IPUMS International: Building a census data time machine

Peter Clark, Minnesota Population Center
http://www.pop.umn.edu
What’s an IPUMS?

- Integrated Public Use Microdata Series
  - USA (1850-2008, not 1890)
  - International (machine-readable, post 1960)
  - Current Population Survey (CPS)
  - North Atlantic Population Project (NAPP)
  - Integrated Health Information Survey (IHIS)

- [http://www.ipums.org](http://www.ipums.org)
Worldwide Participation in IPUMS-I
<table>
<thead>
<tr>
<th>Add to cart</th>
<th>Variable</th>
<th>Variable Label</th>
<th>Type</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RELATE</td>
<td>Relationship to household head</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>ERELATE</td>
<td>Relationship to head, Europe</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>AGE</td>
<td>Age</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>AGE2</td>
<td>Age, grouped into intervals</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>SEX</td>
<td>Sex</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>MARST</td>
<td>Marital status</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>EMARST</td>
<td>Marital status, Europe</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>CONSENS</td>
<td>Consensual union</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>POLYGAM</td>
<td>Polygamous union</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>AGEMARR</td>
<td>Age at first marriage or union</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>MARRYR</td>
<td>Year of first marriage</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>MARRNUM</td>
<td>Number of marriages or unions</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>SUBFREL</td>
<td>Relationship to head of subfamily</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>SUBFNUM</td>
<td>Subfamily membership number</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>BIRTHYR</td>
<td>Year of birth</td>
<td>P</td>
<td>codes</td>
</tr>
<tr>
<td></td>
<td>BIRTHMO</td>
<td>Month of birth</td>
<td>P</td>
<td>codes</td>
</tr>
</tbody>
</table>
What’s so hard about that?
What’s so hard about that?

Censuses have different structures:
What’s so hard about that?

Censuses have different structures:
Canada (2001)
Censuses have different structures: Canada (2001)
What’s so hard about that?

Censuses have different structures:
Canada (2001)
USA (2010)

For over 300 years, the census has painted a picture of our people and the places in which we live. This census, on May 15, 2001, will continue the tradition by providing information needed by community groups, businesses and governments to develop plans for education and training, seniors’ housing, day care, fire protection, public transport, and many other programs that are important to all of us.

As Canada’s national statistics agency, Statistics Canada uses census data for producing statistical tables, analytical reports and for selecting samples or following up respondents for some of our surveys. These uses are strictly for statistical purposes and no one outside of the agency can have access to your identifiable information.

By law, Statistics Canada must take a census every five years, and every household must fill in a census form. Also, by law, Statistics Canada must protect the confidentiality of the personal information you provide. Our employees, including census takers, are personally liable to fines or imprisonment should they break the confidentiality of your information.

Please complete your census form and mail it back on May 15, 2001.

Thank you for your cooperation.

Ivan P. Fellegi
Chief Statistician of Canada

For information and instructions see the Guide 1 800 591-2001
TTY/TDD users call: 1 877 881-8301
Censuses have different structures:

Canada (2001)
USA (2010)
## Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---


---

### What’s so hard about that?

---

### Censuses have different structures:

- **Canada (2001)**
- **USA (2010)**
- **Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?

---

### Censuses have different structures:

**Canada (2001)**

**USA (2010)**

**Iraq (1997)**

---

### What’s so hard about that?
What's so hard about that?

Censuses have different structures:
- Canada (2001)
- USA (2010)
- Iraq (1997)
What’s so hard about that?

Censuses have different structures:
Canada (2001)  
USA (2010)  
Iraq (1997)  
Armenia (2001)
What’s so hard about that?

Censuses have different structures:
Canada (2001)  
USA (2010)  
Iraq (1997)  
Armenia (2001)
Adding data to IPUMS-I

- Data Acquisition
- Data Cleanup and Reformatting
- Data Integration
- Data Dissemination
The Problems: Acquisition
The Problems: Acquisition
The Problems: Cleanup (Physical)

1973 census tape #1: manual spacing required to recover data
The Problems: Cleanup (Logical)

• Reformatting - from whatever to Household/Person hierarchical format
• Draw Sample (if necessary)
• Apply initial confidentiality measures (if necessary)
• Scan original documentation, and translate to English
The Problems: Cleanup (Logical)

- Recode variables to solely numeric values
- Check codes and marginal frequencies
- Document what we did

<table>
<thead>
<tr>
<th>CR840018</th>
<th>label</th>
<th>cos1984</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Marital status</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
</tr>
</tbody>
</table>

<tt>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NIU</td>
<td>B=Under age 10</td>
</tr>
<tr>
<td>1</td>
<td>Consensual union</td>
<td>1=Consensual union</td>
</tr>
<tr>
<td>2</td>
<td>Married</td>
<td>2=Married</td>
</tr>
<tr>
<td>3</td>
<td>Separated</td>
<td>3=Separated</td>
</tr>
<tr>
<td>4</td>
<td>Divorced</td>
<td>4=Divorced</td>
</tr>
<tr>
<td>5</td>
<td>Widowed</td>
<td>5=Widowed</td>
</tr>
<tr>
<td>6</td>
<td>Single</td>
<td>6=Single</td>
</tr>
<tr>
<td>9</td>
<td>Missing</td>
<td>0=[undocumented]</td>
</tr>
<tr>
<td>9</td>
<td>&quot;</td>
<td>8=[undocumented]</td>
</tr>
</tbody>
</table>
</tt>
The Problems: Data Integration

- Variable integration
- Constructed variables
- Variable documentation and descriptions
## Integration - Marital Status?

<table>
<thead>
<tr>
<th>China 1982</th>
<th>Columbia 1973</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never married</td>
<td>Consensual Union</td>
</tr>
<tr>
<td>Married</td>
<td>Married</td>
</tr>
<tr>
<td>Widowed</td>
<td>Separated or divorced</td>
</tr>
<tr>
<td>Divorced</td>
<td>Single</td>
</tr>
<tr>
<td></td>
<td>Widowed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Kenya 1989</th>
<th>Mexico 1970</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>Married: Civil and Religious</td>
</tr>
<tr>
<td>Married</td>
<td>Married: Only Civil</td>
</tr>
<tr>
<td>Monogamous</td>
<td>Married: Only Religious</td>
</tr>
<tr>
<td>Polygamous</td>
<td>Consensual Union</td>
</tr>
<tr>
<td>Widow</td>
<td>Widow</td>
</tr>
<tr>
<td>Divorced</td>
<td>Divorced</td>
</tr>
<tr>
<td>Separated</td>
<td>Separated</td>
</tr>
<tr>
<td></td>
<td>Single</td>
</tr>
</tbody>
</table>
## Integration: Table-driven

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>code</td>
<td>label</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>SINGLE/NEVER MARRIED</td>
<td>1=never married</td>
<td>4=single</td>
<td>1=single</td>
<td>9=single</td>
<td>6=never married</td>
</tr>
<tr>
<td>200</td>
<td>MARRIED/IN UNION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>210</td>
<td>Married (not specified)</td>
<td>2=married</td>
<td>2=married</td>
<td>3=monogamous</td>
<td></td>
<td>1=married</td>
</tr>
<tr>
<td>211</td>
<td>Civil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3=only civil</td>
</tr>
<tr>
<td>212</td>
<td>Religious</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4=only religious</td>
</tr>
<tr>
<td>213</td>
<td>Civil and religious</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2=civil and religious</td>
</tr>
<tr>
<td>214</td>
<td>Polygamous</td>
<td></td>
<td></td>
<td>3=polygamous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>220</td>
<td>Consensual union</td>
<td>1=free union</td>
<td></td>
<td>5=free union</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>SEPARATED/DIVORCED</td>
<td>3=sep. or divorced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>310</td>
<td>Separated</td>
<td></td>
<td>6=separated</td>
<td>8=separated</td>
<td>3=separated</td>
<td></td>
</tr>
<tr>
<td>321</td>
<td>Legally separated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>322</td>
<td>De facto separated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>330</td>
<td>Divorced</td>
<td>4=divorced</td>
<td></td>
<td>5=divorced</td>
<td>7=divorced</td>
<td>4=divorced</td>
</tr>
<tr>
<td>400</td>
<td>WIDOWED</td>
<td>3=widowed</td>
<td>5=widowed</td>
<td>4=widowed</td>
<td>6=widowed</td>
<td>5=widowed</td>
</tr>
<tr>
<td>999</td>
<td>UNKNOWN/MISSING</td>
<td>0=missing</td>
<td>6=unknown</td>
<td>B=blank</td>
<td>1=unknown</td>
<td></td>
</tr>
</tbody>
</table>

**Countries:**
- China 1982
- Colombia 1973
- Kenya 1989
- Mexico 1970
- U.S.A. 1990
VAR: yrschl
   // VAR: yrschl = Years of schooling.
switch(dataSet){
   case ke1989a: {
      int d = ke89a412(ln); // educ attainment
      int e = ke89a411(ln); // whether in school or ever attended
      int g = ke89a403(ln); // age
      if (g >= 20 && e != 1){
         // For persons age 20+ who are not currently in
         // school(?), reduce their years of schooling by 1 year
         // if their educational attainment was Form1
         // through Form4.
         switch(d){
            case 11: a = 8;  break;
            case 12: a = 9;  break;
            case 13: a = 10; break;
            case 14: a = 11; break;
         }
      }
   } break;
MARST describes the person's current marital status according to law or custom. Individuals who remarried should report the status relevant to their most recent marriage. Census instructions rarely explicitly limit marital status to strictly legal unions.

The first digit of MARST is largely comparable across all samples. The second digit differentiates consensual unions from other forms of marriage and distinguishes among the categories separated, divorced, and married-spouse-absent. The final digit provides additional detail within the married and separated categories.

The divorced category includes both legally registered and unregistered divorces. All persons under age 15 are coded single.
Dissemination: Policy

• USA, CPS, IHIS - all US Domestic data
  – freely available to anyone

• IPUMS-International & NAPP
  – data usage governed by agreements with donor countries
  – only for academic use, must have an approved application to use

• No cost to use any MPC data project
Research Impact

- 5556 registered IPUMS-I users
- 36909 registered IPUMS USA users
- 526 bibliography citations for IPUMS-I
- 284 Google Scholar hits for IPUMS-I
- 4390 hits in Google Scholar for IPUMS
  - For comparison: 390K for Mathematica
  - and 1060 hits for “wolfram alpha”
FIGURE 7. Cumulative distribution of siblings of children aged 9-11, eight countries: Proportion with specified number of surviving children or fewer
SOURCE: Estimates from IPUMS-International census microsamples (Minnesota Population Center 2007).
FIGURE 7. Cumulative distribution of siblings of children aged 9-11, eight countries: Proportion with specified number of surviving children or fewer
SOURCE: Estimates from IPUMS-International census microsamples (Minnesota Population Center 2007).
FIGURE 7. Cumulative distribution of siblings of children aged 9-11, eight countries: Proportion with specified number of surviving children or fewer
SOURCE: Estimates from IPUMS-International census microsamples (Minnesota Population Center 2007).
FIGURE 7. Cumulative distribution of siblings of children aged 9-11, eight countries: Proportion with specified number of surviving children or fewer
SOURCE: Estimates from IPUMS-International census microsamples (Minnesota Population Center 2007).
Future

• Expecting to add another 100 samples over the next three years to IPUMS-International

• Word/Excel-based metadata system getting creaky
  – New metadata system (currently under development) will store metadata directly in database, allowing for checkin/checkout for editing.

• Use column-oriented db for actual data files

• Better leveraging of Big Data tools - hadoop?
Thank you!

Peter Clark
pclark@umn.edu

http://www.pop.umn.edu
http://www.ipums.org
Integration: Constructed Variables

(Simple household)

<table>
<thead>
<tr>
<th>Pernum</th>
<th>Relate</th>
<th>Age</th>
<th>Sex</th>
<th>Marst</th>
<th>Chborn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>head</td>
<td>46</td>
<td>male</td>
<td>married</td>
<td>n/a</td>
</tr>
<tr>
<td>2</td>
<td>spouse</td>
<td>44</td>
<td>female</td>
<td>married</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>aunt</td>
<td>77</td>
<td>female</td>
<td>widow</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>child</td>
<td>15</td>
<td>female</td>
<td>single</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>child</td>
<td>13</td>
<td>female</td>
<td>single</td>
<td>n/a</td>
</tr>
<tr>
<td>6</td>
<td>child</td>
<td>11</td>
<td>male</td>
<td>single</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Spouse’s Location

<table>
<thead>
<tr>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Mother’s Location

<table>
<thead>
<tr>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

Father’s Location

<table>
<thead>
<tr>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

(Colombia 1985)