NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Wee Kim Wee School of Communication and Information

College of Humanities, Arts, and Social Sciences

# Social Science Quantitative Research Data Reuse: A Data Repository Perspective

**SUN Guangyuan (Dr)**

Research Data Librarian

Library

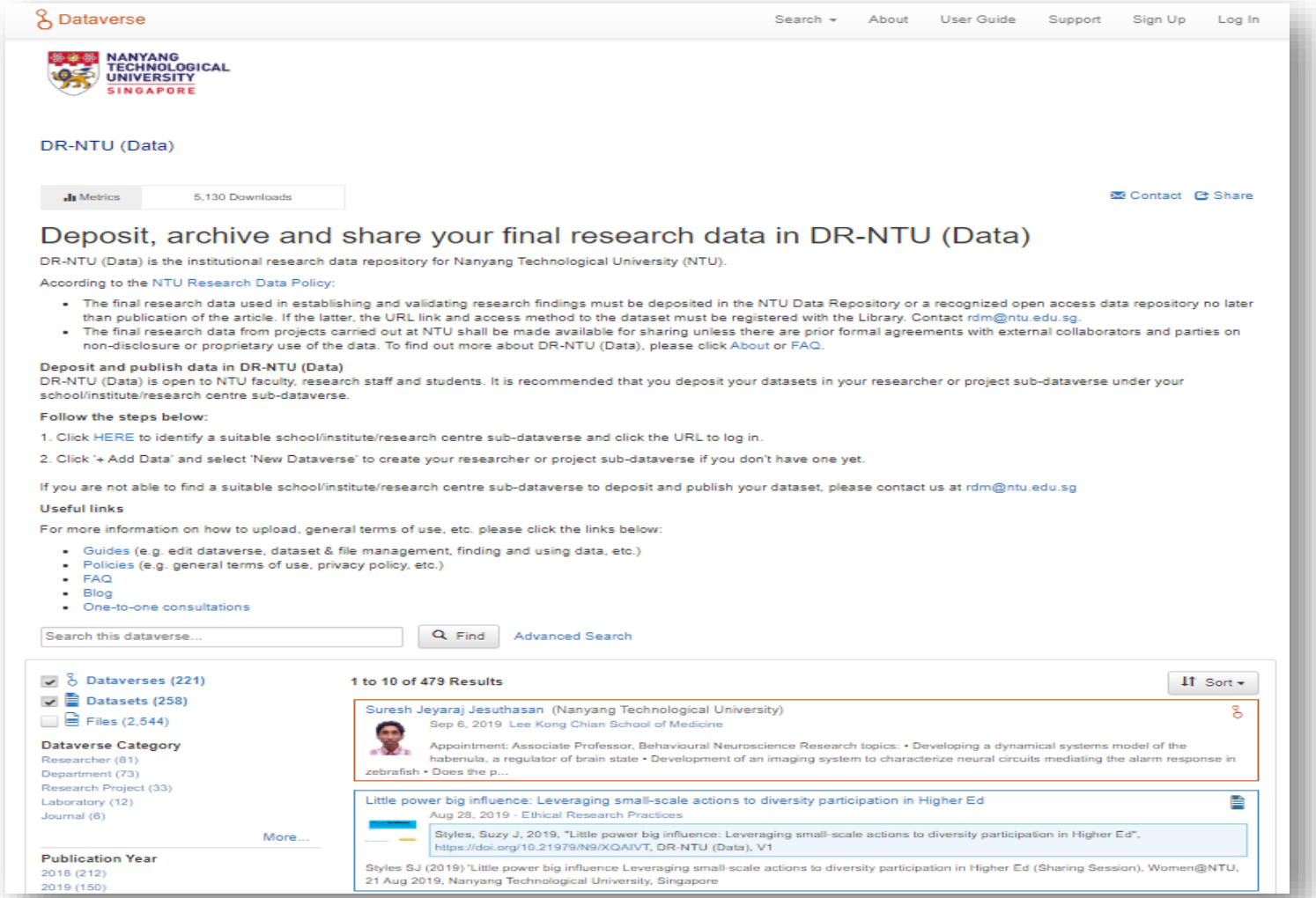
Nanyang Technological University, Singapore

17-Dec-2020

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# The trend of data curation & research data management

- NTU Policy – Scholars file Data Management Plan for funded projects & deposit research data in NTU data repository
  - To protect **research integrity** – e.g., in cases of accusation of data falsification, there is data for review
  - 2nd purpose – allow **reuse of data** beyond original study design, to exploit maximum value out of exiting data



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# The trend (cont.)

## Biological sciences

- DNA DataBank of Japan (DDBJ)
- European Nucleotide Archive (ENA)
- GenBank
- dbSNP
- European Variation Archive (EVA)
- dbVar
- EBI Metagenomics
- NCBI Trace Archive
- NCBI Sequence Read Archive (SRA)
- NCBI Assembly

## Earth, Environmental and Space sciences

- NASA Goddard Earth Sciences Data and Information Services Center
- NERC Data Centres
- SIMBAD Astronomical Database
- UK Solar System Data Centre

## Materials science

- NoMaD Repository
- Materials Cloud

## Social sciences

- Archaeology Data Service
- Harvard Dataverse
- openICPSR
- Open Science Framework
- Qualitative Data Repository
- UK Data Service

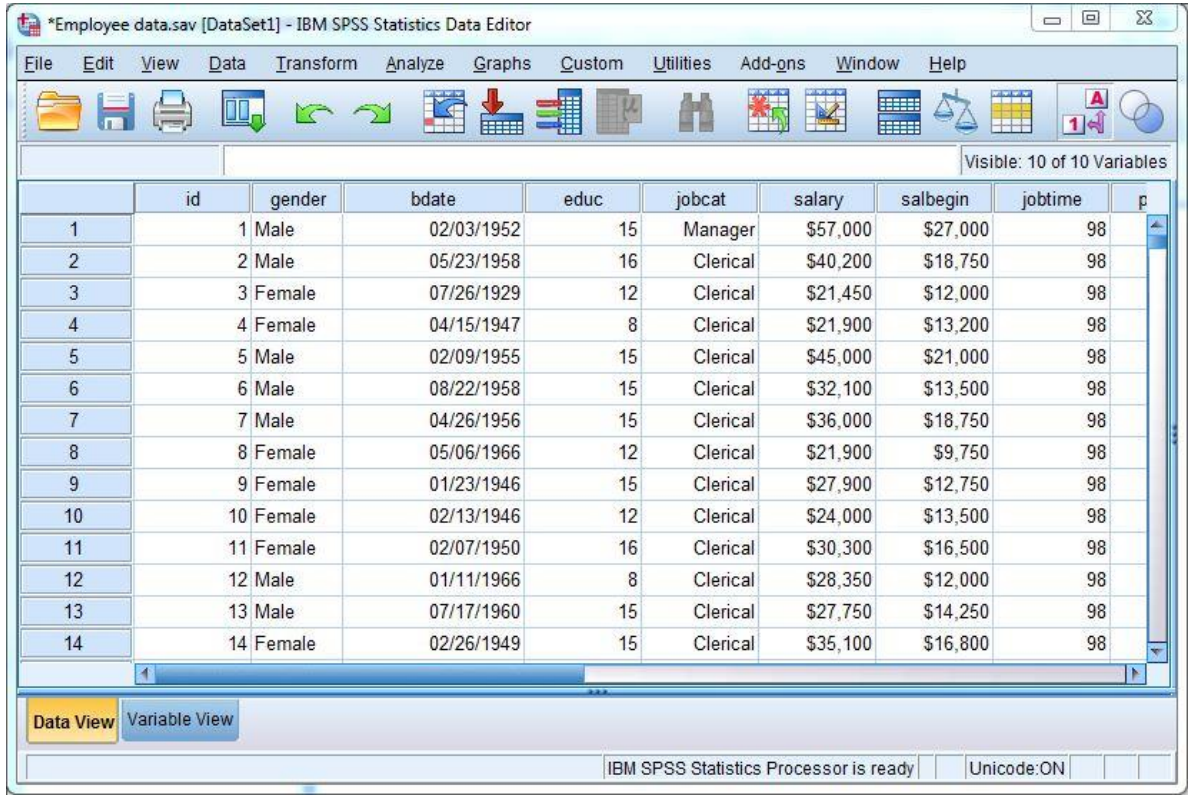
- ❑ A trend of **data curation & research data management**
- ❑ Data repositories—share & reuse
  - ➢ *Do the repositories really support data reuse?*
  - ➢ *How to design a repository to better support reuse?*

*Source*: Retrieved from the webpage “Recommended Data Repositories” https://www.nature.com/sdata/policies/repositories#envgeo

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Quantitative Social Science Research Data Set

- My research interest →
- From questionnaire survey
- Tabular form (columns, rows)
- To support reuse of such data




| | id | gender | bdate | educ | jobcat | salary | salbegin | jobtime | p |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Male | 02/03/1952 | 15 | Manager | $57,000 | $27,000 | 98 | |
| 2 | 2 | Male | 05/23/1958 | 16 | Clerical | $40,200 | $18,750 | 98 | |
| 3 | 3 | Female | 07/26/1929 | 12 | Clerical | $21,450 | $12,000 | 98 | |
| 4 | 4 | Female | 04/15/1947 | 8 | Clerical | $21,900 | $13,200 | 98 | |
| 5 | 5 | Male | 02/09/1955 | 15 | Clerical | $45,000 | $21,000 | 98 | |
| 6 | 6 | Male | 08/22/1958 | 15 | Clerical | $32,100 | $13,500 | 98 | |
| 7 | 7 | Male | 04/26/1956 | 15 | Clerical | $36,000 | $18,750 | 98 | |
| 8 | 8 | Female | 05/06/1966 | 12 | Clerical | $21,900 | $9,750 | 98 | |
| 9 | 9 | Female | 01/23/1946 | 15 | Clerical | $27,900 | $12,750 | 98 | |
| 10 | 10 | Female | 02/13/1946 | 12 | Clerical | $24,000 | $13,500 | 98 | |
| 11 | 11 | Female | 02/07/1950 | 16 | Clerical | $30,300 | $16,500 | 98 | |
| 12 | 12 | Male | 01/11/1966 | 8 | Clerical | $28,350 | $12,000 | 98 | |
| 13 | 13 | Male | 07/17/1960 | 15 | Clerical | $27,750 | $14,250 | 98 | |
| 14 | 14 | Female | 02/26/1949 | 15 | Clerical | $35,100 | $16,800 | 98 | |

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Social Science Data Repositories

- Inter-university Consortium for Political and Social Research (1962)
- UK Data Archive (1966)
- Consortium of European Social Science Data Archives (1976)

## Social Science Data Repositories

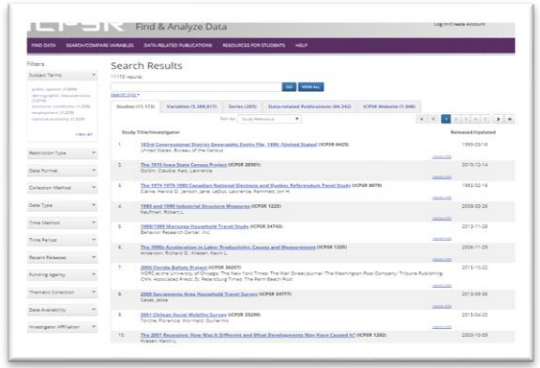
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

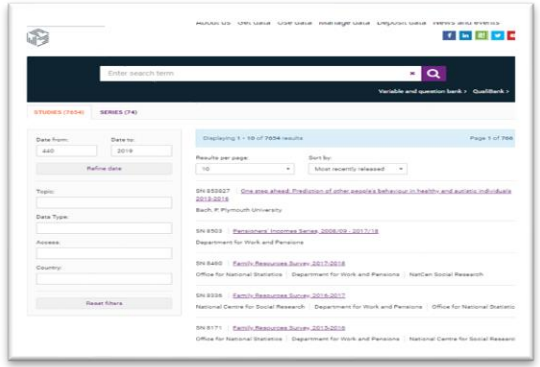
**Filters to refine search results**

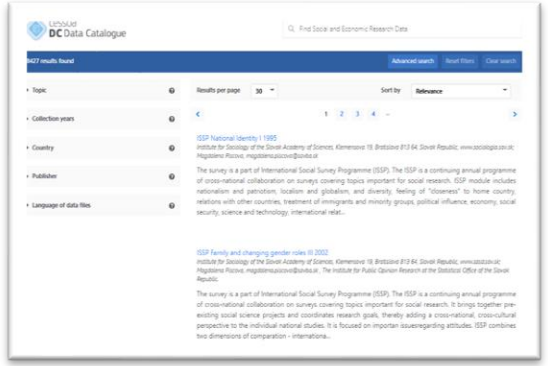
- Filter 1
- Filter 2
- Filter 3
- …

**Summarized Data Sets Search Results**

1. Data set A surrogate record
2. Data set B surrogate record
3. Data set C surrogate record
4. Data set D surrogate record
5. Data set E surrogate record
…





## Social Science Data Repositories

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

- Topic (e.g., society and culture; health)
- Data type (e.g., survey data; experimental data)
- Data accessibility (e.g., open; controlled)
- Country of data collection
- …

**Filters to refine search results**

- Filter 1
- Filter 2
- Filter 3
- …

**Summarized Data Sets Search Results**

1. Data set A surrogate record
2. Data set B surrogate record
3. Data set C surrogate record
4. Data set D surrogate record
5. Data set E surrogate record
…

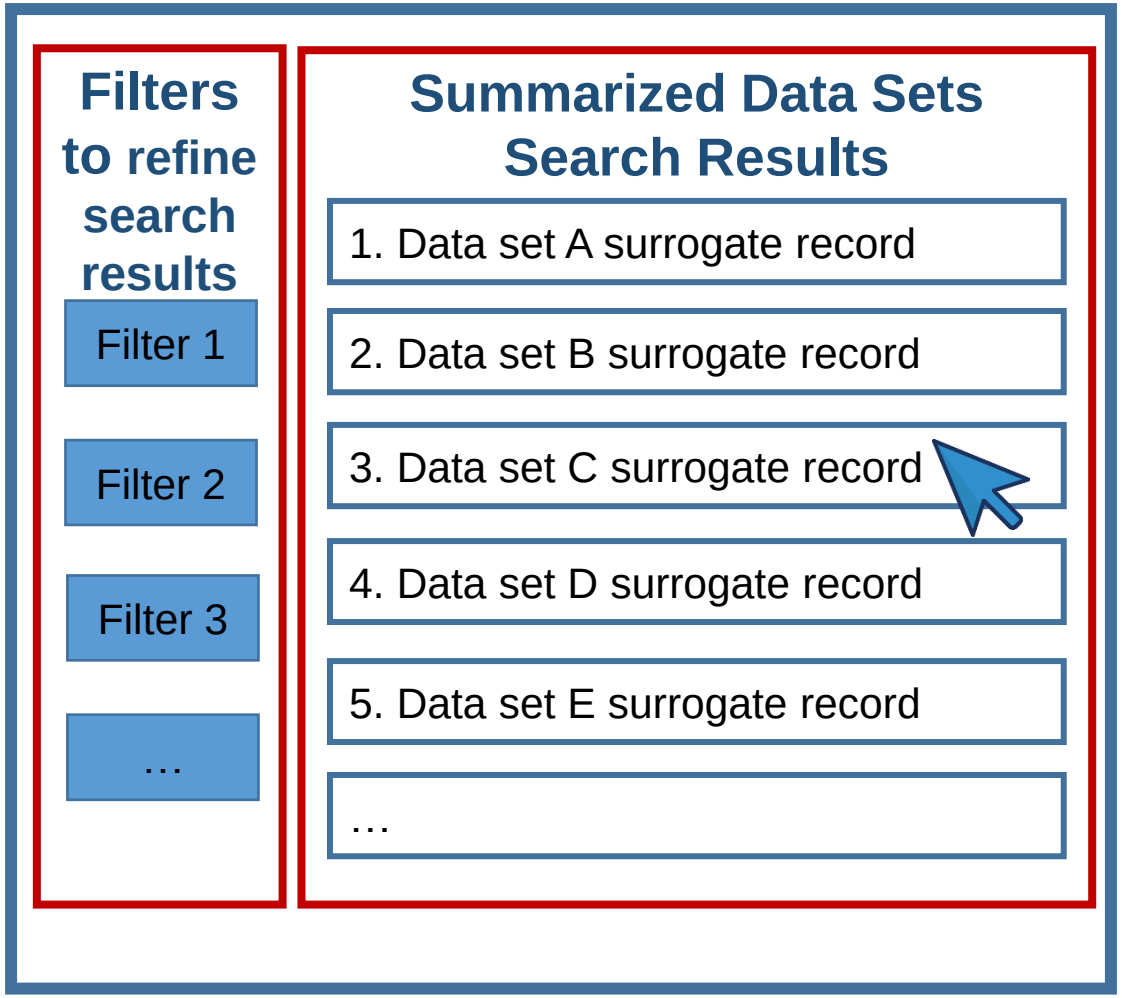
Metadata elements for **each** data set surrogate records

- Dataset title (e.g., study name)
- Data collector(s) name (e.g., study's principle investigators)
- Data collector(s)' affiliated institution

Sun, G., & Khoo, C. S. G. (2017). Social science research data curation: issues of reuse. *Libellarium: journal for the research of writing, books, and cultural heritage institutions*, *9*(2).

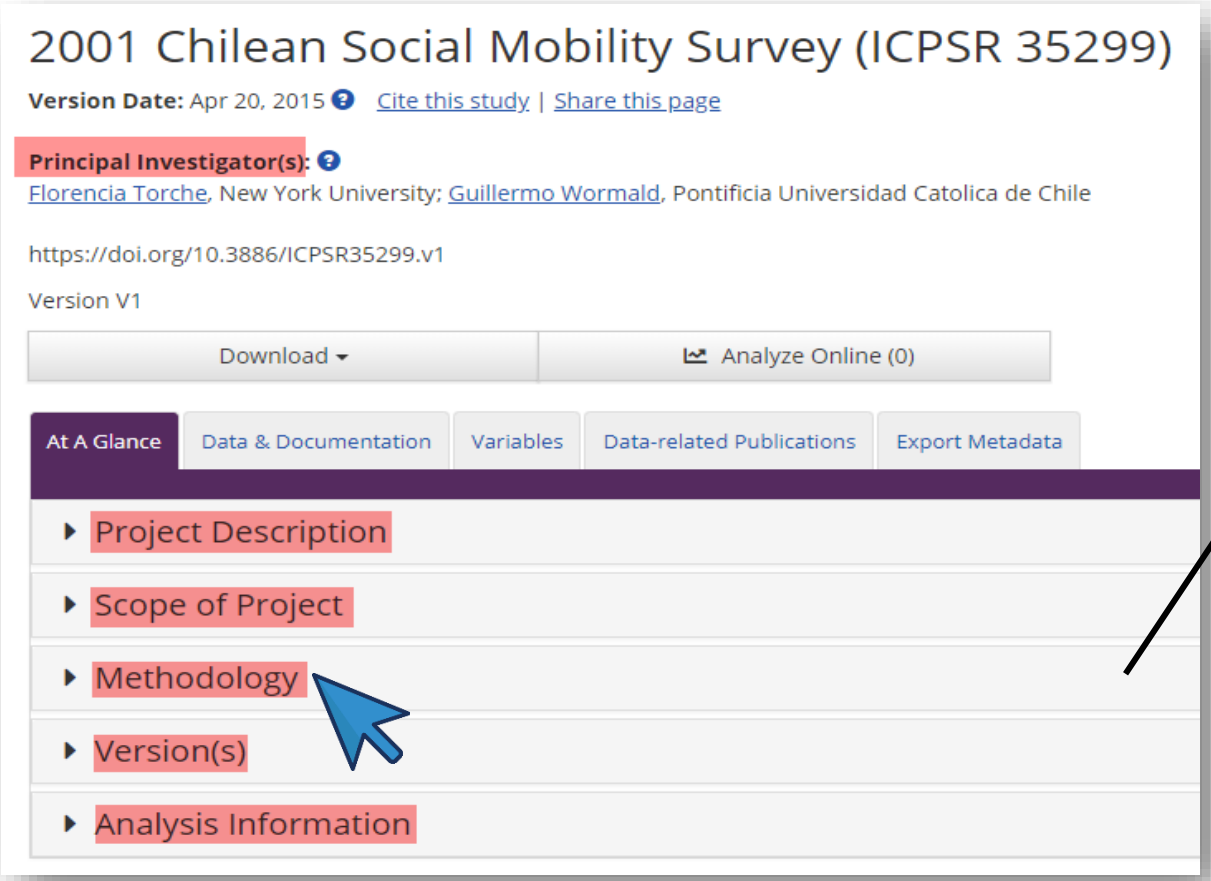
Social Science Data Repositories

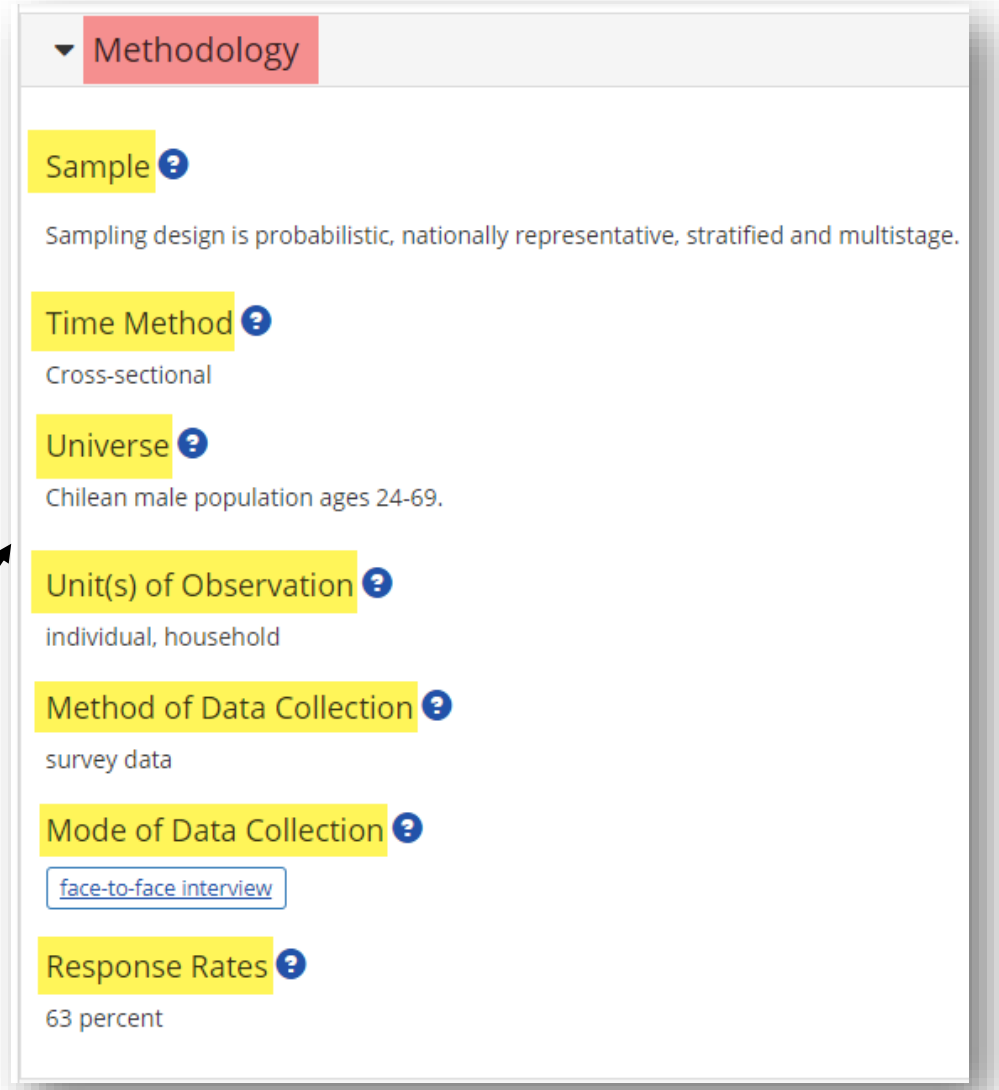
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

## Individual Detailed Surrogate Record



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Semantic Challenges of Data Reuse

- ***What is your sex?***

**1. Male**

**2. Female**

| | A | B | C | D |
|---|---|---|---|---|
| 1 | ID | Sex | Att1 | Att2 |
| 2 | 1 | 1 | 2 | 1 |
| 3 | 2 | 2 | 3 | 2 |
| 4 | 3 | 1 | 5 | 5 |
| 5 | 4 | 1 | 1 | 4 |
| 6 | 5 | 2 | 4 | 2 |
| 7 | 6 | 2 | 3 | 99 |

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Semantic Challenges of Data Reuse

- ***Which of the options best describes how you think of yourself?***

**1. Heterosexual or Straight,**

**2. Gay or Lesbian,**

**3. Bisexual,**

**4. Other**

| | A | B | C | D |
|---|---|---|---|---|
| 1 | ID | Sex | Att1 | Att2 |
| 2 | 1 | 1 | 2 | 1 |
| 3 | 2 | 2 | 3 | 2 |
| 4 | 3 | 1 | 5 | 5 |
| 5 | 4 | 1 | 1 | 4 |
| 6 | 5 | 2 | 4 | 2 |
| 7 | 6 | 2 | 3 | 99 |

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Semantic Challenges of Data Reuse

- Thus, **variables** with the same name may refer to different concepts.
- The semantics of **categorical values** may not be obvious to users.
- Relationships between variables may not be readily apparent (e.g., dummy coding).

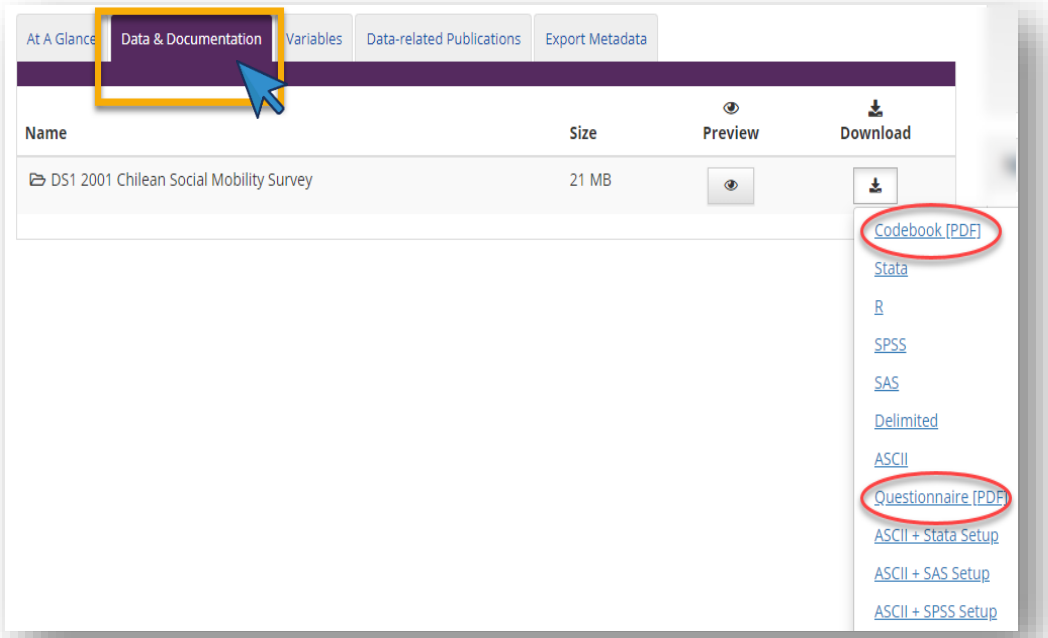
| | A | B | C | D |
|---|---|---|---|---|
| 1 | Chinese | Malay | Indian | Others |
| 2 | 1 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 |
| 5 | 0 | 0 | 0 | 1 |

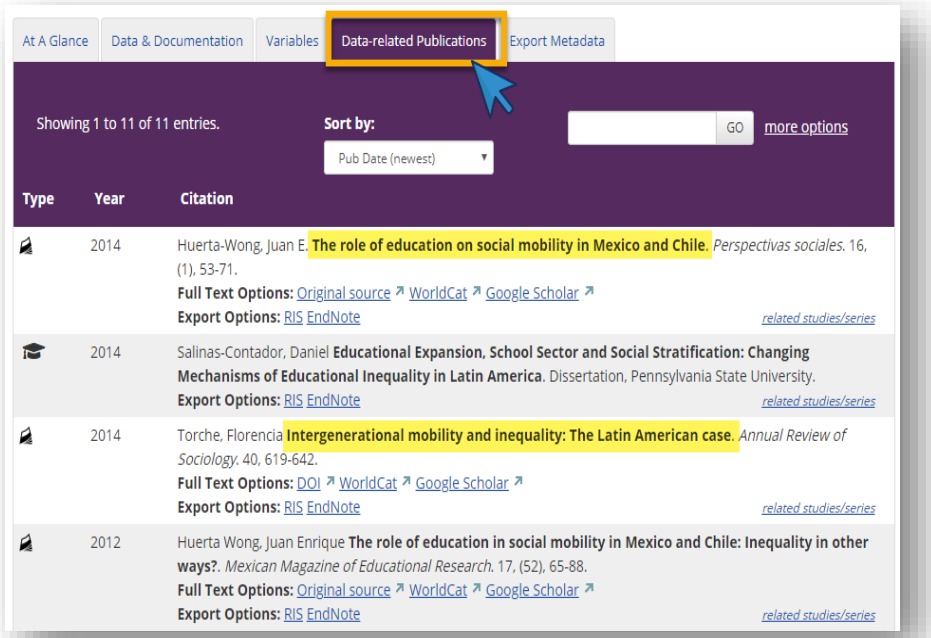
| E |
|---|
| Race |
| 1 |
| 2 |
| 3 |
| 4 |

Social Science Data Repositories

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

## Individual Detailed Surrogate Record




*Current social science data repository*: users need to download supporting documentations (usually lengthy PDF files) **individually and separately** that help them to understand and reuse data sets.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

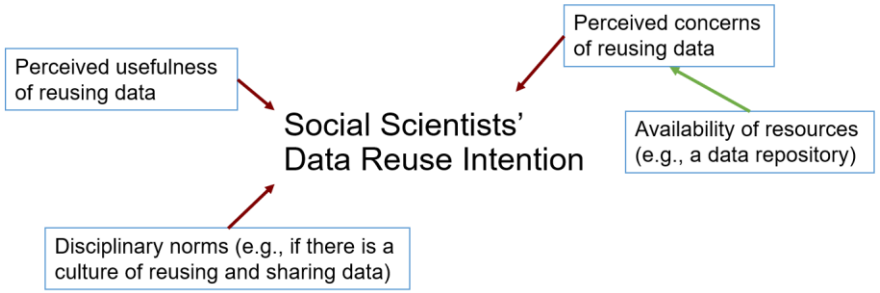
- Mischo et al.'s (2014) found that researchers generally underutilized domain-specific data repositories in their disciplines.
- This includes the social sciences data repositories.

Mischo, W.H., Schlembach, M.C. and O'Donnell, M.N. (2014). An analysis of data management plans in University of Illinois National Science Foundation grant proposals. *Journal of eScience Librarianship, 3*(1), Article-3.

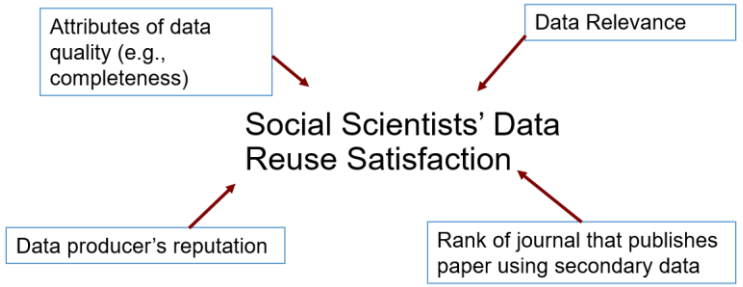
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Prior Studies on Social Science Data Reuse

Yoon & Kim (2017)




Faniel, Kriesberg & Yakel (2016)




Lack of literature on social scientists' data reuse behaviour in relation to **data repository systems**.

- How: *browse, search, and evaluate data reusability in a data repository*
- What: *challenges and unsupported needs*

Yoon (2017)

Social Scientists' Data Reuse Trust Development Stages

1. Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library & Information Science Research*, 39(3), 224-233.
2. Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology, 67*(6), 1404-1416.
3. Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology, 68*(4), 946-956.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Research Objectives

1. to identify the user requirements of social scientists for a data repository system to support the reuse of curated quantitative social science research data.
2. to develop a knowledge representation system for the support of data curation of quantitative social science research data to support data reuse.

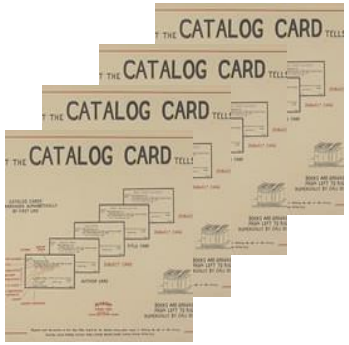
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Three Studies

## Study 2

### Knowledge representation system design & development

To design **metadata and ontology** to support data reuse in social science data repository systems



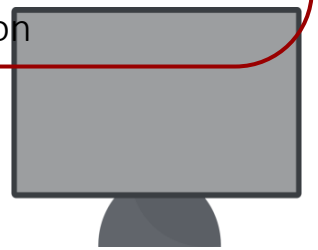

Data set records

## Study 3

### User evaluation study

To test the usability of the developed knowledge representation system to support social scientists data reuse

- ❑ **Prototype system design**
- ❑ **Task-based** user evaluation



Interface (the user-end of a data repository system)

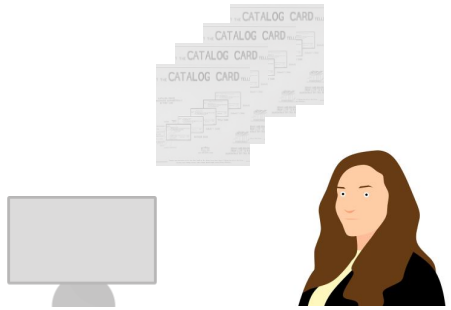
Social scientists

## Study 1

### Semi-structured interview study to collect user requirements

Understand social scientists' data reuse behaviour when interacting with a data repository

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE



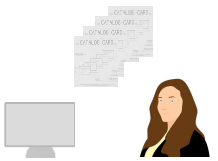

# Study 1

## Semi-structured interview study

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Research Question

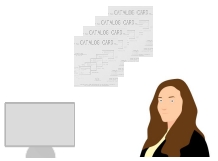
- ❑ What are the social scientists' data reuse behavior?
  - types of users, types of reuse, types of secondary data reused, and sources of the secondary data?
- ❑ What are the **types of information** that social scientists pay attention to when
  - **locating** secondary data of interest, attempting to **understand** the data, and **evaluating** the reusability of the data?
- ❑ What are the issues faced by social scientists in reusing quantitative research data?
  - How does the current data repository system influence social scientists' intention to reuse data?



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Method

- Face-to-face
- Semi-structured interviews
- 21 social scientists
- 6 social science fields
- Audio-recorded
- IRB approved (IRB-2017-03-046)
- Email invitation (Invitation sent to 94 faculty members and 6 research staff members)
- Aug – Nov 2017
- Interview duration: Average 50 min



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Method

- Basic list of questions
- Intent of the study: **Exploratory**
- Mainly asked "**How**" questions
- **Follow-up** questions varied from case to case

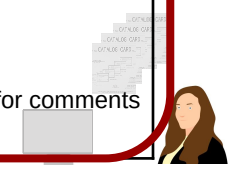
Part 1 – recall data reuse experience

Part 2 – access ICPSR to look for data of interest

Part 3 – discuss challenges met, perceptions and concerns of reuse (non-reuse)

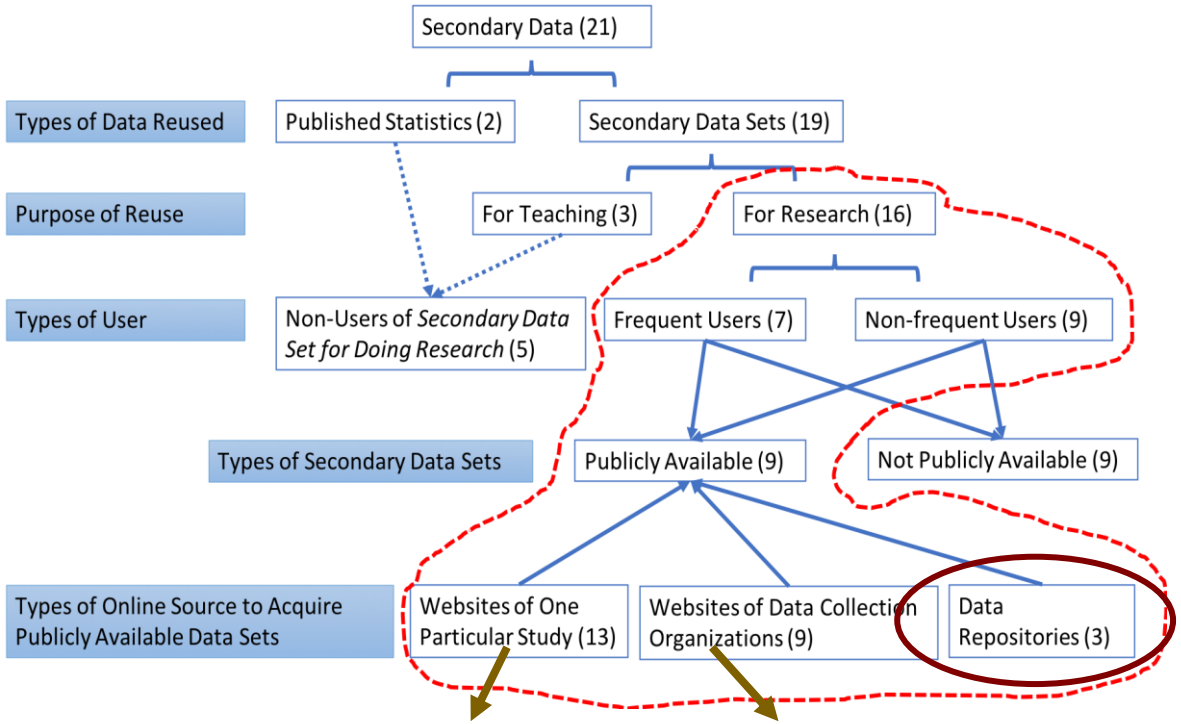
## List of Basic Questions

| | |
|---|---|
| B1 | Have you used secondary data (research data or publicly available data) for your own research? (Secondary data are data collected by someone other than you.) |
| B2 | Can you recall the types of secondary data you have reused?<br>(2a) Data stored in data repositories that are publicly available. They can be collected by individual researchers, research groups, and governmental or inter-governmental organizations.<br>(2b) Data collected and stored by other researchers. Not publicly available.<br>(2c) Others. Please elaborate. |
| B3 | Can you recall the data repositories (or online sources) that you have used to download data? |
| B4 | Can you recall the last time you reused data?<br>*Follow-up questions:<br>• When was it?<br>• What was the context?<br>• Which data repository did you use? How did you identify the repository? Why choose it? |
| B5 | (Access the repository and ask) Can you describe how you searched for and identify the appropriate data?<br>*Follow-up questions:<br>• How did you search? (hint: keywords used)<br>• How did you browse? (hint: browsing dimensions preferred)<br>• How did you understand the data? (hint: metadata elements)<br>• How did you decide whether to use the data or not? (i.e. preliminary data usability evaluation)<br>• What challenges did you face? Any suggestions how the interface or system may be improved? |
| B6 | How did you reuse the data?<br>*Follow-up questions:<br>• How did you analyse the data? (hint: online statistical analysis function)<br>• How did you integrate the data?<br>• What challenges did you face? Any suggestions how the interface or system may be improved? |
| B7 | What do you think of data reuse? (hint: usefulness, challenges met, concerns) |
| B8 | For interviewees who have never reused data, ask:<br>• Why haven't you reused data? (hint: concerns)<br>• If you were to reuse data, what are the challenges would you expect?<br>• Show the participant one data repository (potentially relevant to his/her research area), and ask for comments on the repository design. |



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Results – Basic Data Reuse Statistics



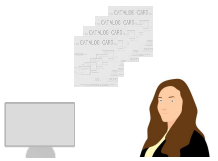

**Data repositories**:

- ❑ Store (not "collect") data sets of **various topics** within *the social sciences* research community.
- ❑ Requires the **additional effort** of searching, browsing and identifying data sets relevant to the social scientist's research topics.

**Discussion: Lack of users of data repositories** (Same to Mischo et al., 2014)

**Website of one particular study / data collection organization**:

- ❑ collect and store data sets of a **clear theme** within a clear topic **boundary**.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Results - Characteristics of Users, Reuse, and Data

**Table 3-4. Types of User**

| | Frequent User (n=7) | Non-frequent User (n=9) |
|---|---|---|
| **Criteria for Categorization** | Self-reported as reusing secondary data **frequently**<br>At the time of the interview, they were **in the process of reusing** some data | Self-reported as reusing secondary data sets **occasionally**<br>At the time of the interview, the last incident of reuse took place **at least several months ago** |

**Table 3-10. Three Types of Online Resources Mentioned to Acquire Publicly Available Data Sets**

| Research Studies | Data Collection Organizations | Data Repositories |
|---|---|---|
| 1. Programme for International Student Assessment<br>2. National Election Study<br>3. Health Information National Trend Survey<br>4. Capital Survey in the United States<br>5. General Social Survey<br>6. National Organization Survey<br>7. The US Census<br>8. Indonesia Family Life Survey<br>9. Indian Human Development Survey<br>10. Health & Retirement Survey<br>11. Chinese Family Panel Studies<br>12. Chinese Household Income Project<br>13. China General Social Survey<br>14. Pew Global Attitude Survey<br>15. World Value Survey<br>16. General Social Survey<br>17. Eurobarometer | 1. Organization for Economic Co-operation and Development<br>2. Pew Research Center<br>3. World Bank<br>4. National Bureau Statistics of China<br>5. National Human Resources Social Security Commission<br>6. World Health Organization<br>7. Bureau of Labor Statistics<br>8. China Statistics Bureau<br>9. Singapore Department of Statistics | 1. ICPSR<br>2. IPUMS<br>3. Global Terrorism Database |

**Table 3-6. Types of Reuse of Secondary Data Sets**

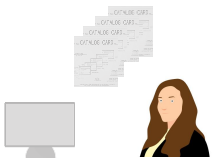
| What Have Been Reused Specifically | Use | Reported in Which Section of Paper | Type of Reuse |
|---|---|---|---|
| Questionnaire | Adapt question items to design own questionnaire (n=2) | Not in the final paper | To support the conceptualization of the study design (7) |
| Data Set | Get research idea quickly (n=2)<br>Pilot research ideas (n=2)<br>Test the rigor of primary study design (n=1) | Not in the final paper | |
| | Provide background information (n=5)<br>Back claims social scientists make (n=2) | Introduction, Literature Review | To complement the main study (9) |
| | Test the external validity of primary study result (n=2) | Results | |
| | To write a full paper by re-analysis of one or multiple (with data integration) secondary data sets (n=4)<br>To write a full paper by comparing secondary data sets with primary data sets (n=3) | Full Paper | To build the main study (7) |

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

## Locate a dataset for potential reuse

1. ***browsing or searching for data sets*** in the data repository that meet some criterion; and then
2. ***selecting*** *one or more data sets of* ***potential relevance*** to examine the surrogate record more closely.

| | Types of information paid attention to |
|---|---|
| *Recall data reuse experience* | **Notes**: *They didn't browse data in repositories. They just knew which data to use based on memory. No types of information was identified.* |
| *Interact with ICPSR website* | Summarized Data Sets Search Results: **research concepts** (inferred from title), **data collector**<br>Individual Detailed Surrogate Record:<br>"**Summary**" section (information on core variables of a data set) |



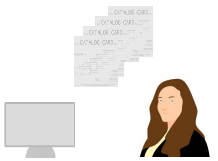
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Understand a dataset

| | Types of information paid attention to |
|---|---|
| *Recall data reuse experience* | **Questionnaires, Codebooks** (as expected)<br>**Published journal papers** (less expected) |
| *Interact with ICPSR website* | Individual Detailed Surrogate Record:<br>"**Summary**" section (main variables investigated in the study) |

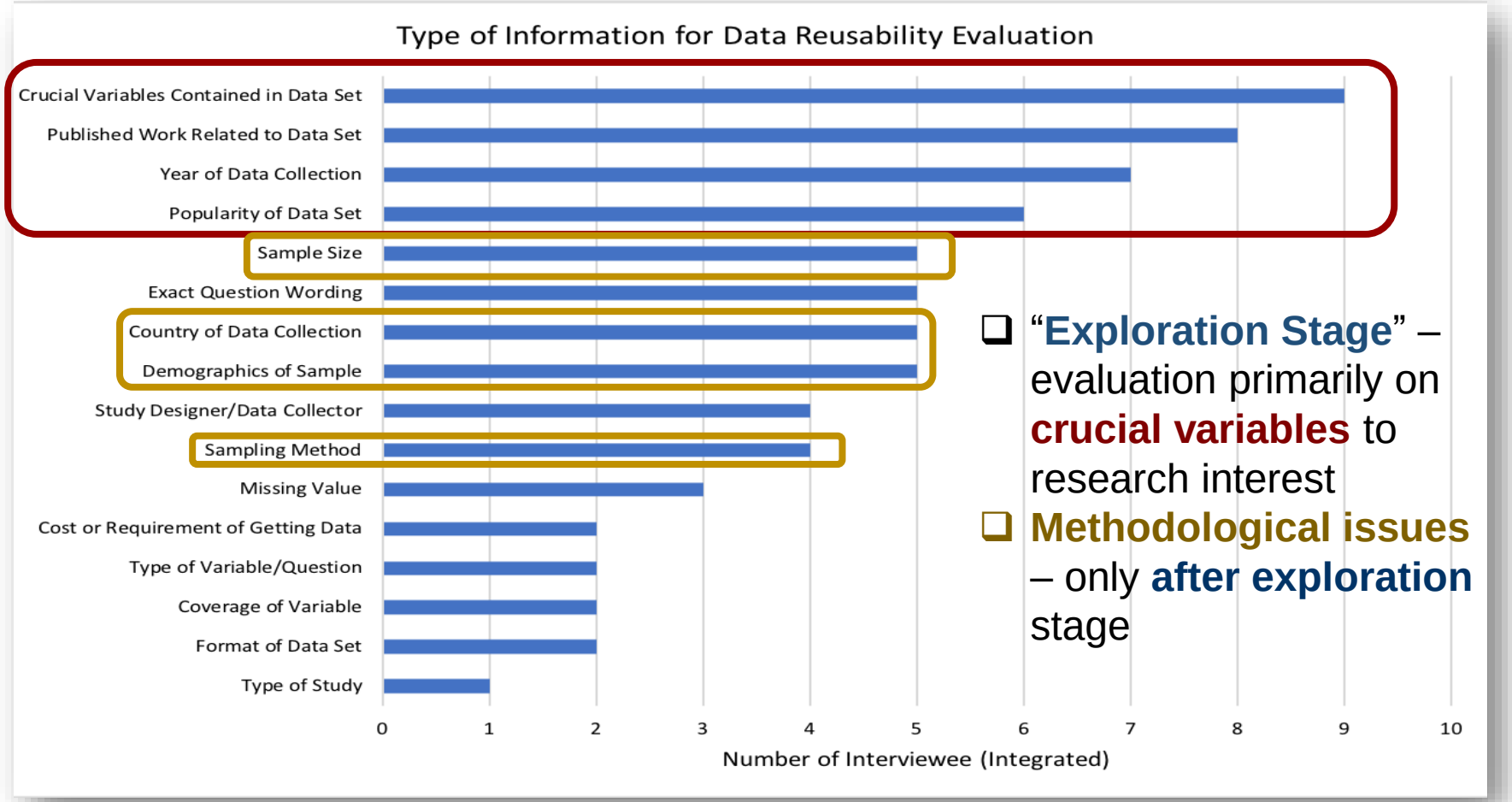
**Implication** for my **knowledge representation system**:

- ❑ design to **reduce** the complexity and **effort** during the process of data understanding.
- ❑ represent information from questionnaires/codebooks, published journal, and data sets in an **integrated view** that save users the trouble to toggle between different documents.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Evaluating the reusability of data sets




- "**Exploration Stage**" – evaluation primarily on **crucial variables** to research interest
- **Methodological issues** – only **after exploration** stage

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Discussion - Evaluating the reusability of data sets

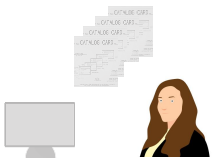
**Exploration stage** → **"Considering" to reuse** ✗ **"Deciding" to reuse**

**My Work**

- ✓ locate data sets of interest
- ✓ gain preliminary understanding of the data, and
- ✓ evaluate potential data reusability
- × stop at the point when the user harvests the data set for more detailed analysis offline.

**Activities in between** (includes but not limited to):

- ❑ scrutinizing **details of the data collection method** and execution process (if reported) to **assess the quality of data** and its **fitness** to their research **methodological requirements**;
- ❑ reviewing published papers based on the data sets to ensure that their **research ideas have not been investigated in extant studies**. This is to ensure the novelty of the social scientists' research contribution;
- ❑ running **statistical analysis** on the data. Social scientists will "decide" to reuse data sets only if **desirable results** are produced.

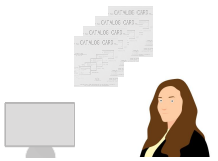


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Results – Data Reuse Issues

**Table 3-13. Issues and Challenges of Reusing Secondary Data Sets**

| Issues | Challenges |
|---|---|
| Variable Semantic Issue (n=6) | Understanding confusing variable name & label (n=4)<br>Understanding what concepts that the variables are measuring (n=2) |
| Research Novelty Issue (n=4) | Ensuring that the data set has not been reused by others for similar topics (n=2)<br>Coming up with new research ideas (n=2) |
| Statistical Analysis Issue (n=6) | Dealing with missing value (n=3)<br>Dealing with out of range value (n=2)<br>Dealing with large-scale secondary data sets (n=2)<br>Cleaning data (n=1) |
| Issue of Mismatches between Users' and Study Designers' Research Need (n=12) | Finding crucial variables needed in data sets (n=7)<br>Accommodating conceptual gap (n=6) |
| Issue of No Control over Data Collection Process (n=6) | Giving trust to data quality relevant to data collection process (n=6) |
| Other Issue (n=1) | Reading lengthy questionnaires(n=1) |



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

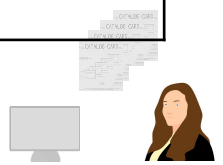
# Study 1

Derived **principles for knowledge representation system design**:

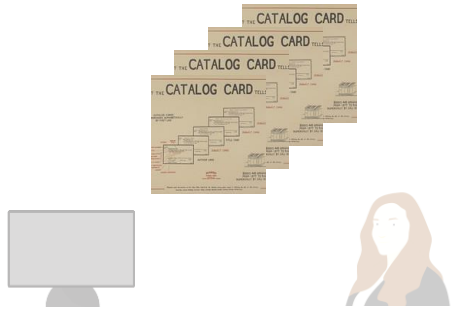
- ❑ To **support data exploration**, thus to promote the use of data repositories by social scientists.
- ❑ To focus on making the **searching** and **browsing** of data sets **based on research concepts** as intuitive as possible.

Identified user requirements:

| **The knowledge representation system** | **The types of information to be represented (The minimal set of information)** |
|---|---|
| To support different types of reuse | ❑ Data set<br>❑ Questionnaire/Codebook of data sets<br>❑ Publications relevant to data sets |
| To support data integration | ❑ possible "key" variables in the data sets that can be used for data set integration;<br>❑ linkages of same or similar variables across data sets |
| To support data exploration | Of data sets:<br>❑ number of published works related to the data sets;<br>❑ year of data collection<br>Of variables:<br>❑ exact questions been asked on the variable;<br>❑ answer choices of the question, which correspond to values of variable in data sets;<br>Of publications related to data sets:<br>❑ research concepts investigated;<br>❑ research objectives;<br>❑ research questions (if exist);<br>❑ hypotheses (if exist); |



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

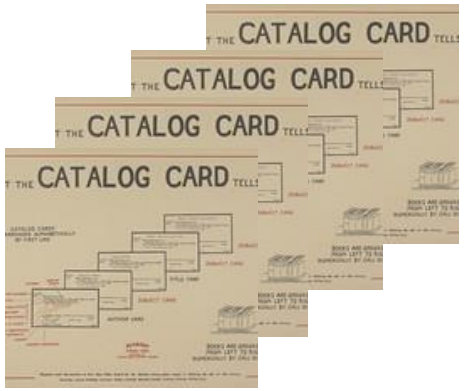



# Study 2

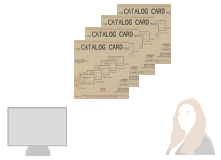
## Design and development of a knowledge representation system

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

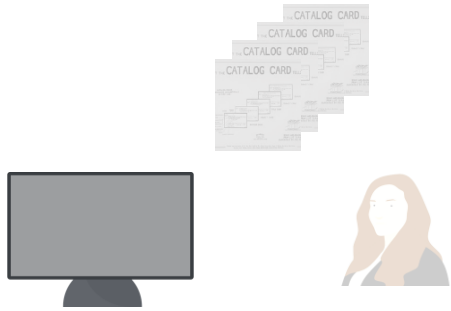
# Research Question




- What is the **minimal set of core metadata elements** needed to support social scientists data exploration?
- Are there any **types of information still lacking** in current metadata standards?
- What are the Resource Description Framework (RDF) **classes** and **properties** needed for the **representation** of **semantic information**?



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE



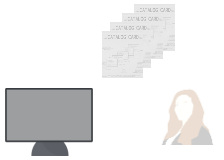

# Study 3

## User Evaluation Study

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Research Question

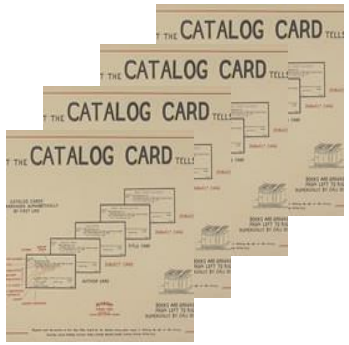
- Can the designed knowledge representation system be applied to a data repository to support an **organic integration and exploration** of social science data **files**?
- Is there any **unexpected functions** that the proposed system can support? What are they?



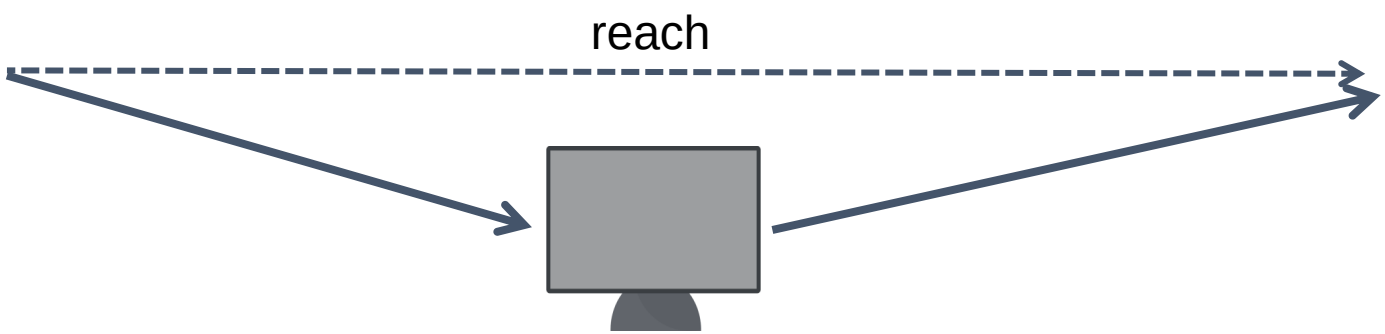
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Background Explanation

* **The user evaluation cannot be carried out directly**:
  - ✓ A prototype system (including a graphical user interface) was developed
  - ✓ Then a task-based user evaluation



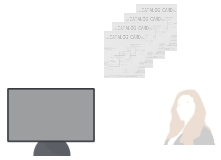

Information and knowledge encapsulated in the knowledge representation system




User

**via**

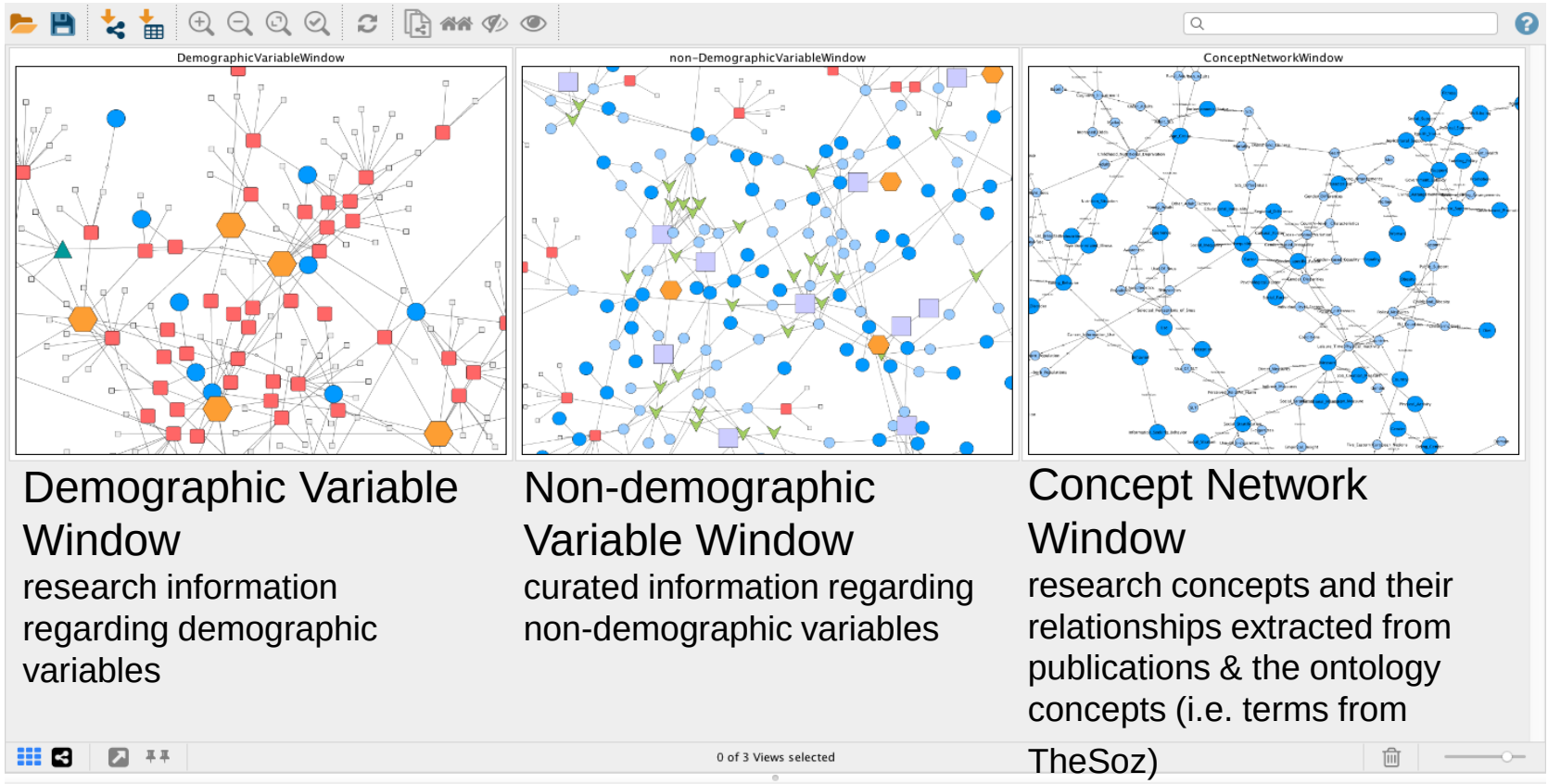
System and a user interface



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# For the task-based user evaluation…

- **4 sets** of visualization interface
- Each set contains **3 windows**
  - All the windows are **text-searchable**
  - Users can zoom in or select a subset of nodes for **closer examination**




**Demographic Variable Window**
research information regarding demographic variables

**Non-demographic Variable Window**
curated information regarding non-demographic variables

**Concept Network Window**
research concepts and their relationships extracted from publications & the ontology concepts (i.e. terms from TheSoz)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Conclusion

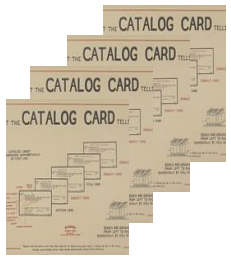
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Contribution

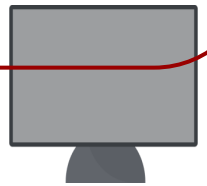
## Study 3 – Contribute to repository interface design

A catalogue of **interface design features** were constructed which can be investigated more thoroughly in future studies.

Yielded insights into what design **features** are **useful** and which are confusing to social scientists.




Data set records



Interface (the user-end of a data repository system)

Social scientists

## Study 2 – Contribute to metadata and ontology design

- Suggest what **kinds of metadata elements** for description of social science data sets are useful for **data reuse**.
- The ontology framework, and issues encountered and solutions adopted in the ontology construction can be **applied to developing** ontologies for quantitative data sets **in other disciplines**.
- **Desirable extensions** to the Web Ontology Language (OWL) 2.0 specification were identified.
- The demographic ontology itself is also a useful contribution to semantic web applications. It yields **insights into the demographic and socio-economic characteristics** of social science research.

## Study 1 – Contribute to social science data reuse literature

- Fill the gap – **relating** data reuse **intentions** to repository **design**;
- Understand & Explore – social scientists' **data reuse behaviour** when **interacting** with a data repository

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

## Major Limitation

- Small sample of users
  - *Generalizability?*
  - **Exploratory** study.
- Manual process of generating metadata records and ontology instances used in the graphical visualization prototype
  - *Current practice*: **manual** metadata creation

## Future Study Directions

1. **Meta-analysis** on the **number of** social science research papers generated using primary data, and those from reusing secondary data to gain substantial implication on the **popularity of data reuse**

2. **Investigate** data reuse behaviours of social scientists working in a **different research environment** compared with Singapore (e.g., U.S., Europe where there is higher number of social science data archives or institutional data repositories).

✓ **Automation** using machine learning techniques—test and select natural language processing algorithm(s) to **build models** to 1) identify research objectives/research questions/hypotheses sentences from publications; 2) to extract research concepts and their relationship from these sentences. Results generated from this computational techniques will still need to be scrutinized and rectified by human manually.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Future Study Directions (cont.)

- Address potential issue of information overload

Investigate the amount of metadata information should be visualize in one screen that best fits people's cognitive ability (i.e. to strike a balance between right amount of information and not overloading users).

- To study the **number (range) of publications** related to a data set that can be **presented at one time** for optimal visualization.
- To Identify **selection criteria for the publications**. (e.g.,

  *e.g.*, a certain types of metrics such as the number of citation?

  *e.g.*, customized selection based on users' historical behaviour? → The future study is to investigate the selection models for publications recommendation on a fixed data set.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Thank you ☺