PREPARING, ARCHIVING
AND ACCESSING NIJ DATA

Patrick Clark, NIJ Data Resources Program Manager
Cheryl Crawford Watson, NIJ Human Subjects Protection Officer
Julia Roach, NACJD Archive Manager
August 20, 2014
AGENDA

- Part One: Human Subjects, Confidentiality & NIJ Archiving Requirement
- Part Two: Overview of ICPSR and NACJD
- Part Three: Data Archiving Lifecycle
- Part Four: Depositing Data
- Part Five: Using the NACJD website
PART ONE

Human Subjects, Confidentiality & NIJ Archiving Requirement
Laws and Regs. re: DOJ-funded Research

- Confidentiality of Information (42 USC §3789g)
- Confidentiality of Identifiable Research and Statistical Information (28 CFR Part 22)
- Protection of Human Subjects (28 CFR Part 46)  
  [The Common Rule]
Confidentiality Basics

• Identifiable data collected with DOJ funds can only be used for research and no other purpose without subject consent.

• Identifiable data must be kept confidential.
  – Full face photographic images, SS#, names, addresses, biometric identifiers, including finger & voice prints, IP addresses

• The consent process is critical!
Department of Justice Confidentiality Statute
(42 USC §3789g)

“[N]o officer or employee of the Federal Government, and no recipient of assistance...shall use or reveal any research or statistical information furnished...by any person and identifiable to any specific private person other than the purpose for which it was obtained. Such information...shall be immune from legal process, and shall not, without the consent of the person furnishing such information, be admitted as evidence or used for any purpose in any...judicial, legislative or administrative proceedings.”
All applicants for DOJ funds must submit a Privacy Certificate (PC):

- describing the research,
- **assuring** that the applicant will comply with the requirements, and
- describing the procedures that will be used to protect the confidentiality of the identifiable information collected as part of the research.
  - To include data archiving plan & data sharing or information transfers agreements, if applicable
Notification (Inform + Consent)

- Research subjects should be informed:
  - the information they provide will be used for research & statistical purposes only;
  - participation is voluntary and can be terminated at any time;
  - study risks and benefits;
  - how the data will be securely maintained;
  - whether the data will be kept for future use, and if so how long;
  - Whether they may be re-contacted for follow-up;
  - how data will be archived/disposed of post-study; and.
  - Who is funding & who is conducting the study.

- Subjects then must give consent to participate and, if applicable, consent for reporting.
Special Grant Condition

Grant Products Due Ninety Days Prior to End of Award Period - Data and Associated Artifacts

- Data collected as part of the NIJ funded study is to be submitted according to the Data Archiving Plan approved by the NIJ grant manager at the time grant funds were awarded.

- Using the Data Submission Checklist, data and other associated documents are to be submitted 90 days prior to the end date of the grant.

- Data must be de-identified prior to submission unless archiving of information identifiable to a private person is provided for in the informed consent process.
1. Describe the formats of data to be collected (quantitative/qualitative/spatial)?

2. Describe software(s) that will be used in data collection, how will data be entered into a database, stored, analyzed and transferred?

3. Describe procedures by which data will be collected (e.g., individuals, locations, interviews, surveys).

4. If the proposed research includes collection of information identifiable to a private person include:
   - Describe all assurances of confidentiality made to those persons.
   - Provide a copy of the consent form that was utilized.
   - Provide a copy of the Privacy Certificate (as submitted to and approved by the funding agency).
   - Provide IRB approval documentation.
Required Data Archiving Plan (cont.)

5. Describe processes used to de-identify qualitative, quantitative and geo-spatial data in preparation for archiving.

6. Describe any information transfer agreement utilized to transfer the data with identifiers.

7. Describe any anticipated variable creation, data transformations, or scale construction that may be needed in the interpretation or analysis of the data by others.

8. Describe the plan for submission of computer programming code or software syntax, provide detail on how the data will be processed, including any significant treatments of the data, such as de-identification, imputation, filtering, or weighting.

9. Describe technical documentation (e.g., data dictionary or codebook) that explains how variables will be designated in the data file(s), such as the use of variable-naming conventions and variable groups, missing data/value designations, variable and category/value labels, operational definitions, and citations as needed for these variables.

10. Provide contact information for the Principal Investigator in the event other researchers should need more information about the study or the data.
As part of the process:

- Make IRB Aware of Archiving Requirement
- Include Archiving in the Privacy Certificate and have IRB review and sign it.
NIJ HSPO & DRP

Contact Information

Patrick Clark
DRP Program Manager

Cheryl Crawford Watson
Human Subjects Protection Officer
National Institute of Justice, USDOJ

Patrick.Clark@usdoj.gov
Cheryl.Crawford.Watson@usdoj.gov
Overview of ICPSR and NACJD
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH: ICPSR
GUIDE TO ARCHIVING

ICPSR
INTER-UNIVERSITY CONSORTIUM FOR
POLITICAL AND SOCIAL RESEARCH

Guide to Social Science
Data Preparation
and Archiving

Best Practice Throughout the Data Life Cycle • 5th edition

http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/
Data Archiving Lifecycle
PHASES OF THE DATA ARCHIVING LIFE CYCLE

I. Proposal Development & Data Archiving Plan
II. Project Start-up
III. Data Collection and File Creation
IV. Data Analysis
V. Data Depositing and Sharing
PROPOSAL
DEVELOPMENT &
DATA ARCHIVING PLANS

Phase 1
PROPOSAL DEVELOPMENT AND DATA ARCHIVING PLANS

NIJ Data Archiving Plan*
- Data description
- Format
- Security
- Responsibility
- Access and sharing
- Ethics and privacy
- Intellectual property
- Identify dataset(s) and variables used

Additional Elements to consider
- Metadata
- Update Data Archiving Plan as needed to include these items

Initial items to consider

- Data file and structure
- Naming conventions
- Data integrity
- Preparing dataset documentation
- Variable construction
- Project documentation
- Data Verification
DATA COLLECTION AND FILE CREATION

Phase 3
BENEFITS OF QUALITY DATA SETS

- For original data collectors
  - Preserves original data and project information
  - Enables continued use of the data at a later date
  - Reduces questions by secondary analysts of the data
- For secondary data users
  - Allows analysis of data without the expense of data collection
  - Provides quality data sets for training new researchers
**PRODUCING A QUALITY DATA SET**

**Items to consider**
- **Choice of software**
  - for storing data and documentation
  - data file structure and content
  - data transformations
- **Documenting decisions made about the data**
- **Handling of confidentiality and disclosure risk issues**
- **Congruence between the data and the data collection instrument and other project materials**
QUALITY QUANTITATIVE DATA

Codes and Coding
- Assign a unique identifier to every record
- Use mutually-exclusive, exhaustive and precisely defined codes
- Distinguish different types of missing data
- Include all cases and variables
  - Archive de-identified data

Data Entry
- Use a data-entry program that can catch errors
- Consider using double entry
- Check the first 5-10% of records entered
- Preserve the original information
- Verification (e.g. versioning, case/variable counts, wild codes)
QUALITY GEOSPATIAL DATA

- Is computer-readable; preferably in a standard package (e.g. ESRI)
- Used in conjunction with files that can produce visualizations (maps)
- Includes geographic data not used in any visualization (e.g. county name, census tract number)
- Archived data must be aggregated to a level so that it is de-identified
QUALITY QUALITATIVE DATA

- Preferred data formats include:
  - Text in Plain text, Rich Text or Word format
  - Databases in Microsoft Access
  - Spreadsheets in Microsoft Excel

- Completely de-identified using pseudonyms or other generalized text before data are archived

- Transcribed data should include:
  - a unique identifier
  - a uniform layout as well as header, page numbers, line breaks
  - speaker tags indicating the question-answer sequence
PRODUCING QUALITY DOCUMENTATION

- Study documentation – describing the data collection as a whole
- Codebook – variable-level information
- User guide – guidance for using the data well
- Technical documentation – additional information about instruments used, decisions or changes made during the data collection process, etc.
DATA ANALYSIS

Master Datasets and Work Files

- Data and Documentation Versioning
- Raw Data vs. Statistical System Files
- File Structure
- Data Backups
We will fully assist in the data submission process for all contributions to NACJD
PART FOUR

Depositing Data
Guidelines for Depositing Quantitative Data at NACJD for National Institute of Justice Sponsored Research

To facilitate secondary use, it is important to fully document and describe variables in the context of the data as well as in the documentation. This document provides guidance and a checklist to assist in preparation for archiving. Accordingly, NACJD staff review deposits for completeness and accuracy and send a report to NIJ. Any outstanding issues will need to be resolved before the submission can be finalized and assigned to a data processor. Please note the following:

Confidentiality — All fields directly identifying research subjects must be removed prior to deposit or transfer to NACJD (unless explicitly permitted and provided for through an IRB-approved process for informed consent). Examples of direct identifiers include names, addresses, and linkable numbers such as social security numbers, driver license number, prisoner identification number, etc. Data submitted with direct identifiers will be rejected and a resubmission of a repaired dataset will be required. Researchers should also carefully consider indirect or secondary identifiers and assess their analytic importance. Such variables make unique cases visible and may need to be removed before deposit. For example, a ZIP code variable may not be problematic on its own, but when combined with race and income variables, it may allow identification of unique individuals (e.g., the extremely wealthy).

Deposit Checklist and Guidelines

The following checklist details the items we are looking for with your submission:

- **Data** — Submit the final, sanitized data.
- **Format** — SPSS system files with embedded variable elements are preferred but other formats are acceptable as well (e.g., SAS, Stata).
- **Secondary analysis of existing data** — For projects that only involve analysis of existing data already publicly available where the product of the project is the analysis alone, data does not need to be deposited. Researchers are encouraged to deposit the code that produced the analysis and indicate that the project is a secondary analysis of existing data on the deposit form.
- **Analysis variables** — Submit the final, cleaned data file(s) used to produce analysis and tables in the final report including transformed, computed, or recoded variables that were constructed or derived from variables collected using the data collection instrument(s).
- **Project documents** — Submit copies of the final, approved Data Archiving Strategy, IRB approval, Privacy Certificate and Consent forms.
- **Syntax** — Describe the source of any transformed variables and the method for deriving them by depositing the recode syntax statements and/or providing a more extensive explanation in the documentation.

... checklist continued on reverse
WHAT TO DEPOSIT

- Data Submission Checklist and User Guide – new!
- IRB approval, Informed Consent and Privacy Certificate
- Data (de-identified)
- Documentation
- Syntax or special coding
- Bibliographies
DEPOSITING DATA

- De-identified
- Used to produce analyses and/or tables in published findings
- Include transformed, computed or recoded variables
- Provide all transformation/compute/recode syntax or code
- For quantitative data, remove unused character variables (e.g. comment/note fields)
Additional technical documentation to include:
- Unused copy of instruments and scales (with questions text)
- Codebook for each dataset
- Secondary Data Analyst’s User Guide (described in the checklist)
- Interviewer guide, coding instrument, etc. (qualitative data)
- Desired formats: PDF, MS Word, Rich Text, Plain text
DEPOSITING DOCUMENTATION: TECHNICAL DOCUMENTATION

**Project Level**
- Interviewing schedule or procedures
- Core questions over several waves
- Strengths and limitations of the data
- Contact information for technical assistance

**File Level**
- Inventory of files
- Data file structure
- Coder manual
- Imputation procedure
- How to aggregate, merge or subset data
- Errata
- FAQs
DEPOSITING DOCUMENTATION: CODEBOOKS

- Variable names
- Value labels for each code
- Missing data designations
- Variable formats
  - Variable type
  - Variable length
  - Decimal specifications
- Exact question text
- Skip patterns
- Weighted or unweighted univariate statistics
- Column locations for fixed-length ASCII data files
- Variable groupings
Recodes and/or formulas to create ALL derived variables
- Information on the creation of weight variables
- Coding scheme/tree (qualitative data)
- Secondary analysis of existing data

Consult the Deposit Submission Checklist and/or Deposit Guidelines for details.
Bibliographies for publications based on the data can be used to

- Identify research that has already been done in an area
- Replicate analyses of data
- Avoid duplicating analysis that has already been done
- Identify cross-disciplinary implications and uses of the data
Grantee’s part:

- Complete Data Submission Checklist
- Locate data files and related documentation
- Access online electronic deposit form
- Complete the deposit form fields
- Upload all files including Data Submission Checklist
- Sign deposit form electronically
NIJ DATA DEPOSIT FORM

Describe the Collection
This section collects basic details about the data collection. On the next page of the deposit form, you will be able to describe this data collection in more detail.

**Title of the Data Collection**
Descriptive titles typically include the time period(s) and geographic location(s) that the data cover.

**Principal Investigator(s)**
Please list the Principal Investigators in order of importance to the study. If the Principal Investigator is an organization, complete just the affiliation field.

- **PI First Name**
- **PI Last Name**
- **PI Affiliation**

**Description or Abstract**
Please give the user a clear sense of what the study is about. The focus should be on questions such as the purpose of the study, the major topics covered, and what questions the PIs attempted to answer when they conducted the study. Note that you should avoid attempting to address issues of how the data might be used, who might be interested in the data, or any evaluative comments about the worth or usefulness of the study.

**Grant Managers**
When applicable, please list the grant manager email addresses to be notified when the deposit has been signed and to receive the manifest of deposited files.

- **Name**
- **Email**
This message is automatically generated in response to files you (or your group) uploaded to ICPSR through the online deposit form. When ICPSR receives files through the data deposit form, we automatically generate a listing of the files received. This list also includes a description of the format of the files uploaded as detected by our software. And for data deposited in SAS, SPSS, or Stata formats, we provide a count of cases and variables in each data file.

Please review the following information carefully to be sure that it corresponds to what you intended to submit to ICPSR. Please contact us with any discrepancies or questions that you have.

Thank you,
ICPSR Acquisitions Staff

Deposit Title: San Diego Labor Trafficking Study Survey Data 2010-2011, Using RDS Sampling Method
Deposit Number: 34427
Link: [http://www.icpsr.umich.edu/cgi-bin/ddf2?key=n5gimQcaagutLlx0DRkSavRQHOrchojM/uvmt6dd&page=suppl](http://www.icpsr.umich.edu/cgi-bin/ddf2?key=n5gimQcaagutLlx0DRkSavRQHOrchojM/uvmt6dd&page=suppl)

Listing of Deposited Files

*Codebook_Main Data_7-31-2013.docx was scanned on 05-AUG-13.*
We think it’s: Microsoft Word

*Specific Trafficking Incident_July-16-2012_revised 7-31-13.sav was scanned on 05-AUG-13.*
We think it’s: SPSS System File TICS DATA FILE 64-bit MS Window
Variables: 99, Cases: 826

*San Diego Labor Trafficking Final report_11-25-12.pdf was scanned on 05-AUG-13.*
We think it’s: PDF document, version 1.5

*Codebook_Specific Trafficking Incidents_7-31-2013.docx was scanned on 05-AUG-13.*
We think it’s: Microsoft Word
NACJD DEPOSIT REPORT

NACJD staff review each submission for accuracy and completeness.

NIJ Grant # 2013-AB-CD-1234
NIJ Project Title: My NIJ Grant
Principal Investigator: John Smith

During the review of deposited materials, the following issues were found. Please work with your grantee to resolve these issues before submitting a resupply.

File: Main Data.sav

These variables are missing variable labels (Please make sure all variables are labeled):
- valid

These variables are missing value labels (Please make sure all values are labeled):
- variable DUP; values 0 and 1
- variable MatchSequence; values 0, 1, and 2
- variable valid; values .00 and 1.00
- variable west; values .00 and 1.00
- variable progcomp; values .00 and 1.00
- variable fail; values .00 and 1.00
Using the NACJD website
OVERVIEW OF HOW NACJD ARCHIVES AND PRESERVES DATA

- Evaluate the data collection for completeness, suitability for public release, and readiness for use
- Prepare searchable study descriptions and bibliographic citations
- Produce data in multiple formats: SPSS, SAS, Stata and R
- Back up and store both the original data and the processed data at an external site
DATA & DOCUMENTATION PROCESSING

- All files will be reviewed, variable by variable, and NACJD data processors will clean the files, working with the Grantee to resolve issues.
- Variable labels and value labels will be created if they do not exist.
- Updates will be made to the files when data users find and report errors or at the Grantee’s request.
- Errors or weaknesses are be discussed in data documentation (e.g. metadata, codebook notes).
Assess research data for indirect identifiers (a single or combination of variables) that could reasonably identify participants.

- Perturb data to prevent disclosure
- Build documentation and data collections for greatest access and least risk
NACJD does not publically release data containing information that might reasonably identify any individuals, organizations, or agencies who provided information under a promise of confidentiality.
STANDARD PRODUCTS DISSEMINATED

- Study Documentation
  - Final Report (linked to NCJRS website)
  - User guide
  - Codebooks
  - SPSS/SAS code
  - Other technical documentation
- Micro-data
- Publications
DATA DISTRIBUTION MODES

- Online/downloadable
  - ASCII with programs to read data
  - “ready-to-go” system specific software files
- Restricted Access (i.e. CD/DVD, secure download)
- Physical Enclave
- Other
Users can search our online database of citations for publications using or related to NACJD data collections.

The database contains over 64,000 citations of published and unpublished works resulting from analyses of data held in the ICPSR archive.

- Over 5,000 citations using NIJ sponsored data collections
EXPLORING THE NACJD WEBSITE
RESOURCES & REMINDERS

- **Guide to Archiving:**
  [http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/](http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/)

- **Submission Requirements**
  - De-identified Data, Documentation, Privacy Certificate, IRB Approval & Consent Forms, Data Archiving Plan, Submission Checklist, Bibliography
  - Available from our NACJD and NIJ websites
  - Webinar recording and slides available on NACJD website soon
CONTACT US: WE ARE HERE TO HELP!

Julia Roach
(734) 615-7833
jmaycock@umich.edu

General
(800) 999-0960
nacjd@icpsr.umich.edu

www.icpsr.umich.edu/NACJD