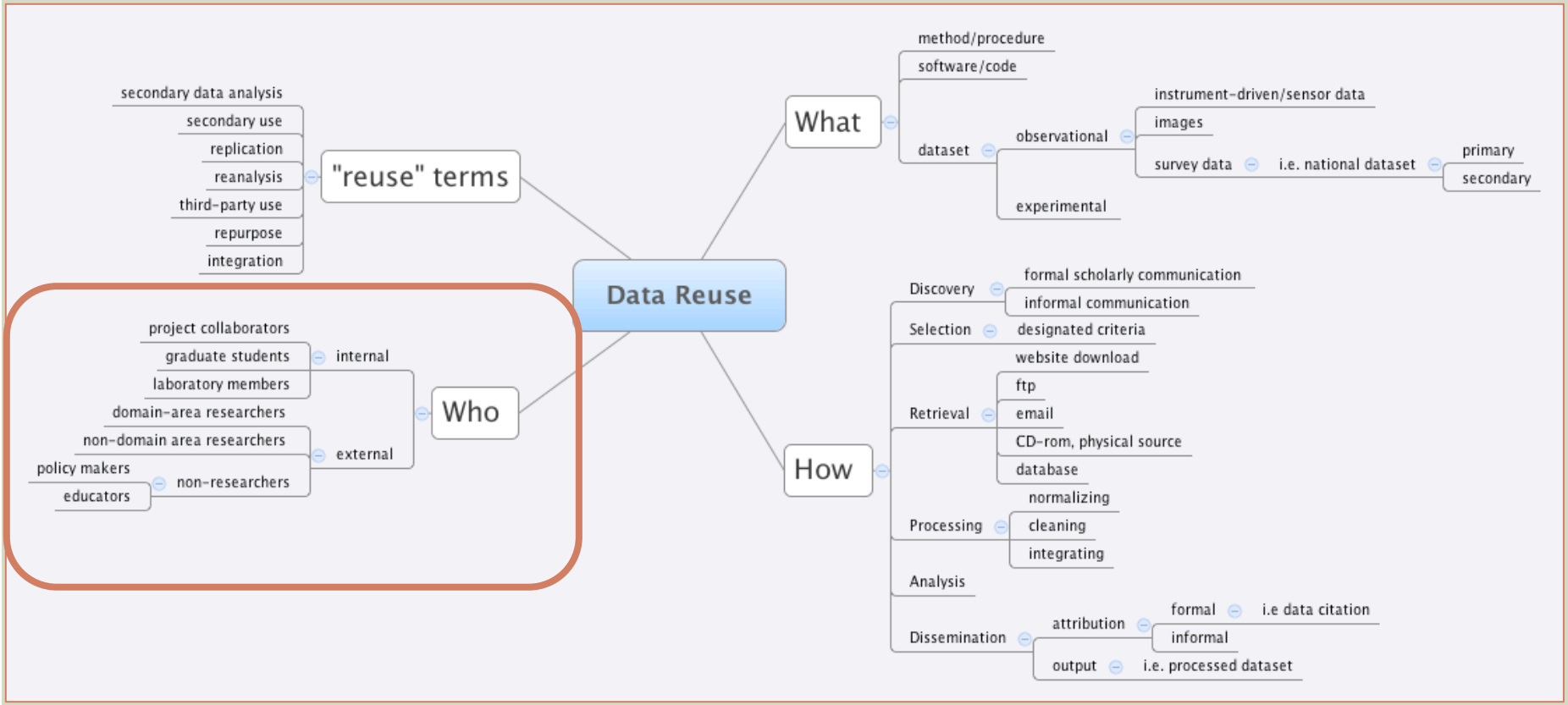

## DATA REUSE FRAMEWORK

What activities and processes do we observe?

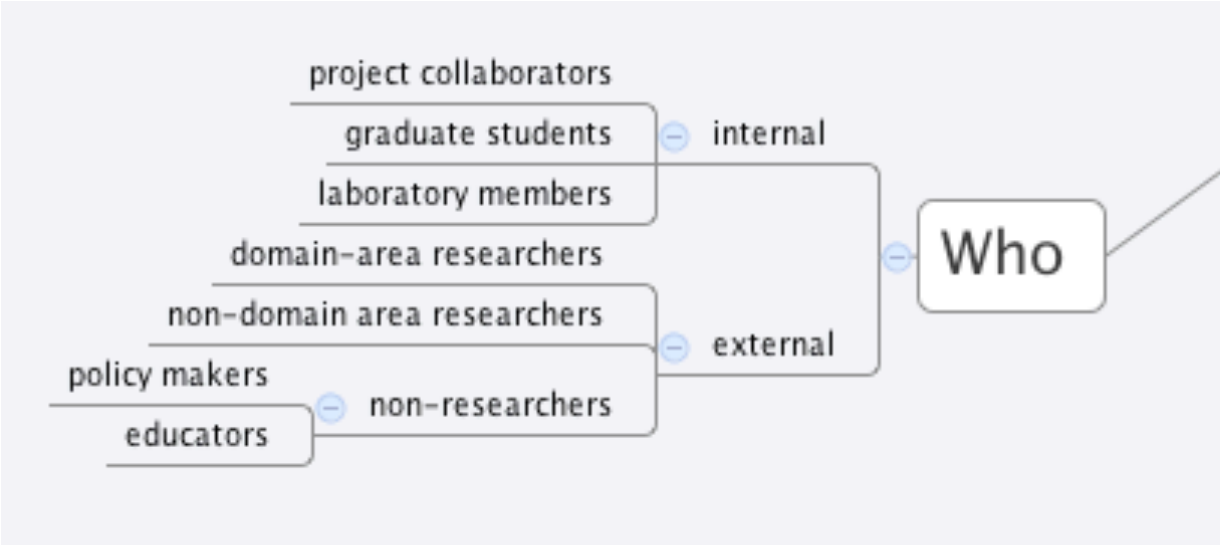
# [REUSE IN PRACTICE]

- ‘data in the rough’: a rare occurrence in the literature

- [discovery]
- [selection]
- [retrieval]
- [processing]- activity unique to reuse: requires contextual information
- [analysis]
- [dissemination]


Data Reuse
"reuse" terms
secondary data analysis
secondary use
replication
reanalysis
third-party use
repurpose
integration
Who
internal
project collaborators
graduate students
laboratory members
external
domain-area researchers
non-domain area researchers
non-researchers
policy makers
educators
What
method/procedure
software/code
dataset
observational
instrument-driven/sensor data
images
survey data
i.e. national dataset
primary
secondary
experimental
How
Discovery
formal scholarly communication
informal communication
Selection
designated criteria
Retrieval
website download
ftp
email
CD-rom, physical source
database
Processing
normalizing
cleaning
integrating
Analysis
Dissemination
attribution
formal
i.e data citation
informal
output
i.e. processed dataset

# Who is reusing data?




## DATA REUSE FRAMEWORK

Who participates in reuse activities?

# [USER COMMUNITIES]

- Internal vs. external communities for data reuse
  - Social science (psychology)

*"In spite of the original investigators' efforts to make the data available, only investigators attached to the original research team in some way submitted articles."* (Foster, 2010, p.974).

# [DISCUSSION]

- "reuse" definitions: fuzzy boundaries
- Acknowledgements are inconsistent, with the exception of quantitative data from established data centers and archives
- in general, the practice of data reuse not well described in the literature
  - Reuse contingent on external factors including adequate metadata (i.e. documentation) and usability of data (i.e. compatibility of formats, damaged data)

# [ DISCUSSION (CONT'D) ]

- Data reusers: an invaluable knowledge source
- Limitations of literature search: very few journal articles that discuss practice
  - Agency reports and dissertations extremely fruitful
  - Position papers/ opinion pieces dominate in search results

# [FUTURE DIRECTIONS]

- Continue to update the DR framework and test against other domains
- Framework provides an organizational mechanism for the study of research and scholarly practice
  - Reveals gaps in research

# Thank you!

Nicholas Weber

nmweber@illinois.edu

@nniiicc

Tiffany Chao

tchao@illinois.edu

# [REFERENCES]


Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: Supporting sharing in science and engineering. *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work,* pp. 339-343.

Carlson, S., & Anderson, B. (2007). What are data? The many kinds of data and their implications for reuse. *Journal of Computer-Mediated Communication* 12(2), article 15.

Cragin, M. H., & Shankar, K. (2006). Scientific data collections and distributed collective practice. *Computer Supported Cooperative Work* 15(2-3), 185-204.

Faniel, I. M., & Jacobsen, T. E. (2010.) Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work* 19(3-4), 355-375.

Foster, E. M. (2010). The value of reanalysis and replication: Introduction to special section. *Developmental psychology*, *46*(5), 973.

# [REFERENCES]


Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., Hinds, C., & Voss, A. (2005). Collaboration and trust in healthcare innovation: The eDiaMoND case study. *Computer Supported Cooperative Work* 14, 369-398.

Niu, J. (2009). Perceived documentation quality of social science data. Dissertation.

Zimmerman, A.S. (2003). Data sharing and secondary use of scientific data: experiences of ecologists. Dissertation.

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries* 7(1-2), 5-16.