Scientific report of the EHPS-net workshop of Working Group 8 – Standards for documentation about databases, 19-20 May 2014, Copenhagen, Denmark

Summary

The goal for this workshop was to come as near as possible to have an agreement on a common documentation for databases with historical, quantitative data. The starting point was a summary of the present situation with the conclusion that the documentation of the databases should have first priority, since IDS itself is documented by way of the metadata system and the extraction software has a documentation recommendation in a first version.

Existing documentation standards were presented with special focus on the Standard Study Description. It originated in the beginning of the 1980'es but the elements from it are still being used in the state of the art documentation standards like DDI 3.2. The documentation of variables and the inclusion of the sources and their history in the way that IPUMS does were also taken into consideration. Finally a questionnaire on existing databases will be integrated.

The workshop resulted in bringing together the elements from the Standard Study Description, some parts from IPUMS and from the questionnaire into a proposal that can be used for documentation of the databases and their sources. The proposal will be presented at the next GENERAL ASSEMBLY.

Scientific content and discussions at the event

The purpose of the workshop was the need to define a common standard for documenting databases with transcribed historical data.

Item 1: Present situation regarding documentation

The debate on the objects to be documented included from the beginning the sources, the databases, the IDS datasets for analysis. IDS itself is documented and the documentation is already presented on the website. The definition for documenting the extraction software has been finished in a version 1 which can be found at the website: http://www.ehps-net.eu/system/files/forum/report_wg4.pdf.

The starting point for the initial discussion was the questionnaire form and the result from the survey carried out with this form. The questionnaire and survey can be found at http://historicaldemography.net/questionnaires.php. The data from the survey has been added and expanded regarding general database introduction on the EHPS-Net website: http://www.ehps-net.eu/databases.

Item 2: Discussion and agreement on the minimum demands for documentation

The purpose of this discussion was to decide on what to document before we began looking at present documentation standards. Should the sources used in the databases be documented independently of the databases but according to the same rules or should the sources be documented as a part of the documentation of the databases? The discussion resulted in a list of requirements and request which were incorporated into the final proposal.

Item 3: Presentation of the Standard Study Description

In 1981 the community for social science developed a standard for documenting surveys in the data archives. Even though this standard was originally made for punch cards the content and items in the standard study description are still used – as can be seen in the latest version from the Data Documentation Initiative.
All the items in the standard were listed and explained. The debate on the items showed that the standard study description was usable for documenting the databases in this project. The complete list will not be given here as it is quite long and many items are included in the proposal from the workshop.

Item 4: Presentation and discussion of DDI-C and DDI-L

The Data Documentation Initiative was presented along with its origin and purpose. The purpose is and was to develop a documentation standard using xml. DDI-C was the first version and based on converting the old punch-card format OSIRIS to an xml-version. The elements from the standard study description were and are included in the DDI-versions as well as the descriptions and definitions of the questions and variables in the codebook. The DDI-C version used to be an xml using a dtd (data type definition) whereas DDI-L is based on schema and documenting a dataset all through its lifecycle. The issues to be decided regarding using xml were manifold. The first issue is whether the documentation of the databases is for the sake of preservation or presentation. The conclusion was that it is for presentation of the databases for secondary analysis users on the website. Another issue was the lack of tool for entering the information and for extracting them to the DDI-C. Therefore it was concluded that the investment for using DDI-xml is too high.

Item 5: Presentation and discussion of IPUMS' documentation of variables

The Minnesota Population Center is the host of a large number of population databases combined in IPUMS and IPUMSi. The project was presented briefly with focus on the way data variables are documented. Every variable in the IPUMS data has an extensive comparability discussion in the documentation that points out consistencies and inconsistencies across all samples in the database. IPUMS metadata also provide documentation of the origin of the sources and their history. Variables are documented with this structure:

- Description
- Comparability
- Universe
- Codes
- Avalability
- Questionare Text
- Flags

Item 6: Decision on documentation standard for historical databases and sources

The result of the presentations and discussions of the workshop was a proposal which includes all relevant items from the standard study description and documentation on sources inspired from previous debates and will include a detailed list of how to state which background variables are included in the database.
The EHPS-Net website will hold the information from the survey carried out using the new documentation proposal. The information will be updated on a regular three-year basis. The most updated information on a database will be on its own website.

Proposal for documentation standard for historical databases and their sources:

Identification

- **Keywords** :
  - Recommended if appropriate: demography, life course, census, church register, civil certificates, population register, history, social science, genetics, migration, occupations
  - Other: if you have data not covered by the recommended

- **Title**:
  - Is mandatory
  - Add a subtitle which brings meaning to the title (scope, place, time)
- **Citation**:
  - Make the citation how you want others to cite your database
- **Institute**:
  - Name of institute or organisation
- **Primary responsible person**
  - Name of institute and/ or researchers
- **Address for database/ source**:
  - Website, email to contact person, post address, phone
- **Main economic funding**
  - Name of organisations who made the grants/who sustains it
- **Abstract**
  - Content of database – (short) description
  - Scope and main objective
  - Sample strategy
  - Main sources
  - Max length: 300 words

Sources

- **Source type**
  - What kind of source is used for the database (register, census, certificates, ...)
  - The purpose of the sources: why was it create(laws, administrative instruments, specific rules of record keeping, information processing strategies of producers, unsystematic characteristics of source origins, ...)
- **Scope**:
  - For whom was the source applicable
- **Content:**
  - What was to be recorded
- **Time dimension**:
  - when was the information in the sources to be recorded
- **Characteristics of storage of sources noted**
  - (completely preserved, partially destroyed by personnel according to systematic criteria, reorganization by producer of the source, reorganization by record linkage procedures, partially destroyed or damaged for other reasons)
- **Documentation and access to sources**
  - (completely documented and accessible by:;partially documented and accessible by:; no documentation but accessible by:)

Time and universe

- **Period covered**
  - give first and last year or date if possible
- **Geographical universe**
  - (local, regional, national, cross-national)
- **Units**
  - (individuals, households, families, institutions)

Sampling design and procedures

- **Selected sources**

- **Sampling units**
    - households, individuals, regions
- **Variables used for selection**
    - (age, gender, ...)
- **Selection method**
    - random, stratified random, total count, clustered, other

The database

- **Number of units**
- **Variables per unit**
- **Completeness**
    - Are all variables from the sources included in the database
- **Current data representation**
    - Database Software (MySql, MsSql, Access, Excel, ...)
- **Language of written material**
    - documentation and original sources
- **Control methods by researcher**
    - Internal consistencies,
- **Access conditions**
    - How do a user get access to the database. What conditions. Limits to delivery of the data.

Data collection

- **Data collection period**:
    - When the data was collected and transcribed
- **Data collection method**
    - Public digital register, Transcription.
    - If transcription: How was the transcription done
        - By individuals, from scanned sources, from LDS's microfilms, proof reading, automatic controls
    - How was the checking of the transcription done
    - When was it done
    - Purpose – LDS, research, genealogy
- **Data collection staff**
- **Publications and reports**
    - About the database itself (max. 5)
    - Main publications on research based on the database by primary researcher (max. 5)

Linkage process

- **Linkage**
    - Which sources and units of observation have been linked: (e.g. birth and marriages, marriage and death)
- **Documentation of linking**
    - programme, manually, ...
    - Name of software if used (and parameters)
- **Rules for linking**
    - Flags definition (list them: age, name, extra knowledge, ...)
- **How is each reconstructed person traceable to the original sources / transcribed data**
- **Linkage percentage**
- **Quality of linkage** (own evaluation)

Variables included in database

- **On individuals**
  - Age, sex, marital status, list further to be made …
- **Household characteristica**
  - Type of household, children present, age and number of children etc.

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

The aim of the workshop was to find a common standard for documentation of the databases that are participating in the EHPS-Net. An important outcome of the workshop was the decision to use the Standard Study Description which has been developed in 1981 in order to document social science surveys. A thorough walk-through of the standard study description resulted in some updates, changes and additions to make it applicable for documenting databases with historical, quantitative data. The result will be presented at the General Assembly in September.
At the workshop the following plan for carrying the decisions further was agreed upon:

1. The questionnaire used for collecting the existing information on the databases will be reviewed in order to identify redundant questions (Paping)
2. The revised and updated standard study description will be given appropriate headlines for each group (Clausen)
3. From IPUMS the most important variables and their descriptions will be selected to see if and how they can be part of the documentation (Erikstad)
4. Integration of the questionnaire and the study description with the data from the EHPS-Net website (Mandemakers)
5. Deadline for the above tasks is July 1, 2014
6. At the General Assembly in September 2014 the proposal for the documentation standard will be presented
7. 2015: Prepare a special issue of the journal where the databases are presented. It is the intention that this issue will be received as an updated version of the book: 'Handbook of International historical microdata for population research'. Ed. By Patt Hall, Robert McCaa and Gunnar Thorvaldsen from 2000.
8. The participants will be asked to add information about their databases according to the accepted documentation standard. The EHPS-Net website will be updated with the new information.

Programme of the meeting

Monday 19 May 2014

12:00 Lunch
13:00 Opening by Nanna Floor Clausen
13:15 Present situation regarding documentation (Kees Mandemakers)
13:45 Discussion on the minimum demands for documentation
15:15 Break
15:30 Agreement on recommended documentation requirements
16:00 Presentation of Standard Study Description (Nanna Floor Clausen)
17:30 Summary of meeting day 1

Tuesday, 20 May 2014

09:00 Presentation and discussion of DDI 2.0 (Nanna Floor Clausen)
10:50 Short break
11:00 Presentation and discussion of IPUMS' documentation of variables (Marianne Erikstad)
12:30 Lunch
13:30 Test the presented standards on databases from the participants

14:00 First version of templates for database documentation
16:00 Summary and conclusion

Full list of speakers and participants

Nanna Floor Clausen (female; Denmark), Danish National Archives, Islandsgade 10, DK-5000 Odense, nc@sa.dk

Prof. Kees Mandemakers (male; the Netherlands), International Institute of Social History, Historical Sample of the Netherlands, Cruquiusweg 31, 1019 AT Amsterdam, kma@iisg.nl

Prof. dr. Gunnar Thorvaldsen (male; Norway), The Norwegian Historical Data Centre, Universitetet i Tromsø, N-9037 Tromsø, gunnar.thorvaldsen@uit.no

Marianne Erikstad (female; Norway), The Norwegian Historical Data Centre, Universitetet i Tromsø, N-9037 Tromsø, marianne.erikstad@uit.no

Dr. Richard Paping (male, the Netherlands), Department of Economic and Social History, University of Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, r.f.j.paping@rug.nl