

***Abstract submitted to NCESS Conference on e-Social Science, Cologne, Germany
24-26 June 2009: <http://www.ncess.ac.uk/conference-09/>***

Performance Evaluation of the Structured Population Event History Register
(SPEHR) Relational Schema for Managing Longitudinal Health and Population Data

Benjamin D. Clark
Samuel J. Clark

Words: 1,473 excluding title, authors and references

Introduction:

The Structured Population Event History Register (SPEHR) (Clark, 2006) is a relational schema designed to standardize the structure and definition of longitudinal health and demographic information collected by long-term demographic surveillance system (DSS) sites. These sites exist primarily in the developing world and host a wide array of health and social science investigations (INDEPTH, 2009). A key challenge facing all of these sites is efficient, accurate management of their data from collection through to publication. This is difficult because the data are inherently temporal, often complex and usually subject to continuously changing structure, content and definition as the sites grow and the questions they address change through time.

SPEHR is based on an abstract, general model of time and uses a strongly metadata-centric design philosophy similar to EAV to be maximally flexible in terms of what can be stored. The core schema consists of entities corresponding to events and states and influences that link events to states in a many-to-many sense. Flexible lists of attributes are linked to both events and states, and a variety of other entities refine the definition and storage of relationships between specific types of states and the temporal integrity of events vis-a-vis states and each other. Because of its metadata-driven, EAV-like structure the SPEHR schema is highly normalized and invariant; i.e. it is not necessary to change the schema to add, modify or delete the ability to store information describing new or modified real-world entities. Because the metadata are necessary to define the database, it is not possible to create a SPEHR-based database without defining the metadata, and this creates a situation where the primary data are always documented. Finally, the metadata themselves become useful in allowing automated, programmed access to the primary data and the standardization of many routine tasks.

In this paper we present the first real test of the SPEHR schema using 'real' DSS data of non-trivial quantity, duration and complexity. The main aim is to demonstrate

that SPEHR works and fulfills its basic design criteria and does not demonstrate any serious problems.

The specific questions we address are:

1. Will SPEHR accommodate the full range of longitudinal data in a real DSS of non-trivial size, duration and complexity?
2. Are data equivalent in the non-SPEHR and SPEHR schemas?
3. Is it possible to conduct the same analysis on the two schemas and produce identical results?
4. Is performance of the SPEHR-based database acceptable?

Methods and Results:

1. Create anonymized 1:10 sample of Agincourt HDSS relational database that maintains relational structure.

The data to conduct this study come from the Agincourt Health and Demographic Surveillance System in South Africa (Kahn et al, 2007). This HDSS has monitored roughly 70,000 people for the past 17 years, describing all vital events, linking relatives to one another and all individuals to places of residence and storing a large quantity of additional information describing specific conditions of individuals and households -- all through time. The full Agincourt data set is larger than needed for this study and also contains sensitive information, so the first step in our investigation is to create an anonymized 10% sample of the Agincourt HDSS database that retains its relational structure. Methods created to accomplish this are general in the sense that the relational structure of the Agincourt HDSS data conforms to the Reference Data Model (RDM) (Benzler et al, 1998) that also underpins many other DSS databases.

The sampling and anonymization are accomplished by creating an application that utilizes the Agincourt HDSS database system catalog to automate the creation of a script that accomplishes the sampling and anonymization while maintaining referential integrity. The sample is effectively a spatial sample of the DSS population drawn by randomly selecting 10% of the locations where people live in the study area and then searching the rest of the database for all data linked to those locations, i.e. households, people, residencies, etc. Data are anonymized by excluding attributes that can easily identify individuals and creating a new system of ID numbers and keys that maintain uniqueness but do not contain the real identifying numbers or codes. The user of this sampling program can specify exactly what datum are included in the sampled database and which ID numbers and codes are anonymized.

The final script generated by this application samples the database, anonymizes and recodes, creates a new relational database with the same structure as the original

and then imports the sampled and anonymized data into the new database. The result is an anonymized 10% sample of the original database that retains the relational structure of the original. This 1:10 sample has all the characteristics we need to investigate the questions that interest us here without worrying about exposing sensitive information or revealing the identity of study participants.

2. Convert the anonymized RDM database to SPEHR structure.

In order to compare the validity and performance of a database utilizing the SPEHR schema compared to one using the RDM schema, it is necessary to translate the data from RDM to SPEHR. Key to this translation is understanding how entities defined by the RDM can be reconceptualized as the more abstract events, influences and states that are the core entities in the SPEHR schema. After trying several approaches, we eventually divided the conversion process into three steps.

First we define and diagram the events and states that exist in the RDM schema and organize the states into parent and child states; i.e. those that share a hierarchical relationship to one another that could be helpful in defining relational and temporal integrity. With the events and states known, we go through the RDM schema at a level of fine detail and define and diagram the influences that appear to connect what we had defined as events and states. After defining the events, influences and states, it is easy to define the domains of types that each can be, and thereby to define the core metadata in the SPEHR schema.

Next we define an intermediate database structure consisting of one table per root event and state type in the new SPEHR schema that can be used to contain each of the event and state hierarchies defined by the metadata in the new SPEHR schema. Next, all possible event-influence-state groupings are created in a separate table and the names of the RDM tables that contain the information associated with each are added. Once populated this table is used to automatically locate event, influence and state information in the RDM database and parse it out into the individual tables in the intermediate database.

The final step is to load data from the intermediate database into the SPEHR database. This is done in an automated way using information in the intermediate database to first load all events and states, then the influences that link them, and finally the attributes that are attached to individual events and states. The conversion is checked by verifying that the same number of events, states and attributes exist in both databases, and further that identical demographic analysis conducted on the two yields the same results, below.

3. Conduct routine demographic analysis on both RDM and SPEHR-based databases and verify that they are equivalent.

SQL scripts are written to extract and/or calculate time-sex-age-specific person-years, counts of deaths, in/out migrations and births by age of mother. Results from each database are compared and found to be identical in the case of counts and equivalent within rounding error in the case of calculated values.

4. Measure, characterize and compare performance of the RDM and SPEHR databases using 'time to complete query' metric.

DSS databases require reasonable performance during data entry so that database-level integrity and consistency checks (requiring more than one table) can be conducted quickly. This type of performance requires retrieving small numbers of fields from a limited number of rows. To evaluate performance of this type we construct a set of queries that retrieve varying numbers of records containing a varying number of attributes from the RDM and SPEHR databases. 'Records' here refers to tuples that describe real-world objects, i.e. require significant reconstruction from the SPEHR schema. The test queries are run once to allow the RDBMS to construct and optimize a query plan and are subsequently run 100 times with the query profiler turned on so that they can be timed. The average time of the 100 optimized runs is then plotted as a function of the number of records and attributes retrieved.

The results indicate that the RDM is generally faster but that performance deteriorates exponentially as the number of records or attributes increases. Although SPEHR is generally slower, performance appears to scale linearly. As expected from its list-like structure, SPEHR is very quick when few attributes are retrieved; what really penalizes SPEHR is retrieval of attributes, which is not surprising given that each has to be retrieved separately.

References:

Benzler, J., K. Herbst, and B. MacLeod. A Data Model for Demographic Surveillance. 1998. http://www.indepth-network.org/publications/zindpubs/DM_for_Demographic.htm.

Clark, S.J., A General Temporal Data Model and the Structured Population Event History Register. *Demographic Research*, 2006. 15(7): p. 181-252. <http://www.demographic-research.org/Volumes/Vol15/7/default.htm>

INDEPTH Network. 2009. <http://www.indepth-network.org>.

Kahn K., et al., Research into health, population and social transitions in rural South Africa: Data and methods of the Agincourt Health and Demographic Surveillance System. *Scandinavian Journal of Public Health*, 2007. 35 (Suppl. 69): p. 8-20.