Scientific report EHPS-net

Workshop in Umeå May 10-11, 2012

Summary

This was a workshop within workgroup ”Extraction Software”. It’s aim is to discuss ways to make an extraction software for the IDS. The suggestions will be discussed within the large group EHPS-net.

The topics at the workshop focused on:

- The making of IDS, and the many ways of creating the database using the IDS standard,
- The system of IDS software, types of modules,
- Documentation issues, such as documenting the program, structure of programs and relation with the EHPS-portal,
- Coordination of writing programs for extraction,
- Levels of advanceness and software tools,
- Dissemination issues, such as publication and handling, how to transfer data to the website, and a discussion of who are the users,
- What we need to do before the next EHPS-net meeting in Budapest, september 2012,
- Next steps – priorities and who will do what.

Description

The discussions at the workshop were lengthy and led to many concrete suggestions to reach a common standpoint in how to handle extraction software as well as important documentation issues:

1. Discussion of rules by which DDB is transferring their data to IDS

In DDB has recorded how individuals are arranged on pages in registers, but the registers do not show who lives in households

- Servants at bottom of page may belong to more than one household
- DDB reconstructs "residential units"
    - Be careful not to call "household" bc don't know that they are in separate hhh
    - e.g. woman who marries re-written to new line on same page with new husband
- -> instead of trying to impute individuals to households and roles in households, describe "conjugal family units" without describing them as households
- servants would be listed as individuals, bc they cannot be assigned to households

2. Create a template database showing all columns in all tables plus METADATA with standard columns

3. Add "TITLE" "TITLE_STANDARD" "TITLE_ENG"

4. Do not use "_ORIGINAL" e.g. "OCCUPATION" means "original occupation"

5. "ESTIMATED" in timestamp

- encourage databases to specify rules in METADATA to explain how dates are estimated
- e.g. "ESTIMATED RULE #1"

6. TIMESTAMP added "Assigned" for cases where a database administrator assigns a date or period

7. A db may add its own metadata row for a STANDARD element to describe their procedure for producing that element. Example, HSN will describe how they standardize family names.

8. Guidelines for IDS Program Documentation (version 1)

1. Program name
2. Author and contact information
3. Date
4. Purpose
5. Keywords
6. Version of program
7. IDS version
8. Software language (e.g. version of SQL)

9. Rationale
10. Background
    a. Where developed
    b. Data used in development
    c. Related documents
11. Requirements
    a. IDS Input
        i. Tables
        ii. Types
        iii. Date/time information
        iv. Other
    b. Other inputs
    c. Method to test for requirements
12. Outputs
    a. Output formats
    b. Tables
    c. Variables
    d. Other
    e. Codebook for outputs
13. Program description
    a. Overview
    b. Step by step description
    c. Program operation
14. Validation
    a. Test data
    b. Validation tests

9. Priorities

1. Validation of IDS
2. Data quality
3. Validation datasets for testing programs
    a. Family reconstitution
    b. Population register
4. Model basic extraction programs
    a. Fertility extraction for age-specific marital fertility rates
    b. Mortality extraction for age-specific death rates
    c. Migration
    d. Marital status
5. Basic analysis datasets to use with students
    a. Episode files for demographic analysis (fertility, mortality, migration) with common variables (marital status, occupation, location)
    b. Individual summary files: One row per individual with all events on one record
6. Imputing occupation over time

10. Differences between versions of SQL

- The SQL standard only covers very basic aspects, and vendors differ in implementation.
- Programs should be kept simple to avoid using SQL features that are specific to vendors.
- Researchers should test programs on standard datasets to be sure that they are getting the same results with their version of SQL.

11. Responsibility for software posted on EHPS website
    - EHPS cannot test software.
    - Software is posted without endorsement.
    - Quality is responsibility of author. Authors should include validation data and tests in documentation.
    - Users can post comments on the website.

12. IDS Extended
    - Need to develop a standard for IDS Extended
        - Table specification for IDS Extended interval files
        - Documentation requirements
        - Validation procedures
    - Create a working group on IDS Extended

13. Can we create data products for people who don't understand event histories?
    - Summary files of counts and rates
    - Synthetic censuses
    - Summary counts of events within lives to age 20, 30, 40, … (e.g. number of migrations, number of children, …)

Assessment of the results

An important aim of the workshop was to come to an agreement on how to move forward. Based on the discussions, we made a schedule for creating the important extraction programs, that will be the core of the extraction software. Each partner in the workgroup was given assignments and there were suggestions on what important topics need to be raised during the next two EHPS-net meetings in September and November.

14. Implementation of schedule

- ICPSR:
    - Validation programs for IDS (December 2012)
    - Validation database for family reconstitution (December 2012)
    - Fertility extraction program (December 2012)
- DDB:
    - Test validation programs on DDB data (May 2013)
    - Mortality extraction program (December 2012)
- HSN
    - Validation database for population register data (December 2012)
    - Migration extraction program (February 2013)
- To be done:
    - Marital status extraction program
    - Imputing occupation over time
    - Standard files for teaching
    - Imputing missing dates

15. Budapest

- 9 working group reports
    - combined report of 3. IDS and 4. Extraction software
        - Guidelines for documentation
        - Priorities
        - Implementation schedule
    - Pre-meeting for new projects and intro to IDS

16. Vancouver SSHA

- Workshop on putting data into IDS version 3
    - Thursday morning 1 November 2012

17. Expanding longitudinal demographic research in new areas

- Create list of historical demographers in Eastern Europe
- Propose 5-day introduction to historical demography

Annexes

Annex 1. Programme of the meeting.

Thursday May 10

8:30 – 8:45 Welcome, who is who

8.45 – 10:00 Making of IDS

- Walk-through making of IDS
- Many ways to do things, experiences Roger Lund.
- IDS clearing issues (Kees)

10:00 – 10:15 Coffee

10:15 - 12:00 System of IDS _ software

- Introduction (Boston presentation Kees, see pdf)
- Types of modules (building blocks)

12:00 – 13:00 Lunch

13:00 – 14:30 Documentation of the program, what do we need? How to document extraction programs?

- Documentation of the program itself
- Structure of programs (combination of modules)
- Relation with EHPS-portal (Kees)

14:30 – 14:45 Coffee

14:45 – 16:30 If many writes parts of the program, how do we coordinate that?

- list of priorities
- time scheme for testing sample
- standard tests of acceptance
- software clearing committee

Friday May 11

8:30 – 10:00 Extraction Software

- Levels of advanceness (programmer participate; Maria Larsson, DDB)
- Levels of Software tools, see APPENDIX A

10:00 – 10:15 Coffee

10:15 - 12:00 Dissemination issues

- Standard dataset, publication and handling
- IDS: How to transfer data to the website?
- Inventory that works interactively
- Standard dataset, publication and handling
- Who are the users (Appendix B)

12:00 – 13:00 Lunch

13:00 – 14:30	FOR BUDAPEST - How to bring IDS forward? - Meeting for new members before EHPS-net meeting - Ways to help those that don't have data: mentor/tutor/together with MOSAIC-project? Get data out of archives, microfilms etc
14:30 – 14:45	Coffee
14:45 – 16:30	NEXT STEPS / CONCLUSIONS - Who will do what, prioritize

Annex 2. Full list of speakers and participants.

Title and Name	Gender	Affiliation	E-mail address
Professor George Alter	Male	ICPSR, University of Michigan, Institute for Social Research P.O. Box 1248 Ann Arbor, MI 48106-1248, USA	altergc@umich.edu
Professor Anders Brändstrom	Male	Demographic Data Base, Umeå University, 901 87 Umeå, Sweden	anders.brandstrom@ddb.umu.se
Nanna Floor Clausen	Female	Danish Data Archives, Islandsgade 10, DK-5000 Odense C, Denmark	nc@sa.dk
Professor Kees Mandemakers	Male	IISG, Cruquiusweg 31 1019 AT, Amsterdam, The Netherlands	kma@iisg.nl
Maria Larsson	Female	Demographic Data Base, Umeå University, 901 87 Umeå, Sweden	maria.larsson@ddb.umu.se
PhD Annika Westberg	Female	Demographic Data Base, Umeå University, 901 87 Umeå, Sweden	annika.westberg@ddb.umu.se