



**IPUMS-International:  
Lessons from 10 years of archiving and  
disseminating census microdata**

**\* \* \***

**Robert McCaa and Wendy L. Thomas  
University of Minnesota Population Center**

**[rmccaa@umn.edu](mailto:rmccaa@umn.edu)**

**[www.ipums.org/international](http://www.ipums.org/international)**

**for additional details, please see:**

**[www.hist.umn.edu/~rmccaa/ipums-global](http://www.hist.umn.edu/~rmccaa/ipums-global)**



# Outline. Lessons from 10 years of archiving and disseminating microdata

	no. of slides
» <b>IPUMS-International: 4 goals</b>	<b>2</b>
» <b>Lesson 1: desire to preserve &amp; disseminate</b>	<b>3</b>
» <b>Lesson 2: data &amp; docs are recoverable</b>	<b>5</b>
» <b>5 steps in integrating metadata and microdata</b>	<b>5</b>
» <b>Lesson 3: data &amp; docs can be integrated</b>	<b>3</b>
» <b>Lesson 4: researchers demand microdata</b>	<b>1</b>
» <b>Conclusion</b>	<b>2</b>



## IPUMS-International: Goals

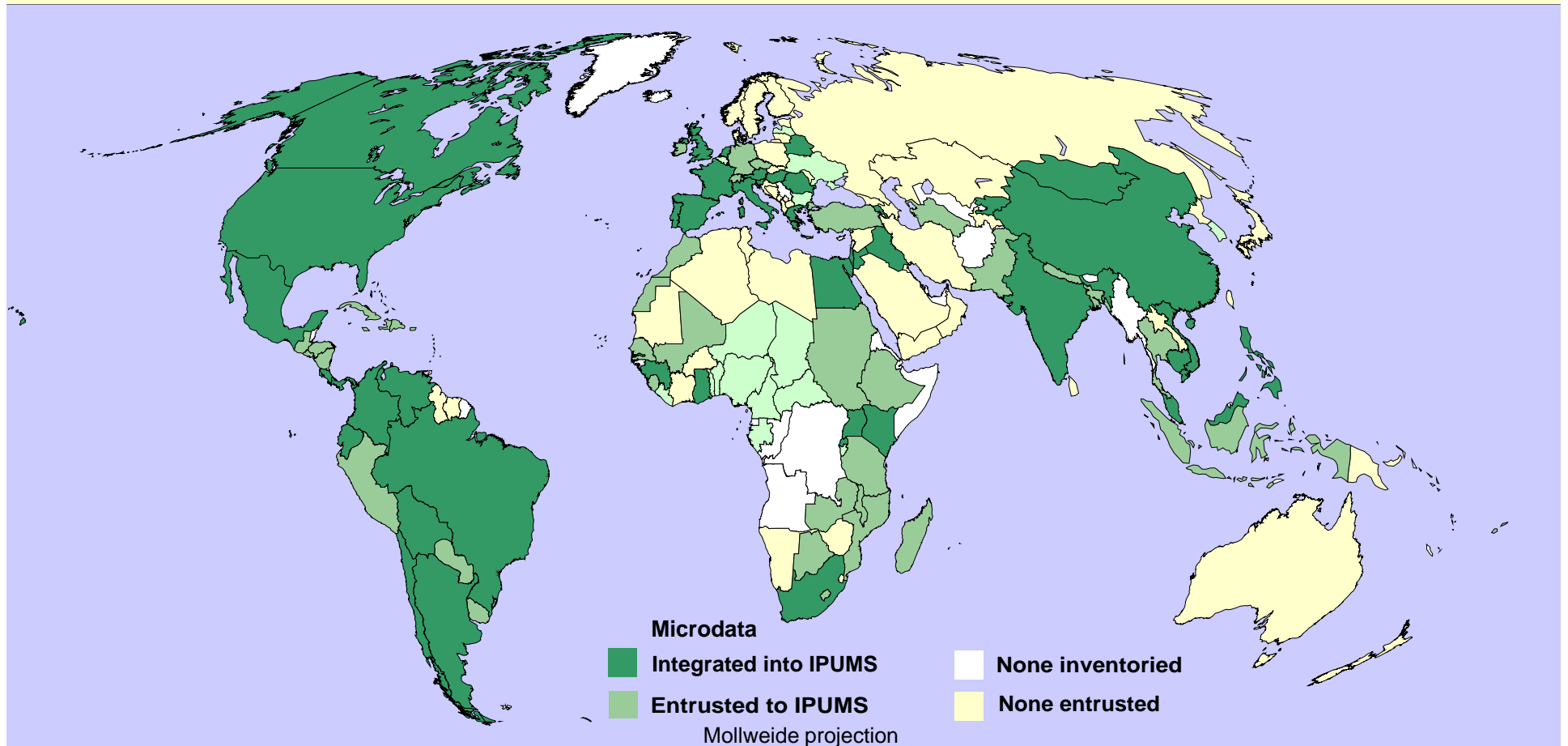
- 1. Inventory census microdata and documentation, world-wide**
  - 2. Recover and preserve at-risk microdata**
  - 3. Integrate census microdata and documentation**
  - 4. Disseminate--without cost--extracts of samples to bona-fide researchers worldwide, regardless of country of birth, citizenship or residence.**
- » **Sustained funding 1999-2015—6 grants of 5 years duration:**
- » **National Science Foundation (USA): 3 successive grants**
  - » **National Institutes of Health (USA): Latin America, Europe, Eur-Asia (Why not Africa?—not enough cooperation from African NSOs)**

## *IPUMS-International*

*dark green = integrated and disseminating*

*(44 countries, 130 censuses, 279 million person records)*

*green = to be integrated (35 countries, 90 censuses, 150 mill.)*





## **Lesson 1: World-wide desire to preserve and disseminate microdata**

- » **Many statistical offices desire to preserve microdata, but may lack technical, material or human resources**
- » **Africa Addendum to UN Principles & Recommendations on Censuses emphasizes importance of archiving**
- » **IPUMS-International offers assistance**
  - » **Legal - uniform Memorandum of Understanding**
  - » **Material – old tape drives and cleaners**
  - » **Technical – software for recovery and conversion**
  - » **Human – experienced personnel to recover bits & bytes, convert binary to ASCII, and retro-engineer codebooks, if needed.**



## All Latin American National Statistical Offices participate: IPUMS Workshop, Panama City, June 2-5, 2008





## Archiving and preserving census microdata

- » IPUMS team is lead by historians—time is our friend
- » 2000: Handbook of International Census Microdata for Population Research
- » Recovery of census microdata from the 1960s, 70s, & 80s—funded by IPUMS-International project
  - » Muller Media Co. recovers microdata for IPUMS
  - » Biggest recovery: Bangladesh 1981, ~300 tapes
  - » Others: Mali, 1976; Romania 1977, Germany (DR) 1971
  - » Recovery from paper: Fiji 1976, Mongolia 1989; Morocco 1961?
- » Microdata are migrated to archival format (ASCII), documented as thoroughly as feasible, and checked for quality.



## **Lesson 2: Metadata and Microdata recovery efforts are often successful**

- » **Metadata preservation (original source documentation)**
  - 1. Census documentation entrusted to the Minnesota Population Center: census forms, manuals, codebooks, data dictionaries, methodological reports, etc.**
    - » **US Census Bureau International Programs**
    - » **UN Statistics Division historical archive**
    - » **Rand-McNally Publishing Co.**
  - 2. Subcontracts to organize and scan archives funded by IPUMS-International**
    - » **East-West Center (Hawaii)**
    - » **Centre de Population & Development (CEPED Paris)**
    - » **CELADE: UN Latin America and Caribbean Center for Demography**
- » **Metadata disseminated on CD and/or internet—free of cost**
  - » **Copies of scanned metadata repatriated to each national statistical office**
  - » **Census Forms, 1960-2000 (56<sup>th</sup> ISI edition)**
  - » **Africa Census documentation (57<sup>th</sup> ISI edition)**
  - » **Latin America: enumerator manuals (U of M Digital Conservancy)**



# Bangladesh Bureau of Statistics Tape Archive

April 14, 2006





## Bangladesh Bureau of Statistics Tape Archive Data recovery in Dhaka, Sep. '08 (funded by IPUMS)





## Bangladesh Bureau of Statistics Tape Archive Data recovery hardware and software installed





# Bangladesh Bureau of Statistics Tape Archive Data recovery in Dhaka, Sep. '08 “Cleaner”





## **Constructing the IPUMS integrated metadata and microdata system**

- » ***IPUMS does NOT hand out CDs of source microdata***
- » **5 step process of integration--2+ years invested in integrating metadata and microdata:**
  1. **\*Confirm the integrity and validity of source microdata and metadata**
  2. **\*Draw and anonymize high precision samples**
  3. **Integrate microdata**
  4. **Integrate metadata**
  5. **Confirm the integrity and validity of the integrated microdata sample and metadata**
- » **\*Steps 1 & 2 conducted by commissioned senior staff**
  - » **Original source microdata never disseminated**
  - » **Violation of confidentiality: subject to civil fine (\$250,000) and/or criminal prosecution**
  - » **Dennis Trewin on-site inspection of IPUMS: “standard of the best statistical offices”**



## 5 step process of integration in the IPUMS system

1. **Confirm the integrity and validity of source microdata and metadata**
  - **Confirm record structure**
  - **Confirm that sample statistics approximate official figures**
  - **Assemble as complete documentation as possible**
  - **Note: not all datasets pass these tests—for others 3 or more years may be required**
2. **Draw and anonymize high precision samples**
  - **Uniform sample design—every  $n^{\text{th}}$  ( $10^{\text{th}}$ ) household, where feasible**
    - **Implicit geographical stratification yields the greatest precision, the most robust, general purpose sample design**
    - **To date, 37 NSOs entrusted 100% microdata for IPUMS to process**
    - **Uniform anonymization procedures (see McCaa, et. al., 2006)**
      - **Suppress low level geography (NUTS4 and below) and other “risky” variables**
      - **Suppress low frequency codes and any “risky” categories**
      - **Randomly swap households across geographical boundaries**
  - **Samples drawn by NSO partner: anonymize upon request**
- ...



## 5 step process of integration in the IPUMS system

### 3. Integrate microdata

- **Composite coding scheme to**
  - 1) **preserve every significant detail and**
  - 2) **harmonize every code**
- **Example: marital status**
  - ...
  - **200 = married**
  - **210 = married, formal**
  - **211 = married, civil**
  - **212 = married, religious**
  - ....
  - **220 = married, informal (consensual)**
  - ...
- **National partners may construct “national-flavored” integrations from the IPUMS-International integrations**



## **5 step process of integration in the IPUMS system**

### **3. Integrate microdata**

### **4. Integrate metadata (XML): Document every census, sample, variable and code:**

- **Source documents (pdf) in official language and English**
- **Dynamic metadata system—compare any combination of countries and samples:**
  - **wording of any census question and instructions to field workers**
- **Characteristics of each census and sample**
- **Describe each variable: “universe”, definition, comparability, etc.**



## **5 step process of integration in the IPUMS system**

### **5. Confirm integrity and validity of each sample**

- **Before launch, each sample is scrupulously checked**
- **Task facilitated by comparing integrated against non-harmonized variables (toggle on variable selection page)**
  - **Researchers gain a more comprehensive understanding of the integration process**
  - **Each integration decision may be checked by any researcher**
- **External evaluation by INDEC-Argentina (commissioned by IPUMS), 4 censuses (1970-2001)**
  - **Compared each variable, code and metadata against original source data and documentation**
  - **Tens of thousands of words, codes, and frequencies tested—only a handful of errors, mis-interpretations or mis-understandings.**



## **Lesson 3: Integration is feasible**

- » **The MPC team is integrating 15-20 censuses per year on average**
  - » **To the highest standards of National Statistical Office partners**
  - » **To the great satisfaction of thousands of users world-wide**
- » **Currently 130 samples are integrated into the IPUMS-International database, representing 44 countries**
- » **Over the next 5 years, the database will expand by one-half or more.**



# The IPUMS team

## May 14, 2009



**(Not present: computer gurus, some researchers, research assistants, civil service employees, and others who were absent from the National Science Foundation Board meeting)**



## **IPUMS dissemination plan, 2010-2014** **samples for 44 countries integrated now, 80 soon**

- » **Europe 12 completed, 5 soon**
  - » **Completed (12):** *Austria, Belarus, France, Greece, Hungary, Italy, Netherlands, Portugal, Romania, Slovenia, Spain, UK*
  - » **Soon (5):** *Czech Republic, Germany, Ireland, Switzerland, Turkey*
- » **Americas 12:12**
  - » **Completed (12):** *Argentina, Bolivia, Brazil, Canada, Chile, Colombia, Costa Rica, Ecuador, Mexico, Panama, USA, Venezuela*
  - » **Soon (12):** *Cuba, Dominican Republic, El Salvador, Guatemala, Honduras, Jamaica, Nicaragua, Paraguay, Peru, Puerto Rico, Saint Lucia, Uruguay*
- » **Africa 7:13**
  - » **Completed (7):** *Egypt, Ghana, Guinea (Conakry), Kenya, Rwanda, South Africa, Uganda*
  - » **Soon (13):** *Botswana, Ethiopia, Madagascar, Malawi, Mali, Mauritius, Morocco, Mozambique, Senegal, Sierra Leone, Sudan, Tanzania, Zambia*
- » **Asia 13:7**
  - » **Available (13):** *Armenia, Cambodia, China, India, Iraq, Israel, Jordan, Kyrgyz Republic, Malaysia, Mongolia, Palestine, Philippines, Vietnam*
  - » **Soon (7):** *Bangladesh, Fiji, Indonesia, Nepal, Pakistan, Thailand, Turkmenistan*



## **Lesson 4: Researchers demand microdata 3,000 accredited researchers—visit IPUMS booth for demo**

- » **3,000 accredited researchers. Country rankings (n=76):**
  - » **USA (1,738), UK (118), Mexico, Brazil, Canada, France, Spain, Colombia (64), Germany, China, Italy, Argentina, Switzerland, Australia (26), Japan, India, Chile, Kenya, Greece, Austria, Netherlands, Belgium (12), Romania, Singapore, South Africa, Ecuador, Ireland, Israel (9), Philippines, Sweden, Uganda, Hong Kong + 44 countries (<5)**
- » **To become accredited submit electronic application**
  - » **Approval is granted to analysts with projects which require access to the data and who agree to the conditions of use.**
- » **Once accredited, make an extract:**
  - » **Select country(ies), year(s), sample density, sub-population(s), and variables**
  - » **Submit selections**
  - » **Receive email when extract is ready**
- » **Download extract and analyze with favorite statistical software**
  - » **Email questions to help desk**



## Conclusion

### » Thanks to:

- » National Statistical Offices for trust and cooperation
- » International organizations for support and encouragement
- » Researchers for enthusiastic adoption of IPUMS integrated datasets

### » Invitation to:

- » National Statistical Offices that are not yet participating
- » Researchers who need census microdata for research:  
[www.ipums.org/international](http://www.ipums.org/international)
- » And...



**...to the 58<sup>th</sup> ISI: Dublin, Aug 21-26, 2011**  
**<http://www.isi2001.ie>**

## ISI 2011



Home

About the ISI

About the CSO

Organising  
Committee

Scientific  
Programme

Registration

About Dublin

Search:

It is with great pleasure that I invite you to the 58th Session of the International Statistical Institute, which will be held in Dublin in August 2011. Ireland is a unique destination, noted for its hospitality, and is guaranteed to offer delegates a creative and memorable experience. Dublin itself is a modern and vibrant city with a long and proud history and has a wonderful array of venues and activities to enjoy. It is also our intention to offer a dynamic Social Programme to allow you to sample our rich culture and heritage.

I look forward to extending the traditional Irish C  ad M  ile F  ilte (one hundred thousand welcomes) to you all to Dublin in 2011 and sincerely hope that you will take this opportunity to visit us on the occasion of the 58th ISI.

**Gerard O'Hanlon (Director General, Central Statistics Office, Ireland)**  
**Chairman - ISI 2011 National Organising Committee**

- » **IPUMS Workshop, Aug 19-20**
- » **Microdata sessions**
- » **IPUMS Funding for delegates from developing countries**
- » **IPUMS booth**



**Thank you!!**

---

**rmccaa@umn.edu**  
**www.ipums.org/international**

---