INFO 7470/ECON 7400/ILRLE 7400
Proposal Writing
and
Access to Restricted Data

John M. Abowd and Lars Vilhuber
February 2013
With thanks to Rosemary Hyson and Lynn Riggs
(U.S. Census Bureau)
Outline

• Why use restricted-access data?
• Two examples of access
• Elements of proposal writing in general
• Elements of the RDC proposal process
• Elements of proposal process for other RDC-accessible data
• What to expect once approved
• What is expected of you
Why Use Restricted-access Data?

- Public-use statistics (tabulations, micro-data)
  - Not detailed enough
    - Aggregated when micro-data required
    - Categories coarsened
  - Leave out critical elements
    - Detailed geography
    - Family structure
    - Longitudinal structure
Use of Pre-Existing Survey Data in Publications in Leading Journals, 1980-2010

Note: “Pre-existing survey” datasets refer to micro surveys such as the CPS or SIPP and do not include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.
Chetty (2012)

Use of Administrative Data in Publications in Leading Journals, 1980-2010

Note: “Administrative” datasets refer to any dataset that was collected without directly surveying individuals (e.g., scanner data, stock prices, school district records, social security records). Sample excludes studies whose primary data source is from developing countries.
Reasons for Restricting Access

• Confidentiality protection
  • Because of laws (Title 13, Title 26, HIPAA, etc.)
  • Because of contracts
  • Because of IRB

• Measures undertaken to protect confidentiality
  • Coarsening/collapsing of categories
  • Aggregation to higher levels
  • Swapping
  • Removal of longitudinal linkages
  • More on this topic later in this class
Examples

– Coarsening/collapsing of categories
  • Geography on NLSY or HRS
– Aggregation to higher levels
  • County Business Patterns, QCEW, QWI, Decennial Census
– Swapping
  • Decennial Census, ACS
– Removal of longitudinal linkages
  • Canadian Survey of Labor and Income Dynamics (SLID)
  • Decennial Censuses, Current Population Survey
– Top-coding
  • Current Population Survey
Solutions to Restricted Access

- Contracts
  - Regarding use of the data
  - Regarding local storage of the data
- Restricted access environments
  - Researcher has physical access
  - Researcher has remote access
  - Researcher has remote access by proxy
Solutions to Restricted Access

Contracts

- Regarding use of the data
  - *NLSY releases more detailed geodata with contractual engagement*

- Regarding local storage of the data
  - *HRS releases more detailed geodata, or linked data, with requirement to set up secure local environment*
Solutions to Restricted Access

- Restricted-access environments
  - Researcher has physical access
    - RDC at Census, BLS, NCHS
  - Researcher has remote access
    - CRADC (one of the first)
    - NORC data enclave
    - German IAB
  - Researcher has remote access by proxy
    - Remote processing facility at Statistics Canada, NCHS, IAB, LIS
    - Staff assisted (NCHS)
Formulate research question
Research available data

Data is public-use?

Is RA data really necessary?

- Justification for access
- Request access to data

Approved

Set up secure environment

- Go to RDC
- Run analysis
- Request release of results

Released

Run analysis

- Write paper
- Publish result
- Win Nobel prize
Obtaining access to restricted data

• Identify and learn about the data
  – Often public-use documentation
  – Not always easy

• Familiarize yourself with the data custodian's requirements
  – Physical access rules
  – Legitimate use rules

• Contact the data custodian
Obtaining Access to RDC Data

• Identify and learn about RDC data
  – Find public-use documentation
  – Not always easy

• Familiarize yourself with the RDC's requirements
  – Physical access rules: where are RDCs?
  – Legitimate use rules: Title 13, IRS criteria

• Contact the RDC administrator
Steps for RDC Proposals

• Initial proposal development
• Submission of formal RDC proposal
• Proposal review
• Approval (or not)
• Post approval process
• Post project activities
Initial Proposal Development

1. Identify and Learn about RDC Data

• CES Website
  – Data descriptions

• General Census website
  – Survey or program pages

• General research
  – Papers and articles that use RDC data (see CES Website for discussion papers, annual report)

• Personal contacts
  – RDC, CES, Census program staff
  – RDC and other researchers

• VirtualRDC
Things to Keep in Mind

• External data can be used with RDC data
  – (you may still need approval from other data custodians, though)
  – Timing, unit, link records need to be compatible

• Data not currently listed on CES website may also be available (-> personal contacts)
  – Space may be a restriction
2. Familiarize Yourself with RDC requirements

- Predominant purpose must be to increase the utility of Title 13, Chapter 5 programs
  - ... always!

- Scientific merit
  - Funding award from a competitive peer review process is sufficient

- Clear need for non-public data

- Feasibility
  - Given methodology, requested data, and environment

- No risk of disclosure (ex ante)
  - Still ex-post evaluation in release process
Excursion: Grant proposals (NIH)

“Get Prepared [link]

To ensure efficient and thorough completion of your application, consider taking the following preliminary steps:

• Review the grant application instructions for important information on the application process and guidance on preparing specific sections of the application.
• Carefully read the funding opportunity announcement (FOA) for any special instructions.
• Solicit feedback from colleagues and/or mentors on your research idea while it is still in the concept state.
• Prepare an outline following the application framework and structure described in the application guide.
• Make sure you have adequate preliminary data.
• Develop a feasible timeline with draft application deadlines. Be realistic about the time it can take to write and revise the application.
• Ask your colleagues or your Office of Sponsored Research for copies of successfully completed NIH grant applications. Examine them closely.
• Contact someone in your institution who can assist you in understanding and completing application materials.
• Make sure that your institution will allow you enough time to accomplish the research, if funded.
• Become familiar with the NIH peer review criteria; reviewers will use them to rate your application.”
Specific NIH Review Criteria

1. **Significance**: Does the project address an important problem or a critical barrier to progress in the field? If the aims of the project are achieved, *how will* scientific knowledge, *technical capability*, and/or clinical practice be *improved*? How will successful completion of the aims *change* the concepts, *methods*, technologies, treatments, services, or preventative interventions that drive this field?
Specific NIH Review Criteria

4. **Approach.** Are the overall strategy, *methodology*, and analyses well-reasoned and *appropriate* to accomplish the specific aims of the project? Are potential problems, alternative strategies, and benchmarks for success presented? If the project is in the early stages of development, will the strategy establish *feasibility* and will particularly risky aspects be managed? If the project involves clinical research, are the plans for 1) *protection* of human subjects from research risks, and 2) inclusion of ..., justified in terms of the scientific goals and research strategy proposed?
Grant Proposals (NSF)

Merit Review Criteria [link]

• **Intellectual Merit**: The Intellectual Merit criterion encompasses the potential to advance knowledge; and

• **Broader Impacts**: The Broader Impacts criterion encompasses the potential to benefit society and contribute to the achievement of specific, desired societal outcomes.
Specific NSF Criteria

1. What is the potential for the proposed activity to:
   a. Advance knowledge and understanding within its own field or across different fields (Intellectual Merit);
   b. Benefit society or advance desired societal outcomes (Broader Impacts)?

3. Is the plan for carrying out the proposed activities well-reasoned, well-organized, and based on a sound rationale? Does the plan incorporate a mechanism to assess success?

5. Are there adequate resources available to the PI (either at the home organization or through collaborations) to carry out the proposed activities?
Initial Proposal Development

3. Contact Your RDC Administrator

• Discuss the project idea with the RDC admin
• The RDC administrator will accompany you the rest of the way
• The RDC administrator will also be responsible for
  – Reviewing output (round-robin assignment)
  – Reviewing input (program uploads, etc.)
• Bring them cookies...
Formal Proposal Submission

1. Elements of a Formal Proposal

- Preliminary proposal information in CES proposal management system (*RDC admin will do this part*)
- Project description
- Predominant purpose (benefits) statement
- Abstract
Formal Proposal Submission

2. Preliminary Proposal Information

- Summarizes proposal information:
  - Researchers involved and affiliations
  - Data sets requested—RDC and external
  - Criteria met by proposed benefits
- Short proposal abstract
- Duration and funding for project
  - Note: extensions rarely granted, except for journal revisions
- RDC location(s) for project
Formal Proposal Submission

2. Preliminary Proposal Information

- Criteria met by proposed benefits
  - To the Census Bureau, benefits might be
    - New estimates derived from existing data
    - Expanded measurement capabilities
    - Improved documentation
    - Consultation with leading researchers
      (from Riggs presentation, 2009, in Atlanta)
  - What is the catch?
Formal Proposal Submission

2. Preliminary Proposal Information

- Criteria met by proposed benefits
  - Under Title 13, Chapter 5, benefits are
    - New estimates derived from existing data - yes
    - Expanded measurement capabilities - yes
    - Improved documentation - yes
    - Consultation with leading researchers - no
    (from Riggs presentation, 2009, in Atlanta)
Research Benefits (choose one or more)

In this step, you will be asked to indicate how your research project best benefits the Census Bureau. Proposals must demonstrate that the research is likely to provide one or more Title 13 benefits to the Census Bureau. A research project must demonstrate that its predominant purpose is to benefit Census Bureau programs. If a project has as its predominant purpose one or any combination of the following, it will be considered to have as its predominant purpose increasing the utility of Title 13, Chapter 5 data.

- Evaluate Concepts/Practices Re Statistical Data Collection/Dissemination
  Evaluating concepts and practices underlying Census Bureau statistical data collection and dissemination practices, including consideration of continual relevance and appropriateness of past Census Bureau procedures to changing economic and social circumstances. (This is not to be used when you have FTI data sets)

- Analyze Demo/Social/Econ Processes Affecting Programs and/or Improve Product Quality
  Analyzing demographic and social or economic processes that affect Census Bureau programs, especially those that evaluate or hold promise of improving the quality of products issued by the Census Bureau. (This is not to be used when you have FTI data sets)

- Increase Utility of Data for Analysis
  Developing means of increasing the utility of Census Bureau data for analyzing public programs, public policy, and/or demographic, economic, or social conditions. (This is not to be used when you have FTI data sets)

- Conduct or Facilitate Data Collection/Processing/Dissemination
  Conducting or facilitating Census Bureau census and survey data collection, processing or dissemination, including through activities such as administrative support, information technology support, program oversight, or auditing under appropriate legal authority. (This is not to be used when you have FTI data sets)

- Understanding or Improving Data Quality
  Understanding and/or improving the quality of data produced through a Title 13, Chapter 5 survey, census, or estimate.

- Improved Methodology
  Leading to new or improved methodology to collect, measure, or tabulate a Title 13, Chapter 5 survey, census, or estimate.

- Enhancing Data
  Enhancing the data collected in a Title 13, Chapter 5 survey or census. For example, improving imputations for non-response, or developing links across time or entities for data gathered in censuses and surveys authorized by Title 13, Chapter 5.

- Business Register
  Identifying the limitations of, or improving, the underlying Business Register, Household Master Address File, and industrial and geographical classification schemes used to collect the data

- Data Collection
  Identifying shortcomings of current data collection programs and/or documenting new data collection needs.

- Sample Frame Maintenance
  Constructing, verifying, or improving the sampling frame for a census or survey authorized under Title 13, Chapter 5.

- New Estimates
  Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5.

- Estimating Non-Response
  Developing a methodology for estimating non-response to a census or survey authorized under Title 13, Chapter 5.

- Statistical Weights
  Developing statistical weights for a survey authorized under Title 13, Chapter 5.
Formal Proposal Submission

3. Project Description

Typical components of any research proposal

- Background, discussion of relevant literature
- Proposed methodology
- Data sets
- Meets certain criteria (grant proposals)

Components specific to the RDC program

- Summary of proposed benefits to Census Bureau programs (again)
- Description of the research outputs you plan to request be disclosed (used to assess ex-ante disclosure risk)
- Description of how you plan to construct the analysis data set if combining multiple data sets, including external ones (feasibility)
- Expected project duration (and cost)
Formal Proposal Submission

4. Predominant Purpose Statement

PPS and benefits to Census Bureau Title 13, Chapter 5 programs

- PPS describes and documents proposed benefits
- Benefits must be the *predominant* purpose of project
- PPS is template for certifying benefits at project end (Post Project Certification-PPC)
Formal Proposal Submission

Benefit Criteria

1. Evaluating concepts and practices
2. Analyzing demographic/social/economic processes affecting Census Bureau programs and/or Improve Product Quality
3. Increasing the utility of Census Bureau data
4. Conduct/facilitate data collection/processing/dissemination
5. Understanding and/or improving quality of data
6. New or improved methodology to collect, measure, tabulate
7. Enhance data
   a) Improving imputations
   b) Developing links across time or entities
8. Business Register or MAF improvement
9. Identifying shortcomings of data collection
10. Sampling frame maintenance
11. New estimates
12. Estimating non-response
13. Statistical weights
Formal Proposal Submission
Benefit Criteria (cont.)

- **Useful links:**
  - “Research Opportunities at the U.S. Census Bureau”
  - “Writing Benefit statements” (CES)
  - For projects involving Title 26 (tax) data: “IRS Criteria Document”

- **Other sources**
  - RDC Administrator, Census Bureau program contacts
  - CES Annual Report, CES Discussion Papers
  - Summary of Census Bureau Research Problems Identified by Senior Staff
Formal Proposal Submission
The IRS

- For projects involving FTI
  - IRS will review the PPS only – need to make sure it is clearly written
  - Only benefits 5-13 apply for proposals that use FTI – as a *predominant* purpose
  - Group benefits in a logical way
Formal Proposal Submission

5. Processing Documents

- RDC administrator follows the process
  - Considers proposal complete
- RDC-specific Project review also has to approve for submission
  - Remember: cookies...
Formal Proposal Submission
6. Approval Process

Reviews by

- CES and Census Bureau subject matter and data experts
- external researchers (scientific merit, may use funding)
- relevant Census Bureau programs/offices (benefits, feasibility)
- RDC Disclosure Officer and other disclosure experts (disclosure)
- other data custodians/sponsors:
  IRS, HUD, BLS (benefits, feasibility)
Formal Proposal Submission

6. Approval Process

● Duration

   ● May be longer if involving FTI
   ● At least 3-6 months as a guideline, but check with your RDC administrator

● Other issues

   ● Special Sworn Status process
Things to Keep in Mind

• RDC environment is different
  – Linux OS (see SSG with KDE environment to get a feel)
  – Restricted set of software
  – No free upload of programs or software to the system
    • Software very difficult
    • Programs through web-based staging process with review
  – All output, including screen output
    • Can only be viewed within the RDC
    • Can only be viewed by authorized persons (this includes your thesis advisor!) - in a narrow sense (listed and authorized on your project, and only that project)
WARNING: DMS bold font metrics fail to match DMS font.

NOTE: AUTOEXEC processing beginning; file is /home/lv39/autoexec.sas.

NOTE: Libref HERE was successfully assigned as follows:
  Engine:  V9
  Physical Name: /export/home/admin/extras

NOTE: AUTOEXEC processing completed.

Stata/IC 11.1
Statistics/Data Analysis

5-user Stata network perpetual license:
Serial number: 50110530211
Licensed to: Cornell University Social Science Gateway

Notes:
1. (-m# option or set memory-) 10.00 MB allocated to data
Post Project Approval

1. SSS Application

- Forms, fingerprinting and documentation
- Training
- RDC Administrator checks forms
- SSS package sent to Office of Security
- May send back corrections/clarifications
Post Project Approval

2. Badge

- Census Bureau computer ID
- Project account setup
- CES Account and Data Request Form
Post Project Approval

3. External Data

- External data form
- CD/media with external data
- Use permission documentation

4. Researcher RDC orientation

--> Project finally starts!
Health Data Available at the RDCs

• NCHS & AHRQ Data Extracts available
  – More relevant for non-DC RDCs
  – Possible to link NCHS and/or AHRQ data to Census data
    • May be more costly and time-intensive
    • May be difficult to justify the benefits

• MEPS-IC is collected by Census for AHRQ and considered Census data (Census application process)
  – Asks employers about provision of health insurance
NCHS Version of Remote Access

- Several more **options**
  - NCHS RDC (Hyattsville, MD; Atlanta, GA)
    - More software choices
  - Census RDC
    - More locations, higher security approval burden
  - Remote access/submission
    - only SAS/Sudaan
  - Staff assisted
    - $$

© John M. Abowd and Lars Vilhuber 2013, all rights reserved
NCHS Version

• Similar process
  – Why not public use data
  – Proposal development
    • Benefit statement
    • CVs
  – Proposal submission
    • By email
  – Proposal review

• New choice
  – Access mode
Other Data Available through Restricted Access

• German IAB data (integrated worker and firm/establishment files)
  – Thin clients at Michigan [Cornell, Berkeley soon]
  – Some remote submission capability

• **Canadian RDCs** (household and survey)
  – No firm/establishment data (exception: WES)
  – Require collaboration with Canadian researcher, crossing the border
  – Some remote submission capability

• **Canadian Centre for Data Development and Economic Research** (CDER, new)
  – Firm/establishment data
  – Access only in Ottawa, but remote submission planned