Entrusting census microdata and metadata for timely integration and dissemination via the IPUMS-EurAsia and IECM initiatives, 2010-2014

* * *

Robert McCaa, Albert Esteve and Patt Kelly-Hall
Minnesota Population Center and Centre d'Estudis Demogràfics

rmccaa@umn.edu; aesteve@ced.uab.es

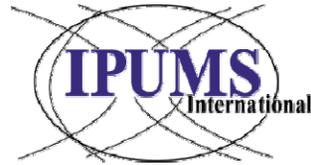www.ipums.org/international

www.iecm-project.org

MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Outline:
Entrusting census microdata and metadata for timely integration and dissemination via the IPUMS-EurAsia and IECM initiatives, 2010-2014

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

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
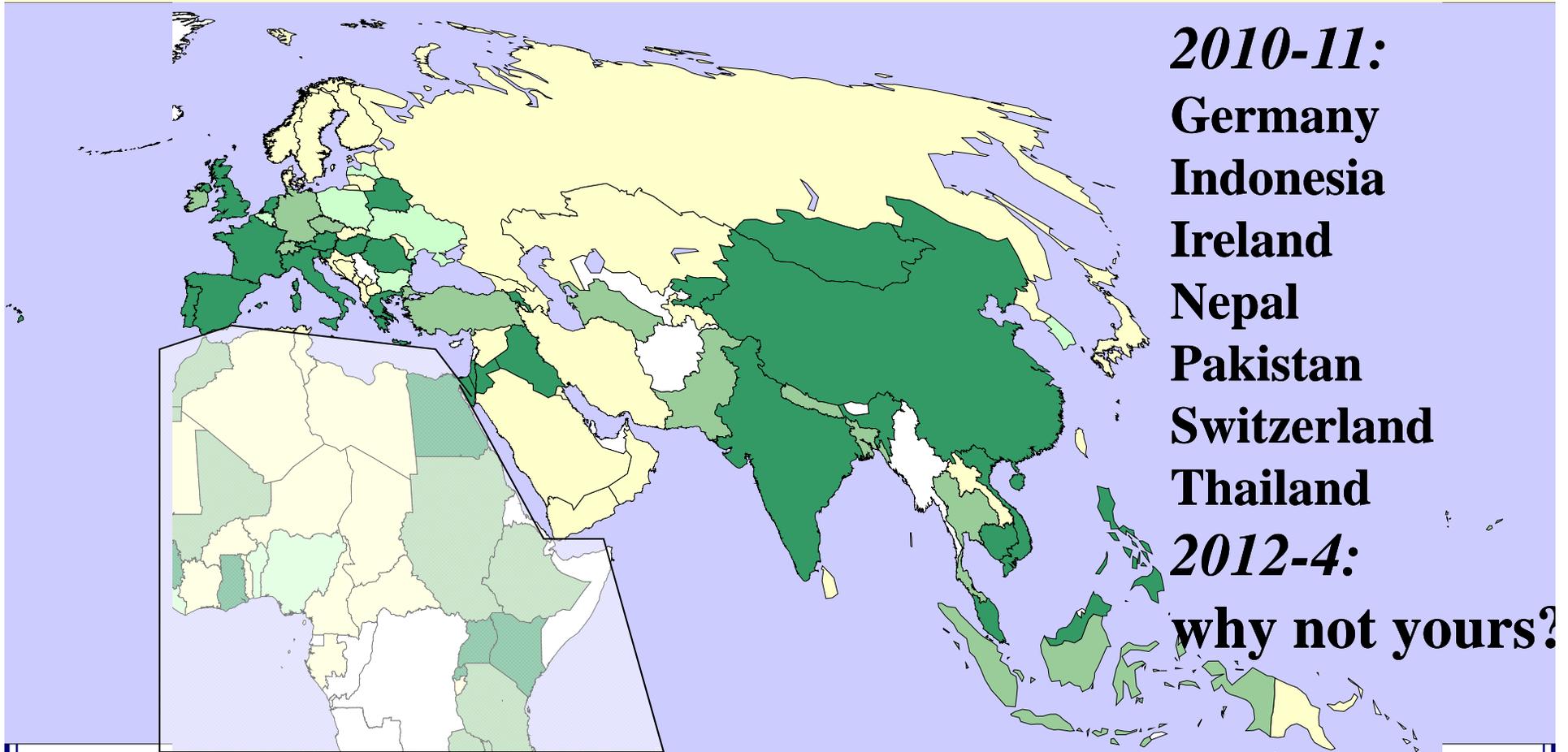
IPUMS-International:

» **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, by the addition of:**

  » **2010 round samples of 17 current partners: Austria, Belarus, Canada, France, Greece, Hungary, Israel, Italy, Kyrgyzstan, Netherlands, Portugal, Romania, Slovenia, Spain, Switzerland, UK, USA, etc.**

  » **Samples for 5 countries currently in development: Belgium, Czech Republic, Ireland, Germany, Turkey**

  » **Future partners? Albania? Bulgaria? Croatia? Estonia? Finland? Kazahkstan? Latvia? Lithuania? Poland? Russian Federation? Serbia? Slovakia? Ukraine? FYR Macedonia? Others?**

IPUMS-EurAsia

dark green = integrated and disseminating
(44 countries, 130 censuses, 279 millon person records)
green = to be integrated (35 countries, 90 censuses, 150 mill.)

2010-11:
Germany
Indonesia
Ireland
Nepal
Pakistan
Switzerland
Thailand
2012-4:
why not yours?

MPC **MINNESOTA**
POPULATION CENTER

UNIVERSITY OF MINNESOTA

The IPUMS-International team
May 14, 2009 with NSF over-sight board

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

Constructing the IPUMS-International integrated metadata and microdata system

» *IPUMS-International NEVER disseminates 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 sample
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

MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

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)**
    - **…**

5 step process of integration in the IPUMS system

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 scruplously 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.

The IECM project
Integrated European Census Microdata
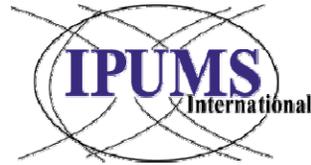
The IECM project
Integrated European Census Microdata

» **Albert: Add text here**

The IECM project--addendum
Prototype of on-line tabulator of integrated variables

- » **For the researcher to avoid having to make an extract, when all that is needed is a few counts**
- » **For any integrated variable, the tabulator produces counts by:**
  - » **Country**
  - » **Census year**
  - » **With or without filters**
  - » **Cross-tabulated by up to 3 control/explanatory variables**
  - » **With or without expansion factors**

  - » **Albert: add some screen shots or whatever you want, but keep your presentation to a max of 5 minutes—5 slides**

Census Output Needs:
1. Succinct description of census and microdata (Form "A")
2. Comprehensive metadata:
questionnaires, instructions, codebooks
3. Encrypted microdata

Ship FEDEX prepaid (email for account #) 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 Spain--censuses of 1981, 1991, 2001**

>   »   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 15-23 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

MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

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

» **2 goals:**

1. **Permanently archive source microdata against loss (copies provided exclusively to the National Statistical Agency owner)**

2. **Integrate high precision, anonymized household samples into database**

» **We prefer 100% microdata, particularly from developing countries where microdata are at risk of loss**

  » Note: some European statistical offices can no longer locate census microdata for 1960s, 1970s, 1980s and even 1990s!

  » Or even where they can locate it, are unable to make the data useable

» **4 modalities for entrusting microdata:**

1. **100% microdata to MPC:**   **38 countries**
2. **Samples provided by National Statistical Office:**   **25**
3. **Multi-use samples also entrusted to MPC:**   **12**
4. **Samples constructed by Research Institute upon request of NSO:**   **6**

» **License fee: US$5,000 for dataset of 1 million plus records**

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

»   **High precision, household samples**
  - »   **10 percent: 70 of 130 samples currently available**
  - »   **5 percent: 28**
  - »   **<5 percent: 32 (8 constitute all that survives)**

»   **Systematic random samples :**
  - »   **every n^{th} private household after a random start**
  - »   **Collective dwellings: every n^{th} person**
  - »   **extremely fine geographic stratification with proportional weighting**
  - »   **NUTS-2, NUTS-3**

»   **Anonymization, performed by NSO or 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**

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Conclusion

- » **Thanks to:**
  - » **National Statistical Offices for trust and cooperation**
  - » **International organizations for support and encouragement**
  - » **Researchers for using of IPUMS integrated datasets**
- » **Invitation to:**
  - » **National Statistical Office partners to entrust 2010 round microdata and metadata with Form "A"**
  - » **National Statistical Offices that are not yet cooperating to participate to integrate pre-2010 census microdata**
  - » **And…**

…to the 58th Session ISI: Dublin, Aug 21-26, 2011
http://www.isi2001.ie

ISI 2011

58th SESSION - AUG 21st - 26th

ISI 2011 DUBLIN

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 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**

UNIVERSITY OF MINNESOTA

IPUMS International

Thank you!!

rmccaa@umn.edu
aepalos@ced.uab.es
pkelly@umn.edu

www.ipums.org/international
www.iecm-project.org

MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

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

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Researchers will use microdata, if it is available
3,000 accredited researchers

» **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, study documentation; then, 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**
   » Study documentation; email questions to ipumsi help desk

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA