



Scientific report of the EHPS-Net workshop of Working Group 4 – Extraction Software for IDS, 14-15 June 2015, Centre for Economic Demography (CED), Lund, Sweden

Summary

The aim of this workshop, organised by Workgroup 4 – Extraction software for IDS, was to discuss the further development of extraction software and extended IDS. Emphases was put on making EHPS-Net more visible by giving priority to IDS-webinars, transposing extraction software output into R - because it is software free for all users - and plan a number of workshops for 2015/2016 for end users of IDS.

Scientific content, discussions at the event and actions to be taken

1. The Chronicle File is adopted as the common format for extraction software. It will be an 'in between' between the IDS and the episode files for analyses. The chronicle files are in this sense the building blocks of the dataset for analysis. The chronological files will have the same basic structure as the IDS.
2. George Alter and Luciana Quaranta will finish the chronological file on *Fertility* and *Marital status*.
3. Going from chronicle files to an episode file or other datasets for analysis will be done by way of developing an R software package. Anders Brändström will contact Göran Broström in Umeå for writing such a package.
4. The Transposer made by George and Emily Merchand was demonstrated. The software maps datasets into IDS by way of an intermediate file in which the mapping is defined (which variable/value is going to which in the IDS). Kees Mandemakers will put the Transposer on the Collaboratory, send an email to invite comments esp. on the building of the mapping files.
5. Extended IDS articles of Luciana.
Luciana will send Kees the articles to be put on the Collaboratory, and ask for comments before the 1st of August. Luciana will then improve the articles before the first of September and send them to be published in the web journal. The programs itself will be put in the Repository.
6. The Validation program on IDS should be put in the Repository. Kees will ask all persons in the EHPS-Net to run it and send the results to the clearing committee. Kees will also announce that Luciana is a new member in the Clearing committee, since Annika Westberg had to withdraw.

7. Webinar on IDS for population registers will be scheduled for spring 2016 and organized by George and Kees.
8. What is really needed for family reconstitution is a workshop or a course. Georg Fertig is interested in doing such a course. EHPS could find a suitable location for such an event.
Action: Proposal to the educational workgroup, with high priority.
9. The suggestion from Kees on the issue of time stamps was discussed and it was decided not to include a new table in the IDS. However, how to handle the issue should be further discussed in the Collaboratory. Kees will set up such a forum.
10. The proposal of Richard Zijdeman for HISCAM variables as part of the official IDS structure was accepted.
11. Money budgeted for programming should be focused on the transformation of extended IDS into R. Göran Broström in Umeå will be contacted by Anders.
12. Top priority for George is to convert the Fertility program to produce chronicle files. George will also think about a possible candidate to write a paper on comparative fertility, using IDS-databases. The paper should be presented in the Leuven historical demography conference in the autumn of 2016.
13. The HISCO workshop in Romania. We will not go into debate and suggest changes in the HISCO code itself. Instead things could be done at the database level of each country, to include regional context. We will have continuing problems with the suggested hierarchy of HISCO in three classes, but we cannot change the hierarchy, just point at the problem.
14. Webinar; one for the Transposer program in particular for people from Eastern Europe on a fairly basic level. This could be done in the Fall 2015 (late October). Presentation will be about 30 minutes and then Questions/Comments. By way of a) chatting (presenter needs an assistant to keep an eye on all traffic), and b) talking after allowance by the chair. People can use a button to make clear that they want to comment; microphones are opened then only by way of the presenter. The whole session is automatically recorded and can be put on YouTube; there is an ICPSR channel on YouTube with their webinars. Installation is easy by sending out a link with an id for each session and the persons bring the session up in the browser (they have to download an extension to the browser and to type in the meeting number). GoToMeeting is in use at ICPSR. When you install it; you can check the performance of the audio and video beforehand.
15. An announcement will be put on the Collaboratory; George recommends to put together a list of all participants of summer schools, list of historical demography society; EHPS-Net itself. List of participants in Alghero. Kees will ask Marja to make such a list of people which can be invited to the webinar.
16. A workshop in Lund on experiences using IDS, as to find common solutions. The target group are databases that have implemented IDS, their programmers and decision makers. It will not be an introductory workshop, but an advanced one. Lund was decided as a convenient place to hold the workshop. Workshop on the mapping issue: 19th and 20th of September. Arriving at the 18th Friday; 1,5 day workshop ending at lunch on Sunday. Attendants: Persons who understand the data very well. Sweden: three ones; Norway: Hilde Leikny Sommerseth, Trygve Andersen.; Leuven: Wim De Mulder. Eastern Europe: Romania: Vlad Popovici; SLS; Cambridge; already using IDS. Luciana will be the contact in Lund. The workshop will be max-

imized to about 20 persons. 17th of August deadline to attend. Costs for the workshop – travel and accommodation – will be covered by the participants themselves. EHPS-Net will cover costs for George, lunches and a dinner for all participants. Marja and Kees will look into the financials.

17. We have to put a thread on Education datasets on the Collaboratory. The Lund data used in their courses are already available on their website (must be linked); exercises for the ICPSR are on the web as well. The same goes with the datasets used in the Umeå summer school. Links should be sent to Marja and Kees.

18. Next meeting of WG 4 on IDS: 27th and 28th of May 2016 in Amsterdam.

Annexes

Schedule

Saturday 13 June 2015

19:00 **Dinner**

Sunday 14 June 2015

9:00 **Tea and coffee**

9:15 **Update on what has happened since the last workshop**

The following schedule was decided/discussed upon at the previous meeting:

Occupation (DDB)

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 available options of how to handle occupations are present.
- A "cleaned" version where choices, where applicable, have been made already. These choices could be, for example, choosing the first occupation if there are more than one within the same record.
- HISCO and contextualization (see documents from Gunnar Thorvaldsen and NAPP)
- HISCO workshop in Cluj, Spring 2016. Input from WG4.

Migration (HSN)

- Still not much done, besides the functional design. There is a need for further development before completing the software for migration.

Marital status (ISPCR)

- Report

IDS 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.

- Demo of the transposing program
- Running it on a synthetic database
- Official version of IDS
- Documentation

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, which uses integration on two IDS databases: HSN and LINKS.

- Report on the building of the IDS structure for HSN, LINKS and integration of LINKS into HSN.

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.

Documentation

12:00 **Lunch**

13:30 **Further development of IDS and extraction software**

Clearing Committee

- New members required
- "Dialects" of IDS
- Scripts for IDS
- User friendliness
- IDS and large data files (PPT by Kees)
- IDS Family reconstitution program

Extraction software

- How to increase productivity: Hiring a programmer
- Structures: Long files (episode files) and/or Wide files
- New ideas for software: List of priorities
- Comparative research on fertility using IDS databases
- Other strategies for WG4

15:30 **Tea and Coffee**

EHPs-Net Closing meeting in Leuven 2016

Wrapping up today's discussion

Drinks

19:00 **Dinner**

Full list of participants

George Alter (USA), ICPSR, University of Michigan, altergc@umich.edu

Anders Brändström (Sweden), Demographic Data Base, Umeå University,
Anders.Brandstrom@ddb.umu.se

Wim De Mulder (Belgium), FaPOS, KU Leuven, wim.demulder@cs.kuleuven.be

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

Hilde Leikny Sommerseth (Norway), Department of History and Religious Studies, The Arctic
University of Norway, hilde.sommerseth@uit.no