fhiso

CFPS 70

(Call for Papers Submission number 70)

Place: Functional Requirements

Submitted by: Johnston, Wesley

Created: 2013-05-07

URL: Most recent version: http://fhiso.org/files/cfp/cfps70.pdf
                This version: http://fhiso.org/files/cfp/cfps70_v1.pdf

Description: This paper identifies several functional requirements for places
                within a modern GEDCOM database. This is a "what" and not a
                "how" paper. The paper does venture into some aspects of
                software implementation requirements, beyond the pure
                representational requirements of GEDCOM. Place is defined as a
                specific geographical location, which is independent of political,
                temporal, administrative, ecclesiastical or other man-made
                divisions but which has those all as potential attributes at any
                given time. It is very important to explicitly state that "place" does
                not necessarily mean "town" or "city"; it can be as specific as a
                building at an address.

Keywords: place, town, city, street, township, county, country

Place: Functional Requirements

by W. Wesley Johnston
May 7, 2013

Abstract

This paper identifies several functional requirements for places within a modern GEDCOM database. This is a "what" and not a "how" paper. The paper does venture into some aspects of software implementation requirements, beyond the pure representational requirements of GEDCOM. Place is defined as a specific geographical location, which is independent of political, temporal, administrative, ecclesiastical or other man-made divisions but which has those all as potential attributes at any given time. It is very important to explicitly state that "place" does not necessarily mean "town" or "city"; it can be as specific as a building at an address.

Context: Purpose of GEDCOM – Original conception vs. modern needs

GEDCOM stands for Genealogical Data Communication. It had – and still has – two essential purposes:

1. Storage of a genealogical data base
2. Sharing a genealogical data base, including the merging of multiple databases

Over the years since the development of GEDCOM, technology and genealogical research and the use of technology by genealogists have seen significant advances, which have greatly expanded the realm of possibility for the handling of places within GEDCOM.

For a technological example, the original GEDCOM conception of place was a town or city, within a hierarchy that included the county, state and country – as they were at the time of the event. This significantly violated the database principle of third normal form, but at the time the implementation of true third normal form for place was technologically a practical impossibility. The result was that the same place had multiple representations, so sorting, searching, aggregating, reporting or analyzing by place was very difficult. Furthermore, there was no conception of place within a city or township. "Place" meant a location with granularity no finer than town or city. Any lower level place was relegated to being an implicit element within the note text for an event, making analysis, searching or sorting by street or other lower contextual level of place impossible within the original GEDCOM. Modern technology can handle the complexities of places within multiple hierarchies, over the course of time, much more robustly, so that the new GEDCOM can support more robust handling of places.

For a genealogical example, the original conception of GEDCOM was as a purely lineage-linked database, aimed primarily at a bare bones tracing of one's ancestors. Most genealogists now recognize that their ancestors lived within a complex milieu, with many kinds of connections not supported in a purely lineage-linked database. For example, witnesses to a marriage have a connection that should be explicit and not implicit within the text of the event. The linkage of a person to a place and a place to another place matters: good genealogists always search the adjacent families in a census, since they know that their family most likely interacted with those families. The people who lived at the same address 20 years apart might differ but might still connect to each other. And the most comprehensive genealogists went a step further, compiling community trees to reconstruct – family by family and house by house – every family within a community, something which the original pure lineage-linked GEDCOM conception could not properly support.

Thus for both technological and genealogical research reasons, GEDCOM can and must now support more robust handling of places.

Paradigms to Keep in Mind when Considering Place

There are two ways in which a continuation of the narrow paradigm of place in the existing GEDCOM must give way to new paradigms about place.

First, the paradigm of place must go below the level of a city or town and include explicit representation of street address or township, range, and section or concession and lot or whatever form the local level of a specific location has. As I note in a separate section, the sub-city place should be a separate contextual level, which may have very different form.

Second, the paradigm of place within only a lineage-linked database must give way to a paradigm of places connected to places: e.g., the house next door, the same house 20 years later. Parish or town reconstructions are a very important – and growing – part of what genealogists are doing with records. The Familien Bucher web site is the best example (www.ortsfamilienbuecher.de), with very impressive coverage. Another is FamilySearch Labs' collection of community trees (histfam.familysearch.org). When a community is being examined and the families connected over the course of several hundred years, the higher levels of the geographic hierarchy really do not matter in understanding the families and houses of the community. What matters is what house people lived in or which nearby town they lived in or who their neighbors were. This paradigm really blurs the line between family history and local history, but there really is significant overlap of these research fields, as any good genealogists knows. Community trees are an extremely valuable way for genealogists to discover the complex relationships that existed beyond the pure bare bones of their blood ancestry. So GEDCOM should support the paradigm well.

In both cases, the user should be able to sort, search, aggregate, analyze or report data by place as an explicit representation.

Place Fields – Both As-Documented and Standardized

A new GEDCOM should treat place with just as much importance as people and events. There really need to be two fields for each instance of a place: one field for the entry of the place as it is specified in the record and one field for entry of the standardized version (which will necessarily be anachronistic in many cases) of this place, so that sorting, searching, aggregating, reporting or analyzing by place are not only possible but are robustly supported.

As will be seen below, I advocate a two-contextual-level approach to place, since the reality of how people use and record places has such a duality. The reason that I raise it here is because when I say "one field" above, I really mean one field at both contextual levels of place.

Going beyond GEDCOM to the software that implements it, the ideal would be that the standardized field would have robust interactive smart suggestion of existing places presented to the user while a place is being typed, so that the standard place name really is standard and not mis-spelled or otherwise mis-entered.

There are efforts to generate a unique identifier for historical places (e.g. gov.genealogy.net/search/index), which would allow a common standard across different databases, supported by software that implements GEDCOM. But the user of the software should be allowed the option to establish their own standard. This will of course complicate the sharing and/or merging of the database.

Places within multiple domains

Place is defined as a specific geographical location, which is independent of political, temporal, administrative, ecclesiastical or other man-made divisions but which has those all as potential attributes at any given time. It is very important to explicitly state that "place" does not necessarily mean "town" or "city" ; it can be as specific as a building at an address.

The same town can be in different counties at different times. The same house can have different addresses at different times. The same house can at the same time be within a religious parish, a civil district, multiple election districts, etc.

Ideally, this complex network of multiple connections of a place to the domains that contain it would have a robust implementation, so that everyone in the same city ward might be searched or sorted or aggregated. This really pushes the limits of GEDCOM as a pure lineage-linked database and even pushes the boundaries of what a modern GEDCOM database would do. But as more genealogists seek to squeeze out the very last drop of information from sources, this grows increasingly important.

Places as two contextual levels

Consider the way we actually refer to places (and thus how letters, documents, personal histories, newspaper articles, interviews, etc. would have referred to them in the past) and also about how GEDCOM has referred to places until now. People use a "natural" dividing line. That line in people's thinking is the town or parish or - for rural areas - the township. It is where they think of the difference between "here" and "there", the difference between "hometown" and the rest of the world.

There really are two contextual levels to place in the way that people think. There are those places within the boundaries of "here". And there is a group all the "heres" at higher levels. This really is how people have come to think about places. You run into this (maddeningly) when you see a clipped-out newspaper article that tells you what street someone lived on but never mentions the town because the town is the other contextual level, the "here" that does not need to be repeated every time you reference a place that is within "here".

So what does this have to do with GEDCOM? It means separating these two contextual levels.

1. We keep the town-and-above contextual level, in a much broader attribute-rich form than the old GEDCOM.
2. We also have an explicit new contextual level that can exist by itself (whether a separate entity or not I leave to the developers) that has never existed in GEDCOM: the contextual level of a place within a "here". This would certainly support recording an address from a clipped-out newspaper article without having to fabricate or guess at a "here" for it. The form of this contextual level can be very different within different "heres", although within any given "here" the form probably will be mostly the same.

Places as research conclusions

Sources can refer to the same place in more than one way. Towns changed their names. We are faced with determining whether two places are one and the same. If we determine that they are the same, then the implementing software needs a robust way to merge places, so that the standard name is then used for both instances, while the as-documented name is kept separate. There are also needs to be a way to document the conclusion that the two places are the same. And ideally, there needs to be some way of undoing the merge of the two, if later evidence reveals that the conclusion that they were the same was not in fact the case.

Universal place identifier generation

As noted above, in the section on as-documented and standard place names, there is a German web site (e.g. gov.genealogy.net/search/index) that now generates a unique identifier for historical places. It now concentrates mostly on northern Europe. FHISO should consider taking on the on-going ownership of such a system for the entire world (which would clearly be a major

perpetual project for FHISO) or else consider supporting the extension of the GOV project to the entire world. A truly global standard would enable software developers to implement robust support of the identifier as the standard identifier for a place at the town level or above. I doubt that the lower contextual level, within a "here", can ever be standardized for all parts of the world.