INTEGRATION OF GENEALOGICAL INFORMATION

Extended Abstract of a Paper by Bruce D. Despain for the Family History Workshop

Authority Systems in the Family and Church History Department of the Church of Jesus Christ of Latter-day Saints have studied carefully aspects of all six subjects categorized under the workshop heading of Integration of Information. This will introduce a context in which it will be possible to say something specific and coherent about merging certain of our legacy data.

To simplify the discussion we use a few special terms hoping to make the concepts easier to understand.

Record: a set of identifiers for an object, including identifiers for certain other related objects. The record is divided into fields.

Field: an identifier in the form of various kinds of data. Fields have values.

Value: the data in a field. The value may itself be a record. The fields of such embedded records are more properly identifiers of other entities. For example, a person's birth date is more properly the date of a birth event, but the date has been abrogated to serve to identify the person, the linkage entity. These attributes of related objects have become part of a *linkage record*.

Linkage: the process of declaring that the data of two records identify the same entity. This process is a prerequisite to merging or integrating data relating to the same entity.

Matching: the fact of two records identifying the same entity. Linkage is not accurate unless the records match, *i.e.*, the entities represented by the linkage records are in fact the same.

Probabilistic record linkage: the assignment of a probability to the declaration that two records are matched.

The first section discusses genealogical resources using as examples the family records implicit in census enumerations and marriage returns. Typically documents need to be related to a common structure expressed in a linkage record with comparable values in certain standard fields.

This more holistic view of integration introduces a second section describing the process of linkage and its various applications to genealogical data. Here we distinguish three basic levels: 1) event record linkage for documents, 2) individual record linkage for persons, 3) family record linkage for nuclear families uniting persons (and events) together in a single structure:

Event: an entity having attributes of time, place, individuals (or families), and relationships to the individuals (or families). These are the kinds of events represented by data in documents involved as the sources for genealogy.

Individual: an entity having attributes of: names of various types, sex, and certain vital events, *i.e.*, birth, death, and various other events. Individuals may have a relationship to multiple families, or to multiple events.

Family: an entity having attributes of: names of various types, and certain events, *e.g.*, marriage, divorce, dissolution, births of children. Families have a relationship to two or more individuals, or to multiple events.

We then discuss these three levels of the process in turn giving details through examples.

The third section builds on typology of record linkage to describe some of the technical details in terms of models and engines and a couple of other objects they need in order to work.

Lineage-linked model:

a data structure that contains records of individuals (either as parents or as children) and families with an indication of the familial relationships (individual/child, individual/parent)

Lineage-linked engine:

a process designed to operate on records of individuals and families based on the relationships between them

Entity class transformation:

an algorithm or function that operates on entities of one class to allow an engine to compare entities, *e.g.*, individual, in one data structure, say, as parent, to the appropriate instances in another, say child, or parent.

Data propagation rule:

a statement of how data may be assigned to a standard but blank field based on 1) the information in one or more other fields, and 2) knowledge of the culture of the individuals and families represented.

The next few sections discuss the results of an analysis of a certain corpus of data — the family group record archives (FGRA). The idea was first to explicate the algorithms required for use on this particular corpus. The corpus available electronically was not structured, except as individuals and families, so we wanted to put them into the lineage-linked form. Originally the sheets themselves where provided with an elaborate cross-referencing system that had been lost. This project used certain principles of analysis that should generalize to apply to record linkage in any other collection. The next section looks at the particular algorithms developed in this project. One would need algorithms similar to these to use the record linkage paradigm of this paper.

In the final section we consider the kinds of data that might be propagated, *i.e.*, guessed at or estimated, in filling out the individual identifiers of a genealogical record linkage entity. In some cases it is necessary to undo the effects of propagation to get at the data that record linkage must take into consideration. The *date* of an event is one of the simplest data structures, yet its propagation is the most involved and precise. Next in complexity is the *locality* of an event. The data with the simplest semantics, and the least amenable to propagation is the *name* of the individual. The semantic base for personal names are the individuals, which are identified by vital events having dates and localities. The vital events belonging to an individual and which a system may propagate are birth, marriage (for two individuals), and sometimes death. It is possible to propagate a death from a probate or burial record, just as one might propagate a birth from a christening or other individual record. It is also possible to propagate in either direction: a possible probate, or burial from a death. In this section we consider only the most basic rules.