CISER

Cornell Institute for Social and Economic Research
A LEADER IN SOCIAL SCIENCE DATA AND COMPUTING

An Early Prototype of the Comprehensive Extensible Data Documentation and Access Repository (CED²AR)

William C. Block and Jeremy Williams,[1]
John Abowd and Lars Vilhuber,[2]
and Carl Lagoze[3]

[1] Cornell Institute Social and Economic Research, Cornell University
[2] Labor Dynamics Institute, Cornell University
[3] School of Information, University of Michigan

Presentation at the 4th Annual European DDI User Conference (EDDI12)
Norwegian Social Science Data Services, Bergen, Norway
3 December, 2012

Cornell University

Outline

The Problem: Curation of Data Locked within a Secure Environment is Difficult

NCRN Solution:

- CED^{2}AR Prototype
- CED^{2}AR Search API
- DDI bridging the boundary between confidential and public metadata

Questions and Discussion

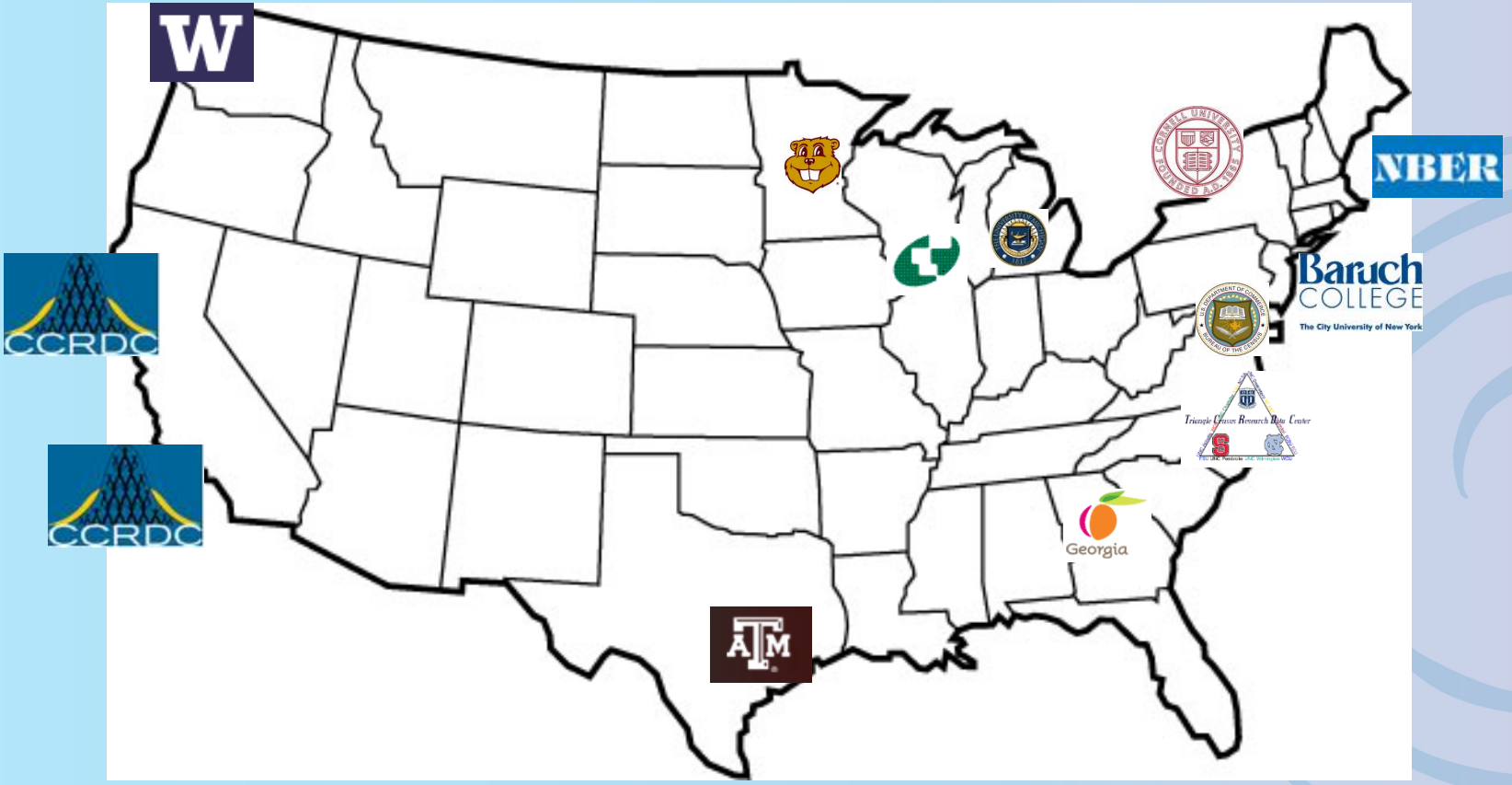
CISER

Curating Data Locked within a Secure Environment is difficult

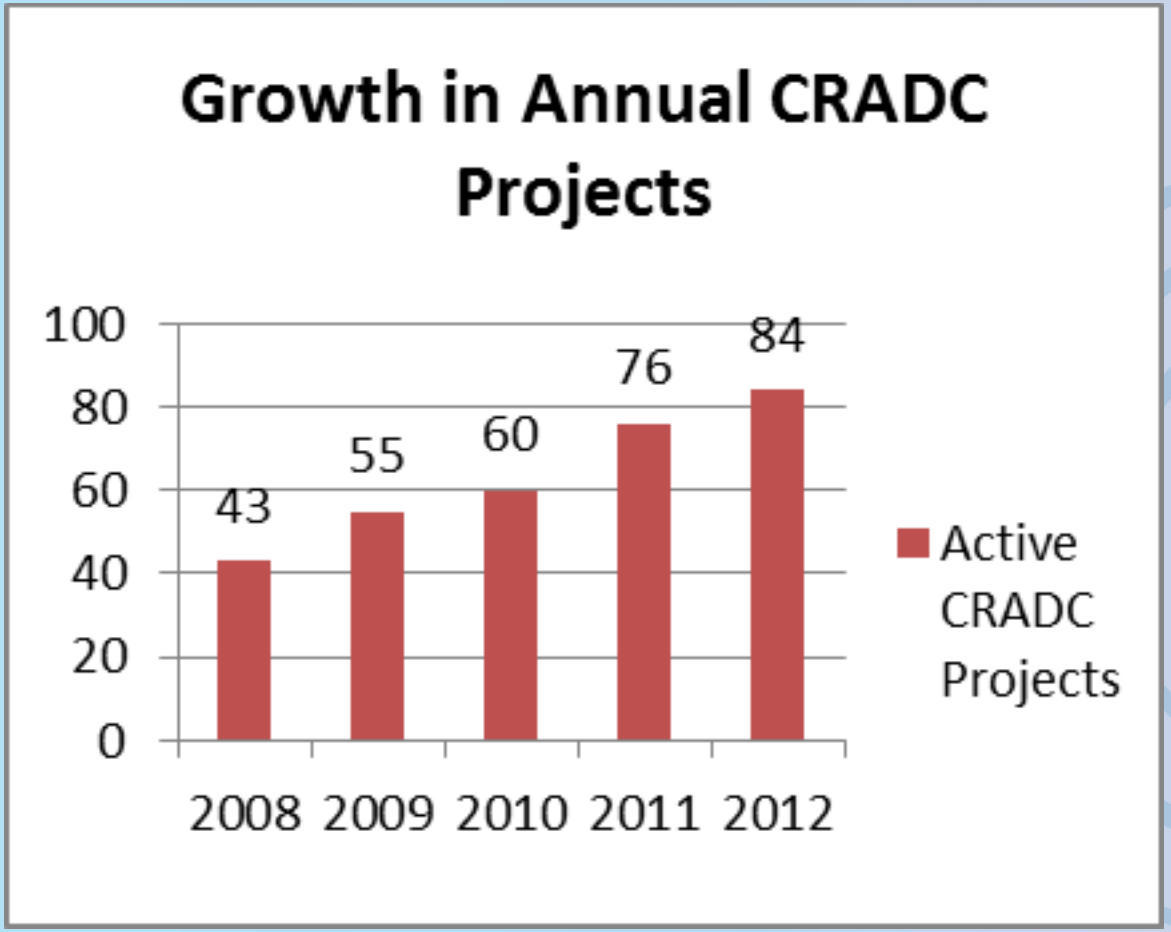
- By definition: Access is *Restricted*
- Lack of Curation throws up a barrier to Future Discovery and Access
- Replication of Results becomes increasing difficult
- Important! The Scientific Method depends on the ability to replicate the results of research

Research Opportunities
at the Cornell Census Research Data Center

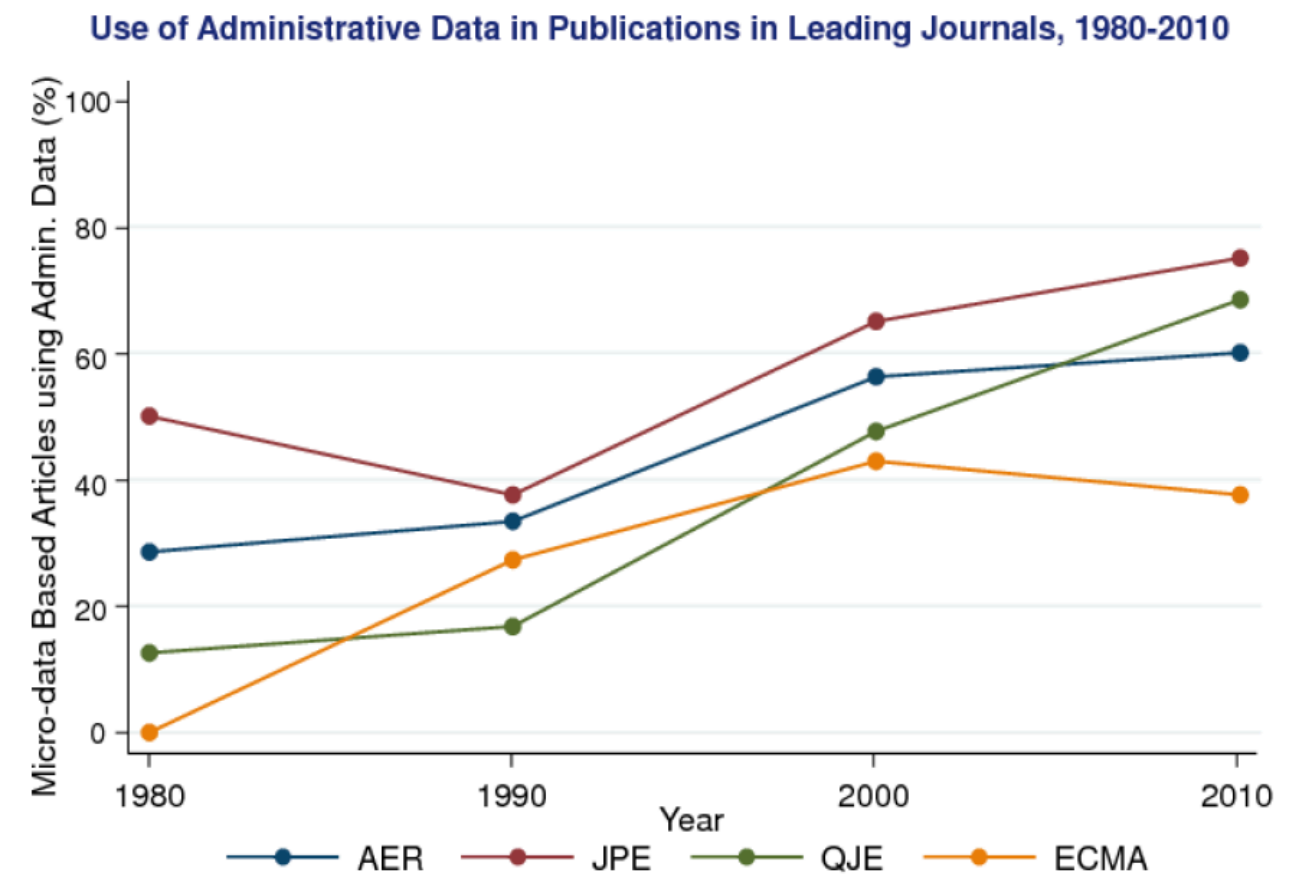
The RDC Network



We see this problem at Cornell: Research with Restricted Data increasing at CISER

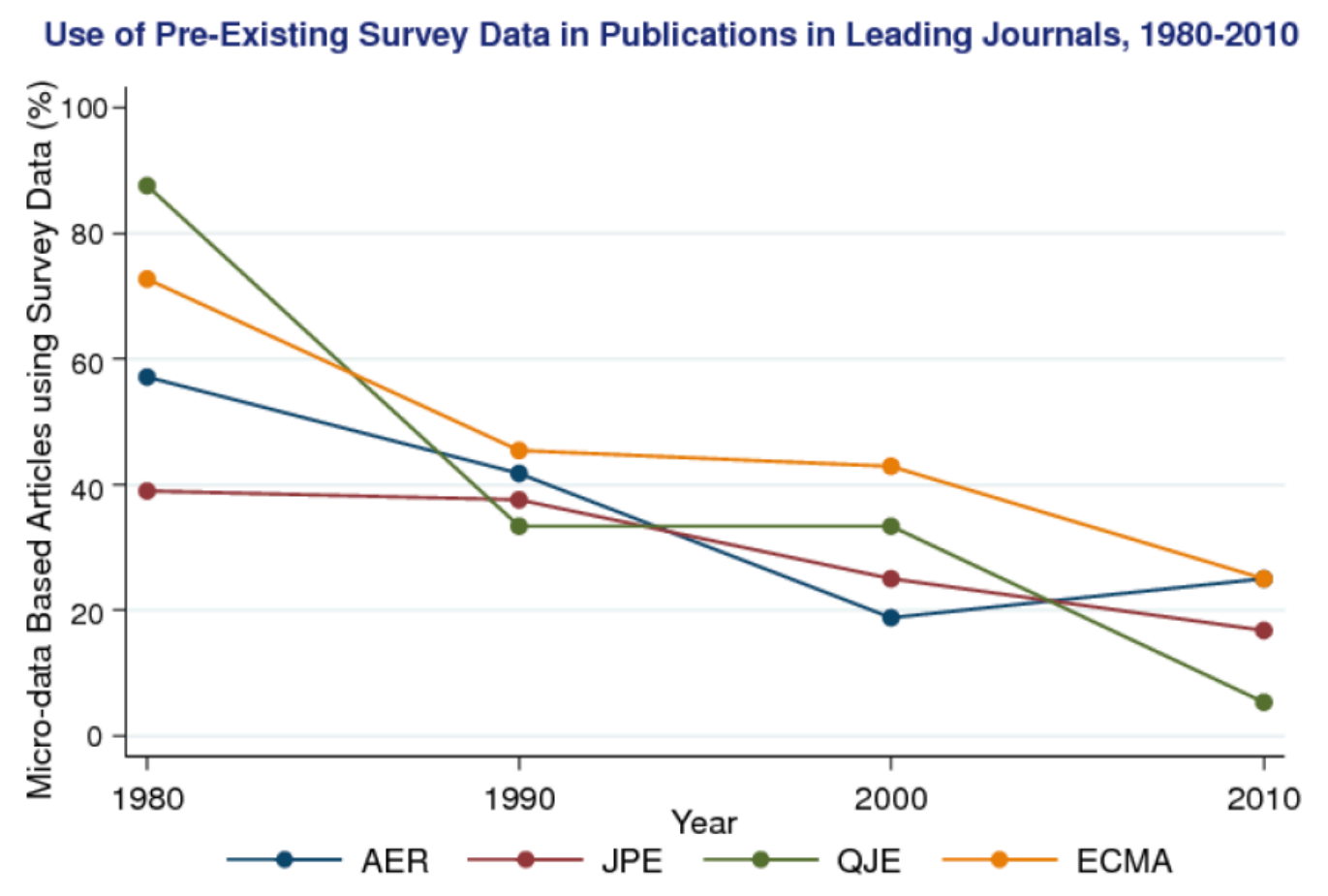


Increasing Use of Restricted Data in Research



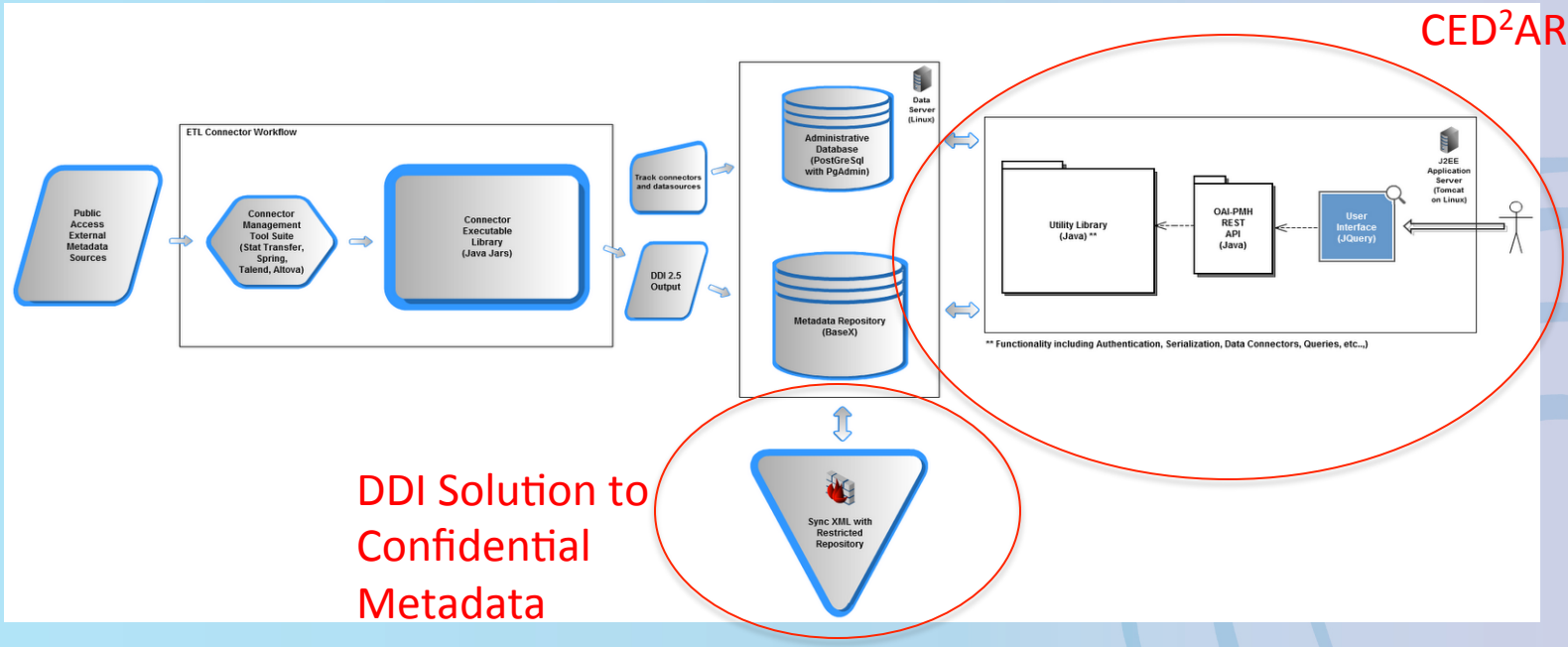
Source: Raj Chetty, http://conference.nber.org/confer/2012/SI2012/LS/ChettySlides.pdf

Use of Public Use Data Declining



Source: Raj Chetty, http://conference.nber.org/confer/2012/SI2012/LS/ChettySlides.pdf

Proposed Solution: Cornell's NCRN Node



Improved documentation and discoverability of both public and restricted data from the federal statistical system

CISER

CED²AR Overview and Goals

- Collect and standardize disparate metadata into a single DDI repository
- Provide a web interface for researchers to access
- Build an API for developers to use
- Use open standards
- Provide thorough documentation

Acknowledging CS 5150 Contributions

Jeremy Williams*
Benjamin Perry
Justin Burden
Chantelle Farmer
Shudan Zheng
Jessica Kane

*CISER and NCRN staff member, EDDI co-author, and coordinator of the CS5150 team

http://rschweb.ciserrsch.cornell.edu:8080/CED2AR_Web/

CISER

CED²AR Search API

The API Supports all of these query functions:

- Return
  - a chosen set of fields within the DDI schema
- Where
  - a chosen set of supported DDI search fields
  - and, or, and not
  - contains, starts-with, ends-with
- Sort
  - descending, ascending
- Limit
  - give me results 10-50 from each codebook

The API makes interacting with the repository easier because it abstracts away the underlying XQUERY necessary to perform the query.

CISER

CED²AR Search API

Some Example DDI things (resources):

- Codebooks:
  - http://rschweb.ciserrsch.cornell.edu:8080/CED2AR_Query/codebooks
- Codebook Named SSB
  - http://rschweb.ciserrsch.cornell.edu:8080/CED2AR_Query/codebooks/SSB
- Variables of Codebook Named SSB
  - http://rschweb.ciserrsch.cornell.edu:8080/CED2AR_Query/codebooks/SSB/variables
- A particular variable in the SSB Codebook named totfam_kids
  - http://rschweb.ciserrsch.cornell.edu:8080/CED2AR_Query/codebooks/SSB/variables/totfam_kids

CISER

CED²AR Search API

Ability to create complex queries across codebooks

- Give me all variables across all codebooks where the variable text contains the word 'house' and the variable label contains the word 'dwelling' but does not start with the word 'number' (and sort it backwards by variable name)
    - http://rschweb.ciserrsch.cornell.edu:8080/CED2AR_Query/search?return=variables&where=variabletext=*house*,variablelabel=*dwelling*,variablelabel!=number*&sort=-variablename

CISER

NCRN DDI Solution at the Variable Level: <dataAccs>

```
<stdyDscr>
    <citation> [8 lines]
    <dataAccs ID="A1">
        <useStmt>
            <conditions>Public</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A2">
        <useStmt>
            <confDec>To download this dataset, the user must obtain Special Sworn Status from the United States Census Bureau.</confDec>
            <conditions>Confidential</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A3">
        <useStmt>
            <confDec>You're never gonna see this data.</confDec>
            <conditions>Need to know</conditions>
        </useStmt>
    </dataAccs>
</stdyDscr>
```

Variable Level Solution (continued)

```
<var ID="V1500" dcml="0" files="F3" intrvl="discrete" name="totfam_kids" access="A1">
     <location width="12"/>
     <labl>Total Number of Children in Family</labl>
     <valrng> [2 lines]
     <sumStat type="vald">1000</sumStat>
     <sumStat type="invd">0</sumStat>
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <varFormat schema="other" type="numeric"/>
 </var>
 <var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A2">
     <location width="12"/>
     <labl>Total Personal Income</labl>
     <valrng> [2 lines]
     <sumStat type="vald">240</sumStat>
     <sumStat type="invd">760</sumStat>
     <sumStat type="min">-278.739</sumStat>
     <sumStat type="max">39515.631</sumStat>
     <sumStat type="mean">1861.779</sumStat>
     <sumStat type="stdev">4015.033</sumStat>
     <varFormat schema="other" type="numeric"/>
 </var>
```

No DDI Solution at the level of a *Value Label*

```
<var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A1">
    <location width="12"/>
    <labl>Total Personal Income</labl>
    <catgry>
        <catValu>0</catValu>
        <labl>5-25k</labl>
    </catgry>
    <catgry>
        <catValu>1</catValu>
        <labl>25-75k</labl>
    </catgry>
    <catgry>
        <catValu>2</catValu>
        <labl>75-125k</labl>
    </catgry>
    <catgry>
        <catValu>3</catValu>
        <labl>125-250k</labl>
    </catgry>
    <catgry access="A2">
        <catValu>4</catValu>
        <labl>250k+</labl>
    </catgry>
    <varFormat schema="other" type="numeric"/>
</var>
```

Small tweak to the DDI Codebook Schema would fix this.

Takk!

Spørsmål?

block@cornell.edu
ncrn.cornell.edu