Technology in data dissemination: IPUMS-International disseminates integrated census microdata and metadata

* * *

Robert McCaa and Patricia Kelly-Hall

Minnesota Population Center

rmccaa@umn.edu; pkelly@umn.edu

www.ipums.org/international

3 goals of presentation

1. Describe IPUMS global integration census microdata and metadata system
   - » Integration: analyze change over time
   - » Free access for bona-fide researchers
   - » No risk for national statistical offices
   - » IPUMS: anonymizes, integrates and manages access
2. Invite: 22 current partners (see map)
   - » to entrust 2010 round census data and documentation
   - » to avoid delay
3. Invite: 31 not yet participating to cooperate in the IPUMS global initiative
   - » Endorse memoradum of understanding, purchase order
   - » Recover historical census data and documentation
   - » Entrust microdata and documentation

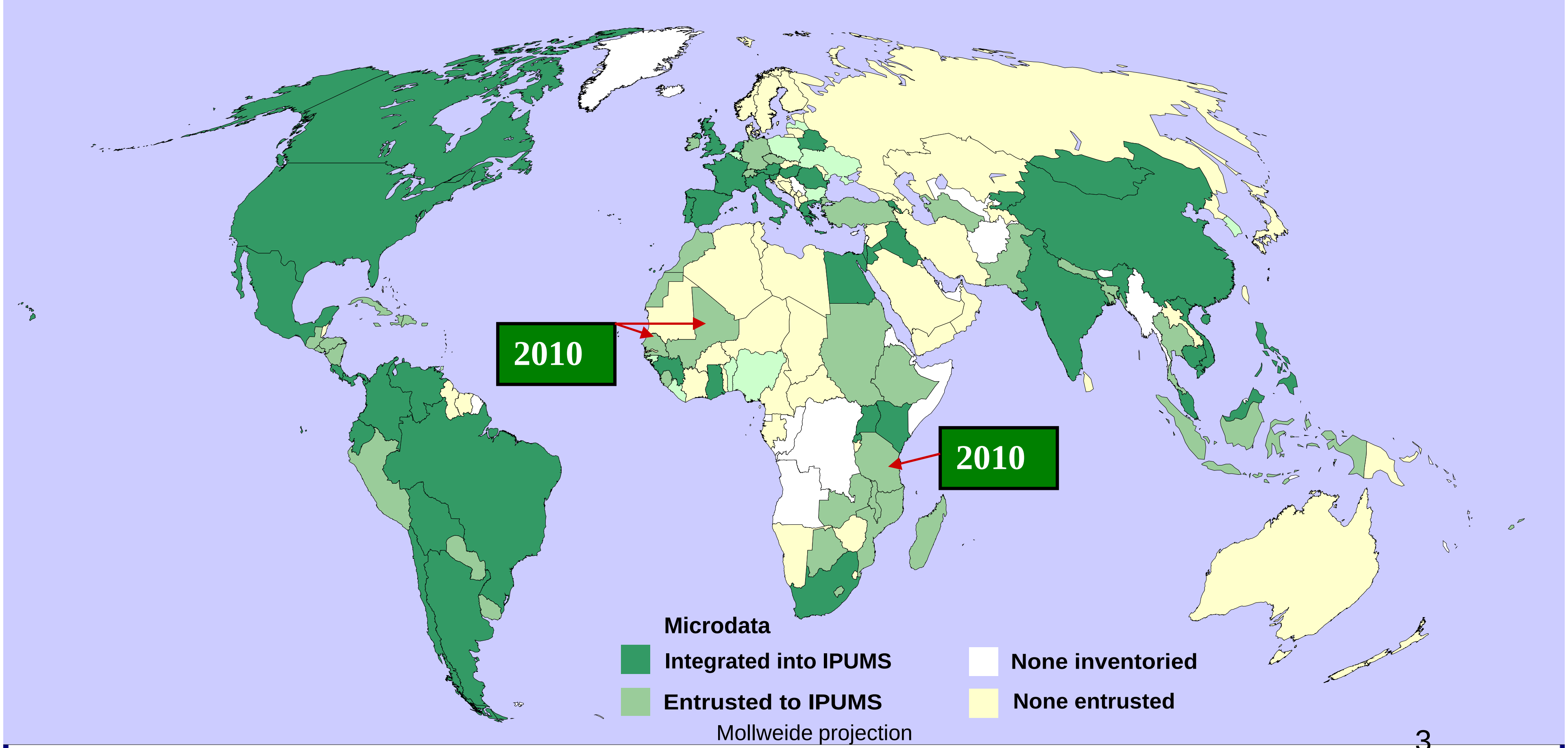
IPUMS-Global (first 10 years)

dark green = integrated and disseminating

(44 countries, 130 censuses, 279 millon person records)

green = to be integrated (36 countries, 93 censuses, 160 mill.)

2010: 10 countries to be integrated, including Mali, Senegal, Tanzania



Outline: IPUMS-International integrated census microdata and metadata

no. of slides

1. **IPUMS-International: “Best practice”** 4
2. **Constructing the IPUMS integrated system** 4
3. **Census output needs:** 5
   - a. **Form “A”: succinct descriptions of both census and microdata**
   - b. **Metadata: questionnaires, instructions, dictionaries, codebooks as images, .txt, .doc, .xls, .pdf, XML, SDMX, CSPro, IMPS, DDI, etc.**
   - c. **Microdata: to prepare, choose 1 of 4 modalities; entrust as encrypted, executable files (email or fax password)**
4. **Conclusion** 5
   - » **IPUMS strengths**
   - » **Africa Time-Line**
   - » **Invitation to Cooperate**
   - » **Invitation to 58th ISI, Dublin, Aug 2011**

What is IPUMS-International?

"…best practice
for a data repository of international statistical data"
--Dennis Trewin

chair UNECE task force on Statistical Confidentiality & Microdata Access

IPUMS Objectives

1. **Preserve census microdata and documentation for all the countries in the world**
2. **Integrate microdata and metadata**
   --a CD of original source microdata is not sufficient
3. **Disseminate--without cost--extracts of samples to bona-fide researchers worldwide, regardless of country of birth, citizenship or residence.**

- » **Sustained, major funding since 1999 by:**
  - » **National Science Foundation (USA)– renewed through 2014**
  - » **National Institutes of Health (USA)– eager to fund IPUMS-Africa When will IPUMS-Africa begin? … when 30 African NSOs endorse Memorandum of Understanding**

IPUMS-International:

- Begun in 1999, IPUMS-International is the world’s largest integrated demographic database:
  - 130 integrated, anonymized census samples (44 countries)
  - 279 million person records; 3,000+ approved researchers
- Database is likely to double over the next five years:
  - 36 countries currently in development
  - 10 current partners: Egypt, Ghana, Guinea, -Conakry, Kenya, Mali, Rwanda, Senegal, South Africa, Tanzania, Uganda
  - 12 African countries currently in development: Botswana, Ethiopia, Guinea-Bissau, Lesotho, Madagascar, Malawi, Mauritius, Morocco, Mozambique, Sierra Leone, Sudan, Zambia
  - 31 Future partners? Algeria, Angola, Benin, Burkina Faso, Burundi, Cameroon, Central Africa Republic, Chad…Zimbabwe

Benefits of IPUMS-International

- ***Preservation – IPUMS provides material and technical resources***
  - ***Recover*** **historical census data and documentation**
  - ***Archive*** **data and documentation to the highest international standards**
- ***Integration – IPUMS does the work***
  - ***Draw*** **high-precision samples to uniform specifications**
  - ***Anonymize*** **microdata to highest international standards**
  - ***Integrate*** **samples according to national practices *and* international principles**
- ***Dissemination – IPUMS manages the risk***
  - ***License*** **samples and documentation in a global initiative (US$5,000 per census of 1 million or more person records)**
  - ***Disseminate*** **microdata with minimal risk and maximum benefit, at no cost**
- ***Rewards – shared by all***
  - ***Promote*** **transparency and equitable access**
  - ***Facilitate*** **highest quality research and fact-based policy making**
  - ***Increase*** **stakeholders and use of census data**
- ***Participate*** **in a global initiative – 2011 workshop in Dublin, Ireland**

Constructing the IPUMS-International integrated metadata and microdata system

- » ***IPUMS-International NEVER disseminates source microdata!***
- » **5 step process of integration—**
  ***2+ years to integrate* metadata and microdata:**

1. **Confirm the integrity and validity of source microdata and metadata**
2. **Draw and anonymize high precision samples**
3. **Integrate microdata sample (next slide)**
4. **Integrate metadata (following slide)**
5. **Confirm the integrity and validity of the integrated microdata sample and metadata**

Step 3 of integration in the IPUMS system

- Composite coding scheme:
  1) preserve every significant detail and
  2) harmonize every code
- Example: marital status
  - …
  - 200 = married/in union
  - 210 = married, formal
  - 211 = married, civil
  - 212 = married, religious
  - ….
  - 215 = traditional or customary
  - 217 = polygamous
  - …
  - 220 = married, consensual union
  - …

Step 4: integrate metadata

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.**

Step 5: confirm integrity and validity

5. Confirm integrity and validity of integration
    - Before launch, each sample is scrupulously checked
    - Test each integrated variable against non-harmonized
        - Each integration decision may be checked by any researcher using integrated vs. non-harmonized
    - 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.

Census Output Needs:

1. Succinct description of census and microdata (Form "A")

2. Comprehensive metadata:

questionnaires, instructions, codebooks, official results

3. Encrypted microdata

Ship FEDEX prepaid by receiver to:

Prof. Robert McCaa
Minnesota Population Center
50 Willey Hall, 225 19th Ave. S.
Minneapolis MN 55455
Tel. 1+612.624.5818, rmccaa@umn.edu

1. Need for succinct, authoritative documentation of census and microdata: Form “A”

- Efficient processing of metadata & microdata
- Form “A”:
  - See Appendix A for details
  - Appendix B is the completed form for South Africa--censuses of 1996, 2001, 2007
  - https://international.ipums.org/international/samples.shtml click the name of a country to view samples
- Describe the census: name, population universe, reference date, field work period, etc.
- Describe the microdata: source, sample design, sample unit, sample fraction, size, weights, etc.
- Define units in the microdata: private household, collective dwelling, included/excluded populations, etc.

2. Metadata needs
see paragraphs 12-19 for additional details

» **Documents in any form: .pdf, .txt, .doc, .xls, .pdf, XML, SDMX, DDI, CSPro, IMPS, etc.**

» **Copies in official language and English: Essential:**

1. **Questionnaires**
2. **Instructions to interviewers**
3. **Codebooks, data dictionaries**

Helpful:

4. **Correspondence tables (e.g., occupation with ISCO08/88)**
5. **Summary official results**
6. **Technical, methodological reports**
7. **Sample design: preferred, every tenth private household; for collective dwellings (e.g., hospitals), every tenth person.**
8. **Boundary files for administrative geography coded in microdata**

3. Microdata needs
see paragraphs 20-27 for additional details

» 2 goals:

1. Permanently archive source microdata
   - » against loss
   - » copies exclusively for the National Statistical Agency owner
2. Integrate high precision, anonymized household samples

» Preferred modality: 100% microdata to IPUMS

- » License fee: US$5,000 per census of 1 million plus person records
- » Note: some European (!) statistical offices can no longer locate census microdata for 1960s, 1970s, 1980s and even 1990s!
- » Or where they can locate it, the data can not be made useable

3. Microdata needs
see paragraphs 24-30 for additional details

- High precision, household samples
  - 10 percent: 70 of 130 samples currently available
- Systematic random samples :
  - every nth private household after a random start
  - Collective dwellings: every nth person
  - extremely fine geographic stratification with proportional weighting
  - NUTS-2, NUTS-3
- Anonymization, performed by MPC
  In addition to sampling, 6 layers of technical protections:

1. Suppress small places or residence, work, school, etc.
2. Suppress codes of social categories with small counts
3. Top and Bottom coding of continuous variables
4. Suppress sensitive variables
5. Swap small % of households into different place of residence
6. Randomly order all household

Conclusion:
IPUMS strengths
Continent of Africa Timeline
Invitation to cooperate
Invitation to 58th ISI: Dublin (Aug 2011)

IPUMS-International strengths: cooperation—national, regional, global

1. Uniform legal authorization with national statistical authorities
2. Access restricted to academics with need who agree to abide by stringent confidentiality protections. Sanctions against individual and institution—denial of access to all microdata for the entire institution
3. Strong technical methods of microdata anonymization
4. Experienced integration teams
5. Proven web-based access management system
6. High producer and user satisfaction
7. Sustainable: MPC, NSF, NIH

IPUMS-International time-line: The continent of Africa

-2009: Egypt, Ghana, Guinea-Conakry, Kenya, Rwanda, South Africa, and Uganda plus 37 countries beyond Africa

2010: Mali, Senegal, Tanzania plus 7 non-African countries

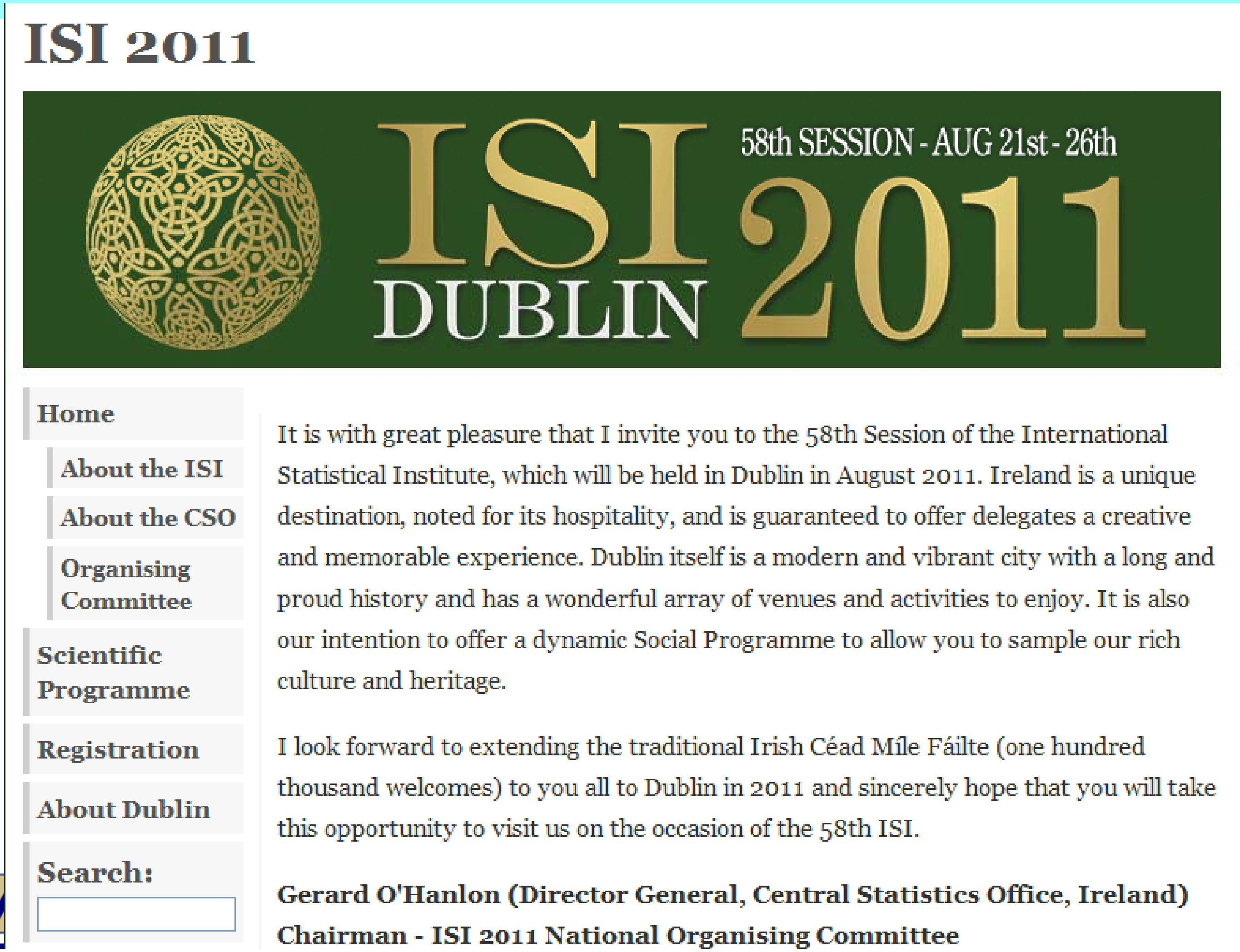
2011+: candidates (all have endorsed IPUMS MOU): Botswana (lack 1981), Ethiopia (lack 2007), Guinea Bissau (just beginning), Lesotho (lack 2006, 1986), Madagascar (lack codebooks 1993), Malawi (lack 2008), Mauritius (lack 1 census form), Morocco (just beginning), Mozambique (just beginning), Sierra Leone (clean 2004 data), Sudan (lack 2008), Zambia (lack 1980?)
--scheduling depends upon when comprehensive microdata/metadata are entrusted

Invitation to cooperate:

- 1. Formalities, endorse:
  - A. Memorandum of Understanding
  - B. Purchase order/Invoice
- 2. Entrust:
  - A. Documentation: instruction manuals, codebooks, etc.
  - B. Microdata: archive 100%
- 3. Payments
  - License fee: $5,000 per census
  - Data recovery (additional funds): 1981? …old tapes, data cartridges, codebooks?
- 4. Integration: MPC
- 5. Dissemination (10% samples): 2011?, 2012?, 2013?
- 6. and beyond: as long as mutually beneficial

…to the 58th Session ISI: Dublin, Aug 21-26, 2011

http://www.isi2001.ie



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 Céad Míle Fáilte (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
pkelly@umn.edu

www.ipums.org/international