

# Improving Methods for Management and Dissemination of Longitudinal Health and Population Data In the Context of Demographic Surveillance

**Samuel Clark**



**Department of Sociology, University of Washington**

Institute of Behavioral Science, University of Colorado at Boulder

Agincourt Health and Population Unit & Computational and Applied Maths, University of the Witwatersrand

# Goal

---

To greatly increase the utilization of longitudinal health and population data

# Assumptions

---

- ▶ The ***philosophical, political*** and ***policy*** issues that often limit access to data are resolved
- ▶ The organization has committed itself to ethical, equitable, responsible and timely ***access / sharing of data***
- ▶ The only remaining challenges are ***operational***

# Specific Aims

---

1. Improve ability to **access** and **share** data
2. Provide ability to easily **pool** data and manage data from multiple sites or studies in a **repository / archive**
3. Make it possible to easily design and implement **multi-site** investigations such that each site:
  - collects and manages a set of core data standardized across all sites
  - can also collect and manage a wide variety of other data without having to maintain two separate data management facilities

# What is the Situation Now ?

---

- ▶ Data from longitudinal studies generally not easy to access or use, *by either the site or those outside*
- ▶ There are no significant repositories containing varied data of this type, apart from special efforts focusing on specific scientific topics
- ▶ It is difficult, time-consuming and expensive to design and properly implement the data management for longitudinal multi-site investigations

**➔ THE DATA ARE SERIOUSLY UNDER UTILIZED**

# What are the Challenges ?

---

- ▶ Significant challenges fall into two broad categories
  1. **Institutional**
  2. **Scientific and Technological**
- ▶ Both must be addressed in significant ways to improve utilization of the data

# Institutional

# Institutional: **IS vs. IT**

---

## ▶ Distinguish **IS** and **IT**

- *Information Science* is creative and potentially scientific in its own right
  - System design, methods development, software design, etc.  
“the collection, classification, storage, retrieval, and dissemination of recorded knowledge treated both as a pure and as an applied science” (M.W. Dictionary)
- *Information Technology* is more technical
  - Hardware, network, software coding & maintenance, etc.  
“the technology involving the development, maintenance, and use of computer systems, software, and networks for the processing and distribution of data” (M.W. Dictionary)

▶ Effective data infrastructure needs both IS and IT

▶ What most sites have now is essentially IT

# Institutional: **IS & the Organization**

---

- ▶ Organizations need IS capacity
- ▶ IS needs to have a high profile in the organization
  - IS personnel should work with scientists from conceptualization and design of projects all the way through to analysis
  - IS personnel should pursue research into data-related theory and methods in and of themselves
  - IT fits largely under IS in the organization

# Institutional: **IS & the Organization**

---

- ▶ IS personnel responsible for:
  - Link between scientists and information system
  - Architecture of overall information system
  - Design of specific components of information system
  - Interfaces between information system and both field and analysts
- Information system consists of:
  - Data
  - Databases
  - Data collection instruments
  - Archives/repositories for all types of data
  - Software and tools to run all of these

# Institutional: Usual Status of IS

---

- ▶ **Common fundamental problem:** entire data infrastructure conceptualized at IT-level and left in the hands of IT personnel – IS usually missing entirely
- ▶ IS is a ***creative, generative*** pursuit that requires the same ability and creativity usually associated with scientists
- ▶ Unfortunately it is common that data personnel are conceived as and treated like technicians

# Institutional: **IS Personnel**

---

- ▶ **IS personnel need:**
  - 1. To be treated as scientists in their own right,**
  - 2. Better salaries,**
  - 3. Career track,**
  - 4. Ability to publish and disseminate their work,**
  - 5. Professional development – training, mentoring, regular meetings to share and develop their field**
  
- ▶ **Necessary to attract the creative, high functioning, ambitious people who are necessary to create the data infrastructure we want**
  
- ▶ **Necessary to produce written record of work in this area**

# Institutional: **Recommendations**

---

- ▶ IS layer needs to be added to organizations
- ▶ **Salaries and support structure for IS personnel need to be augmented significantly** (compared to current IT) in order to attract, develop and retain the type and quality of person who is necessary
- ▶ Supra-site infrastructure needs to be developed for IS personnel:
  - Training and professional development for IS in this context
  - Publication outlets
  - Professional association
  - Regular meetings
  - Mentoring, internships, exchanges

# Scientific and Technological (S & T)

# Scientific and Technological

---

1. Standards
2. Policies and Procedures
3. Software Tools

# S & T: Standards

---

- ▶ Fundamental to simultaneously addressing the three specific aims –
  - access and sharing
  - pooling and operation of repository/archive, and
  - facilitating design and implementation of multi-site investigation
- are a set of coherent **STANDARDS-BASED** technologies and methods

# S & T: Standards: **Minimum Standards**

---

## ▶ Standards for ***defining***:

- Semantics of data
- Behavior or data, how they interact with each other
- Documentation of data
- Data collection instruments
- Data quality rules and checks

## ▶ Standard ***structures*** for:

- Operational data
- Analytical data
- Data in repositories and/or archives

# S & T: Standards: **Hypothesis**

---

**IF** a data system is separated into components related to:

- 1. the meaning (semantics) of data, and**
- 2. the structure of data;**

and:

- 3. the representation and manipulation of time is made fundamental to the design;**

## THEN:

1. the ***storage structure*** of the primary data and the *metadata* that give them meaning can be easily ***standardized***
2. the data system will be able to store and manipulate the descriptions of a diverse set of real-life entities ***through time***, thus making it ***flexible, adaptable and extensible***
3. the ***metadata*** will become essential and serve collectively as a ***built-in description of the data system and data dictionary*** for the primary data

4. the *metadata* can be used to **standardize and automate many procedures** conducted on the data, including accuracy and consistency checking and *preparation of analytical datasets*
5. the rigid, invariant structure and careful definition of meaning required by such a system will **prevent many types of common data corruption and inconsistency**
6. the inherently more **abstract** nature of a system of this type can and will likely result in:
  - a fully **normalized** database structure,
  - **poor** bulk retrieval **performance**,
  - significantly **more effort setting up the database** for specific uses, resulting from the need to create high quality, consistent definitions for data and their behavior

7. The ***operation life of the data system will be extended***, resulting from enhanced adaptability and extensibility
8. ***Analytical output will increase in quality, quantity and timeliness of production***, through automation of extraction and consistent higher overall quality of data
9. It will be possible to ***easily share, pool and archive*** data stored using the resulting *compatible standards*
10. It will be possible to archive, manipulate and ***share metadata***, resulting in the ability to ***easily and quickly add new 'modules'*** to systems that adhere to the standards

# S & T: Standards: **SPEHR**

---

- ▶ The **Structured Population History Register** tests this hypothesis
- ▶ SPEHR is relational database implementation of an abstract model of temporal reality
  - appropriate for DSS
  - tests all parts of the hypothesis
  - test currently being conducted on Agincourt DSS database; initial results favorable and positive

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

# S & T: Standards: SPEHR: **Temporal Model**

---

- ▶ General, abstract framework to unify the representation of *time* and *structure* of modeled reality
- ▶ Measures of time with meaning that concern us:
  - **State**
  - **Event**
  - And others ...

## : Temporal Model: Temporal Entities – **STATES**

---

- ▶ Everything we want to consider has a valid “lifetime”
- ▶ This makes them **States**:
  - Well-defined beginning
  - Well-defined end
  - Constant, meaningful “mode or condition of being” between the beginning and the end
- ▶ Can be either tangible (physical) or intangible:
  - People
  - Marital (conjugal) unions
  - Places
- ▶ Associated with being **CONSTANT**

# : Temporal Model: Temporal Junctures – **EVENTS**

---

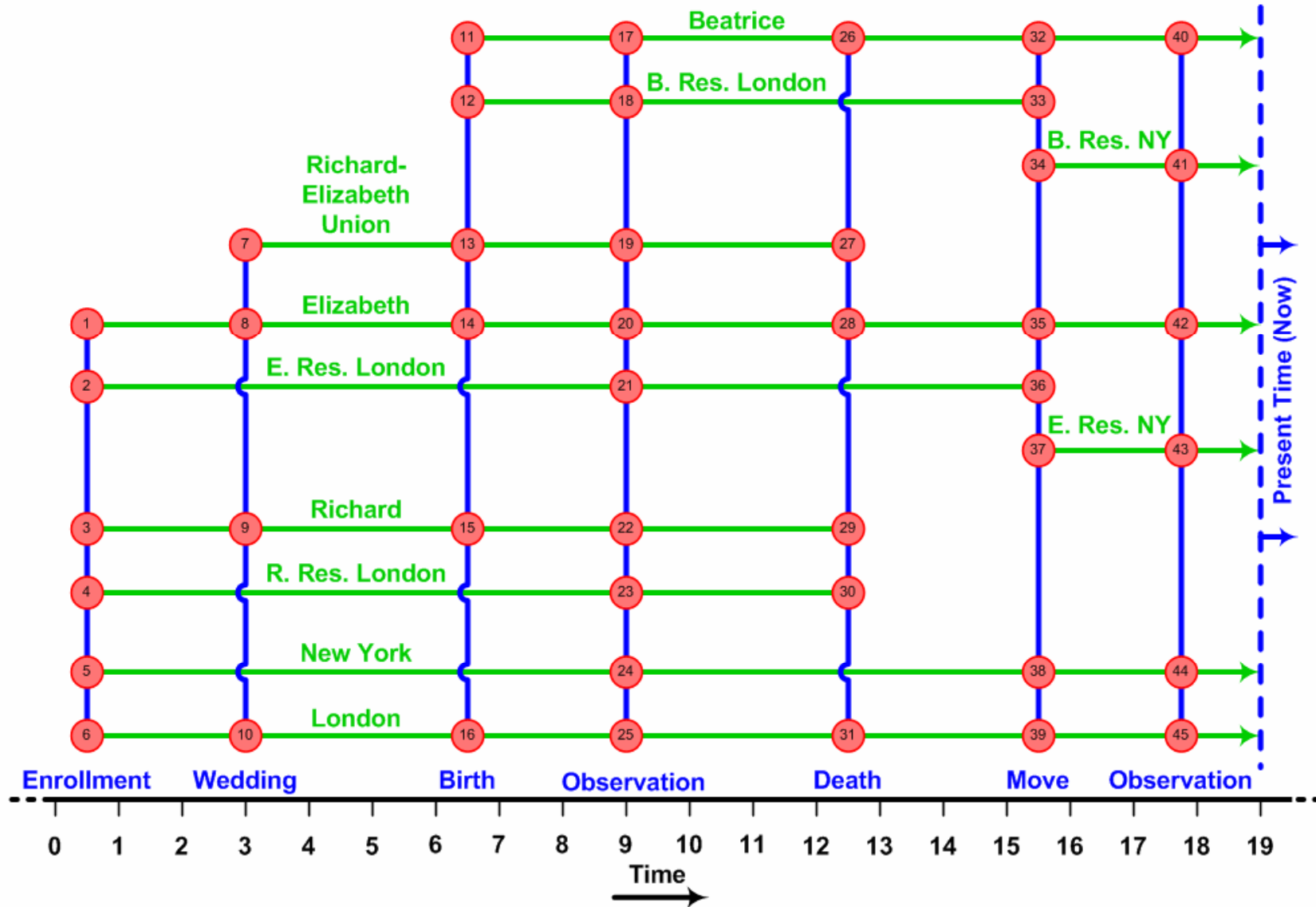
- ▶ Events bring about **CHANGE**:
  - Occur at a well-defined *point* in time
  - Bring about some change
  - The change influences one or more States that we wish to consider
- ▶ Examples:
  - Birth
  - Death
  - Wedding
  - Observation
  - Move
- ▶ Associated with **CHANGE**

## : Temporal Model: Temporal Nexus – **INFLUENCES**

---

- ▶ Influences are an explicit representation of the *link* between States and the Events that *influence* them
- ▶ Influences:
  - Represent the influence of a specific type of Event on a specific type of State
  - Are linked to exactly one Event and exactly one State
- ▶ Events can influence more than one State:
  - ⇒ Individual States and Events can be linked to many influences
  - ⇒ Through their individual links to the *same* event, the States are linked to each other
  - ⇒ ***Forms a representation of temporal relationships between States***

# : SPEHR: Temporal Model: Example



# S & T: Standards: **SPEHR**

---

- ▶ Is a relational database schema that implements the Event ↔ Influence ↔ State concept
- ▶ Can be implemented in any relational database
- ▶ Is metadata-driven
  - Does not itself model any specific reality
  - Provides the framework in which to define the reality that is to be modeled
  - Each individual SPEHR database can model a different, perhaps overlapping, reality
  - Provides a means through which to easily share and compare data describing the “overlapping” reality
- ▶ Allows easy pooling and joint management of data describing different realities

## S & T: Standards: SPEHR: **Metadata**

---

- ▶ Each entity in SPEHR is associated with a *metadata* table
- ▶ The metadata tables contain rows that describe the different *types* of entities that can be stored in a SPEHR database
  - Individual users define the metadata and hence what types of entities they want to model
  - Primary data tables reference these to specify the type of each row in a primary data table
- ▶ **Provide a means to document the “primary” data**
- ▶ Allows the schema (table structure) of the database to remain constant
- ▶ Facilitates data sharing and pooling

# S & T: Standards: SPEHR: Metadata: States

---

- ▶ A metadata table containing a list of the types of States that can be represented in a specific SPEHR-based database
- ▶ A primary State table to contain “instances” of the State types stored in the States metadata table
- ▶ A referential integrity constraint on the States table forcing every State instance to be associated with a valid State Type
- ▶ No other information in the State table:
  - No dates,
  - No attributes
  - ...

# S & T: Standards: SPEHR: **Metadata: Events**

---

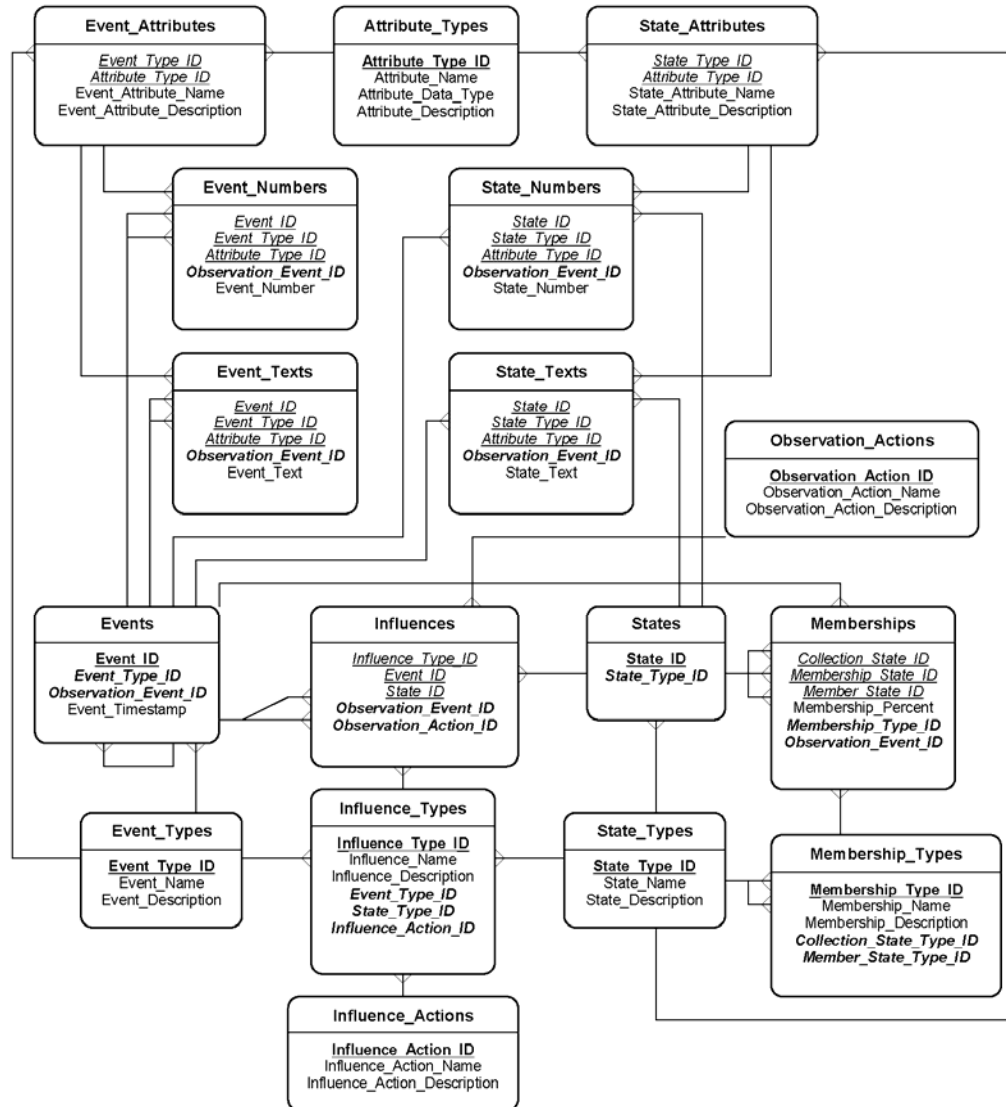
- ▶ Very similar to the States tables
- ▶ Event types metadata table
- ▶ Primary Events table to contain instances of specific Event types
- ▶ Link between the two
  
- ▶ **Critical addition of a single timestamp** attribute in the primary Events table to store the timestamp that describes *when* the Event took place

# S & T: Standards: SPEHR: **Metadata: Influences**

---

- ▶ Same idea ...
- ▶ Metadata table to store Influence types
- ▶ Primary Influences table to store instances of Influence types

# S & T: Standards: SPEHR: Schema



# S & T: Standards: SPEHR: **Normalization**

---

- ▶ *Normalization* relates to the amount of duplication of information in a database
- ▶ The SPEHR schema is highly normalized
- ▶ No information stored more than once
- ▶ Most importantly, no timestamps (dates) stored more than once in different parts of the database

# Observations

---

- ▶ Observations are special Events that update information on the entities being modeled
- ▶ Allow correct censoring during analysis
- ▶ Very important
- ▶ SPEHR has special “Observation” Event type and stores links to observation Events wherever primary data are stored

# S & T: Standards: SPEHR: **Sharing Data**

---

- ▶ ***Only metadata differ*** between two individual SPEHR-based databases
- ▶ Database schema (table structure) remains constant
- ▶ To share data:
  - Simply append data from two or more SPEHR-based databases into one SPEHR-based database
  - Data will only be appended for which valid metadata exist; i.e. for the data that are defined in all contributing databases
- ▶ **Requires globally - across all databases - unique and consistent IDs for metadata**
- ▶ **Likely to require a metadata 'BANK' to store the reference copies of metadata**

# S & T: Standards: SPEHR: Multi-site SPEHR

---

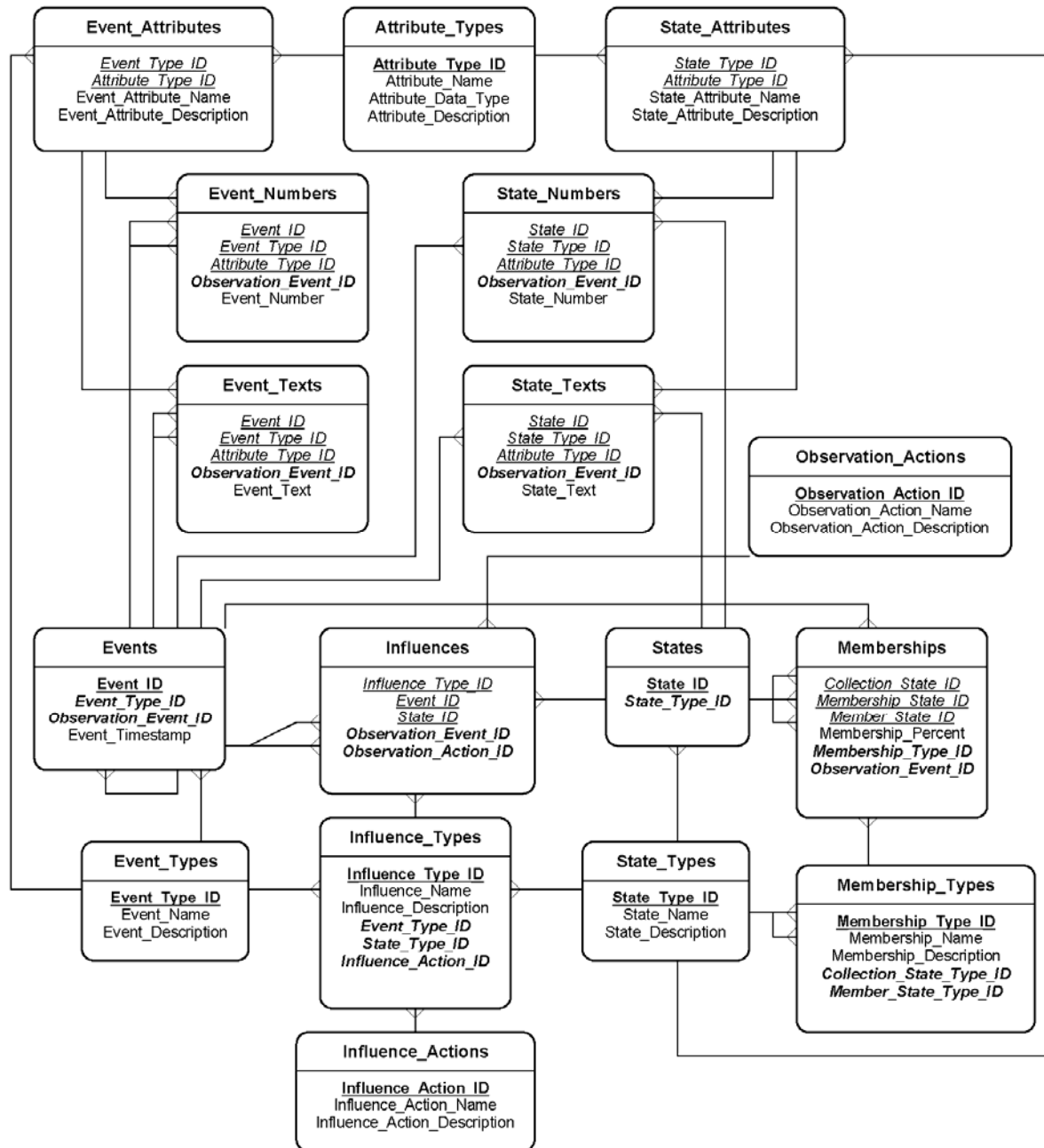
- ▶ Because the SPEHR schema doesn't change, it is possible to manage many sites' data in one SPEHR database → **repository / archive**
- ▶ Add one table to store a unique identifier for each site
- ▶ Link the sites table to all the primary data tables so that primary data is differentiated by site
- ▶ Results in multi-site SPEHR database that can simultaneously contain and manage diverse data from many sites
- ▶ **REQUIRES** metadata bank and consistent use of common metadata by all contributing sites

# S & T: Standards: SPEHR: Preliminary Results

---

- ▶ Conversion of Agincourt database (120,000 people) to SPEHR:
  - Not too difficult
  - Identified inconsistencies in the data that we would never have found otherwise
  - Forced a new level of documentation → provided lots of useful metadata
- ▶ Performance:
  - As expected, slower at bulk retrieval
  - Unexpected, much faster at retrieving specific data
  - In general appears to be good where it needs to be
- ▶ Continuing to test and document ...





# S & T: Policies and Procedures

---

- ▶ Policies and procedures that govern data access and work with software tools need to be developed:
  - Disseminating data access policies and procedures
  - Disseminating metadata and data dictionary
  - Disseminating ‘sample’ data to aid in formulation of data requests
  - Managing and verifying data access eligibility
  - Managing application for data & possible payment for extraction/dataset creation
  - Creation and archiving of analytical data sets
  - Dissemination and tracking of use of analytical datasets

# S & T: Specific Roles of Information Scientists

---

- ▶ Roles of Information Scientists sharpened in a standards-based, metadata-driven system
- ▶ Information Scientists:
  - Are explicitly responsible for creating, *managing, maintaining the metadata* – i.e. conceptual level
  - Function as *the link or interface* between the data system and the other scientists/investigators
    - **Must acquire some domain-specific knowledge in order to function in this role**
    - Responsible for properly translating other science questions into efficient representations in the data system
    - Responsible for ensuring timely, documented, accurate output from data system

# S & T: Software tools

---

- ▶ Software ***must*** be developed and maintained by professional software developers, ensures:
  - Testing & quality control
  - Documentation
  - Maintenance
  - Up-to-date technology
- ▶ In break with past tradition, should be designed from output-to-input, rather than reverse, to guarantee that high quality analytical data in the correct format can be easily obtained
- ▶ Recommend that the design make heavy use of metadata-driven approach → flexibility, extensibility, documentation & longevity of system

# S & T: Software tools

---

- ▶ Integrated software tools required to manage:
  - Data
  - Data collection instruments
  - Workflows: data collection and data dissemination
  - Quality control and data integrity
  - GIS integration
  - Sharing, pooling, archiving
  - **Analytical datasets:**
    - **Creation**
    - **dissemination,**
    - **archiving,**
    - **Tracking**

# Key Issues

---

- ▶ Institutional change
- ▶ Human resource development
- ▶ Standardization:
  - Metadata
  - Metadata bank
  - Documentation
- ▶ Software tools
  
- ▶ **HAVE NOT ADDRESSED COSTS**
  - Likely to be significant, ~ 10 ( $\pm 5$ ) % of total operating costs

**END**