Scientific report of the EHPS-net workshop of Working Group 4 – Extraction Software for IDS, 16-17 June 2014, International Institute of Social History, Amsterdam, the Netherlands

Summary

The aim of the workshop, organized by Workgroup 4 – Extraction software for IDS, was to discuss the development of extraction software, extended IDS and the handling of documentation of the progresses made. The workgroup also discussed the planned summer schools that EHPS-net will organize (Cluj 2014, Lund 2014 and Umeå 2015) and a proposal for Horizon 2020.

Scientific content and discussions at the event

Extraction software:

During the previous meeting, we assigned tasks to be completed to the current meeting of the working group; occupation, migration and marital status.

Occupation (DDB)

During the previous meeting in October 2013, the working group agreed on the development of software for occupation. The Demographic Data Base presented a first version that was discussed. The handling of occupation is very complex, and we expect that users of IDS have a very diverse experience of working with occupation and for the software to be used for a wide range of researchers, we decided to create two occupational scripts:

- A complex one that looks like the sources. All avaliable options of how to handle occupations are present.
- A ”cleaned” version where choises, where applicable, have been made already. These choises could be, for example, choosing the first occupation if there are more than one within the same record.

Migration (HSN)

The functional design is done and there will be some further development before completing the software for migration.

Marital status (ISPCR)

This software is being worked on at the moment. It will be presented later this year.

Transposing program:

George Alter has previously presented a program that will help databases transpose their own database into the IDS format. Any number of tables in original form can be put into the program and

help convert the database into IDS. It has been tested with some good results and it will be further developed during this year.

We agreed that DDB will start by testing the program on a new parish, that is currently not in an IDS format. After the validation of the program is made, there should be a good documentation written and put on the collaboratory for the EHPS-net community to use.

Checking the IDS of HSN

At the previous meeting, HSN agreed to continue the work with converting the HSN database an IDS database. It has not been fully completed yet, but there is a plan to introduce parts of the HSN database into CLARIAH, wich uses integration on two IDS databases: HSN and LINKS.

Extended IDS:

Luciana Quaranta discussed a series of concepts and programs that can be used to select data from the IDS to construct a file that is ready for statistical analysis.

- **Extended IDS table maker**: creates empty tables for the Extended Intermediate Data Structure, which can be used to store constructed extended variables.
- **Household size**: is an example of a program aimed at creating contextual level extended variables.
- **Select type**: produces Excel tables containing a list of all unique Types that are stored in the INDIVIDUAL, INDIVIDUAL_EXT, CONTEXT or CONTEXT_EXT tables, in order to allow the user to select any of the stored Types for their analysis.
- **Append individual variables**: appends the selected individual variables to the Event chronicle and Variable setup files. The Event chronicle file contains the extracted data, while the Variable setup file contains information on these extracted variables.
- **Append contextual variables**: appends the selected contextual variables to the Event chronicle and Variable setup files.
- **Episodes file creator**: converts data extractions into a rectangular episodes table that is ready for statistical analysis.

She also discussed two articles written on these concepts and programs, which she will send for publication in the fall.

Documentation:

Documentation of databases, programs and scripts are of utmost importance for EHPS-net. The workgroup discussed in what ways we best can manage documentation these. The project has budgeted for the developing of documentation, of around 25 % of a full time. The working group wrote the ad for a position as a documentation specialist that will be sent out as soon as possible. Our ambition is to have someone hired in September 2014.

During the WG4-workshop in Umeå in May 2012, we decided on a template, version 1, for documentation. These were checked and a few changes were made. The document will be published on the Collaboratory.

The version 2 of the Guidelines for IDS Program Documentation is:

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, such as programs, data or intermediate results
    c. Instructions for users to check a database to determine whether the data meets the requirements of the programs
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

In order to disseminate the documentation of databases, we discussed the possibility to write documentation in the format of articles that might be submitted to the EHPS-net journal; Historical Life Course Studies. The WG for documentation has made a proposal of a template of the documentation of databases that might be used.

Summer schools

During 2014 and 2015, EHPS-net will organize a couple of summer schools with the aim to educate students and researchers in methods used in longitudinal studies. They will be organized so that they will build on the previous course and give progression in the student's knowledge. George Alter presented some experiences from the Ann Arbour summer courses, and what the organizers of future summer courses might consider:

- Expectations on the project reports could be clearly stated.

- Emphasis should be on running statistics rather than data management.
- Limiting choices of data – the students should only work with data that are given at the course.
- Give the students variables and possibly a set of questions and work in teams of 2-3 in the projects, *OR* assign them 1 question and 5 datasets.
- A poster presentation is better than a report because it needs less text, which makes people focus on the data.

- **Summer school in Cluj 2014**

  Kees Mandemakers presented the plan for the introductory summer school in Cluj 2014. Last year, there were problems with the students not having enough language skills. In order to control this, Kees interviewed all applicants over Skype before accepting them. Some of the contents are: Introduction to historical demography, working with Access and Stata, life course transitions, life- and mortality tables, household classification systems, social stratification and an introduction to IDS and building datasets in IDS.

- **Summer school in Lund 2014**

  Luciana Quaranta presented the plan for the summer school in Lund August 2014. This will be an introduction in historical demography using register-type data. The summer course will discuss sources, methods and theories of historical demography, focusing on, for example: Vital registration, life tables, mortality, marriage and family, social stratification and social mobility, long-term effects of early life conditions, impact of short-term economic stress.

- **Summer course in Umeå 2015**

  Annika Westberg presented the first draft of the summer course that will be held in Umeå in June 2015. At the moment, the draft contains a "wish list" for what the course contents could be. Since this course aims at more advanced methods, but still builds on the previous courses, we discussed what level and methods will be reasonable as well as what expertise the students should possess in order for it to be on a good level.

Project proposal to Horizon 2020

This topic was added to the meeting programme because we have an opportunity to have Saskia Hin working on a project proposal for 2 months. KU Leuven will be the proposal coordinator. At a meeting with EU-project administrators at Brussels, possibilities to apply for a Horizon 2020 call. They advised us to focus on the INFRADEV and the call "Design" (http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/60-infradev-1-2014.html). It recommends applying for 1-3 million euro, with deadline on Sept 2nd. KU Leuven will be the proposal coordinator. During the next couple of months, we will keep discussing what kind of infrastructure we want to build and who the key partners will be.

Assessment of the results and impact of the event on the future

Extraction software
The group discussed if there, at the moment, are suggestions for new extraction softwares. As we already have developed extraction software for fertility, mortality and occupational structure, and extraction software for migration and marital structure are under way, we decided to consolidate the things that have been done so far and finish programs that are being worked with.

EHPS-net round table at SSHA in Toronto, november 2014

EHPS-net will organize a round table at the SSHA conference this year, where we plan to show what the IDS can do. So far, we agreed that George Alter will demonstrate a few US family reconstitution datasets and Annika Westberg will demonstrate some research on social mobility using the occupational extraction software. Kees Mandemakers will introduce the IDS structure and give an overview of where EHPS-net is currently and what we are planning for the next years to come.

Changes in IDS version 4

There has been some important changes in the IDS version 4 compared with previous versions. It was suggested that the programmers should have a meeting and discuss these matters. It was also suggested that the meeting could be organized as a webinar – a meeting over internet:

- October 2014:
  - Organize a webinar with the main topic being changes and mapping in IDS 4 and the possibility to use the transposing program.
  - Participants: HSN, ICPSR, DDB, SEDD and Leuven.

- January, 2015
  - Webinar with an extended initiative, where we should invite databases broadly to participate.

Next meeting of WG4

The next meeting of the workgroup 4 was decided to take place in late April 2015. On the agenda is:

- Result of webinar,
- Put extraction software on the Collaboratory,
- Documentation of extraction software and databases,
- Inventory of new extraction software,
- Etc.

Programme of the meeting

Monday 16 June 2014

9:00 Coffee

9:15 Update on what has happened since the last workshop. The following schedule was decided upon at the previous meeting:

ICPSR:

- Documentation for the validation program (could be done by someone else)
- Documentation for the fertility software
- Extraction software for marital status (May 2014)

DDB:

- Check if DDB has tests for logical relations (birth dates before death dates, etc.) that can be used within IDS.
- Documentation on extraction software mortality (could be done by someone else)
- Extraction software Occupation (May 2014)

HSN:

- Check the IDS of HSN
- Extraction software Migration (no deadline)

SEDD:

- Extended IDS on the collab (January 2014)
- Extended IDS (May 2014)

EHPS-net:

- There is funding for extraction software and documentation. George suggests that we should spend it on documentation rather than programming. Writing and checking the documentation, version 1, for IDS as well as documentation for validation- and extraction software.
- Pass the question on the need for standard files for education to workgroup 7 to discuss it further.
- Workgroup 1 will take care of posting the links on the website on where to find the datasets for helping people create IDS.

12:00 Lunch

13:00 Extended IDS

14:30 Tea

15:00 Documentation, continued discussion

16:00 Wrapping up today's discussions

18:00 Dinner

Tuesday 17 June 2014

9:30 Coffee

9:45 New extraction software for IDS?

12:00 Lunch

13:00 Defining standard sets for teaching purposes (input from WG7?)

14:30 Tea

15:00 Meeting of developers, example files/test files on the website of EHPS-net

16:00 Next steps and assignments

16:30 Other matters

16:45 Wrapping up the workshop and next meeting of WG4

19:00 Dinner

Full list of speakers and participants

George Alter (USA), ICPSR, University of Michigan, altergc@umich.edu
Anders Brändström (Sweden), Demographic Data Base, Umeå University, Anders.Brandstrom@ddb.umu.se
Kees Mandemakers (the Netherlands), International Institute of Social History, Historical Sample of the Netherlands, kma@iisg.nl
Luciana Quaranta (Sweden), Centre for Economic Demography, Lund, luciana.quranta@ekh.lu.se
Annika Westberg (Sweden), Demographic Data Base, Umeå University, Annika.Westberg@ddb.umu.se