A Few Thoughts about GEDCOM and Better GEDCOM

I was asked to put down a few of my thoughts about the Better GEDCOM effort and what I thought it should encompass. After a bit of an introduction I have listed five steps I think are important to make Better GEDCOM successful.

Introduction

I'm a retired software developer with advanced degrees in Computer Science and 44 years of software experience. I worked for Bell Labs for 17 years, where I was a Distinguished Member of Technical Staff and Technical Supervisor. Before Bell Labs I had other software jobs, including Head of the Computer Programming Section at the Geophysical Institute, University of Alaska. After Bell Labs I spent eleven years as a software architect for three internet startups, two of them successful and going strong.

LifeLines

I've been writing genealogical software as a hobby since 1988. In 1990 I released LifeLines, a genealogy program for UNIX systems. LifeLines uses GEDCOM as its database format, but does not enforce any GEDCOM standards, so it can import GEDCOM data from any source without loss. LifeLines users can structure their GEDCOM records to any depth using any tags they choose. If they use the lineage-linking tags of GEDCOM 5.5 LifeLines understands the relationships between persons, the standard events, and acts like a "normal" genealogical program. LifeLines is now an open source project hosted at Source Forge.

I added support for Event (tag EVEN) records in LifeLines, and I also allow users to create their own top level records with tags of their own choosing.

Structured Flexibility

Using LifeLines I gained experience structuring GEDCOM records using unofficial tags and sub-structures. I did this because the 5.5 standard was too limiting to hold all the data I wanted to record. In so doing I realized that the hierarchical tree-based structure of GEDCOM is an excellent solution to a problem I had come to call the "structured flexibility" problem. This is the problem caused by two opposing needs – the need for regularly structured data to enable effective computer processing, and the need for flexible data to enable the recording of highly variable and unpredictable genealogical data.

I wrote a short article based on these ideas, “Structure and Flexibility in Genealogical Data,” and was invited to give a talk based on that article at the 1994 GenTech conference in Dallas, Texas. Partly as a result of discussions generated by my talk, GenTech put together the effort that eventually produced the GenTech genealogical data model.

GenTech

I became a critic of the GenTech model early on when it became clear that it was complex and esoteric, with a number of unnatural data types, including the difficult to understand assertion data type that had to be used as the glue to hold together most of the other data types. The model was designed in terms of normalized data types intended for direct implementation in relational databases, which was something I was strongly opposed to.

DeadEnds

Based on my experiences with LifeLines, in my GenTech talk I advocated a model based on GEDCOM-like hierarchical objects that would both extend the set of existing GEDCOM records into the realm of evidence objects, and allow for richer and more flexible and possibly user definable substructures within the records. I advocated a small set of object types that were natural to the process of genealogy. The Event record has always been high on my list of necessary genealogical records.

I started work on my own genealogical data model about the same time, both to enhance the ability of LifeLines to handle a richer set of data, and to be the model I thought the GenTech model should have been. I called it the DeadEnds model.

Better GEDCOM

Better GEDCOM is an effort that got underway in late 2010, and I’ve been contributing comments and ideas to the effort effort. The goal of Better GEDCOM is to define an archival and transport file format that allows users of different genealogical programs to record and share all their data independent of any specific genealogical software program. It is early in the Better GEDCOM process so it is not yet clear how this goal will be met. Some contributors believe that a formal clarification of the current GEDCOM 5.5 standard with additions to fill in some gaps is what is needed. Others believe that Better GEDCOM provides the opportunity to create an entirely new transport file format based on a more complete model of genealogical data processes, taking into account the full evidence to conclusion research process. And there are others who see Better GEDCOM as an evolutionary effort, one that starts by patching up GEDCOM as its first product, continuing

on through more steps to the larger goal of a more complete genealogical data model and its transport format.

The elephant in the Better GEDCOM living room is whether any of the standards it produces would have any impact on the commercial genealogical software industry, an industry that has demonstrated a surprising lack of will in wanting to solve the data sharing problem for their users. The Better GEDCOM team optimistically believes their work will cause a ground swell of demand for better sharing, and that the big money players like Ancestry and FamilySearch will be compelled to come on board. Personally I believe this is a naïve point of view, but it is still a laudable goal.

Problems with GEDCOM

Any effort to replace GEDCOM as the de facto genealogical data transport format should probably start with a list of GEDCOM's shortcomings for that purpose. Here is my list.

GEDCOM Is Restricted to a Simple Model of Genealogy

The information that can be conveyed in a GEDCOM 5.5 compliant file is restricted to simple information about persons and families. The information that can be specified includes basic (e.g., date and place) information about a number of vital events (e.g., birth, death, burial), a number of obvious properties (e.g., sex, name, occupation), and references to the families the person was either a child or a spouse in. Any other information must be relegated to unstructured notes.

GEDCOM Records Hold Conclusions

GEDCOM 5.5 is based on Person and Family records (INDI and FAM) that hold information about conclusion objects. Each Person record is intended to hold all the information that a researcher has discovered about a real person. There is no facility in GEDCOM to transport the evidence behind the research that led to the conclusions. For example, if birth information for a person were extracted from a number of sources, say by estimation from census records, from the inscription on a gravestone, and from an official birth record, there is no easy way to include this information in a GEDCOM Person record. GEDCOM Source records don't convey this information because Source records specify where genealogical information is located, not the information itself. At best the evidence must be put into unstructured NOTE lines in GEDCOM records.

GEDCOM Has No Multi-Role Events

GEDCOM has no concept of a multi-role event (other than marriage events that are treated only in the context of Family records). All event information in GEDCOM must be placed in vital event substructures inside the Person record of the primary role player. Take for example a birth record. Information from the birth record is used to create the birth event substructure in the record of the person who was born. This includes only information about the place and date of the event and very little else. What if the birth record also named the person's father and mother, their birthplaces, and the father's occupation? Where does this information go? With GEDCOM this information can only be used in the Person records of the person's parents; the evidence must be distributed into many different records and its integrity and identity is lost. How would the information about the father's birthplace be represented? In GEDCOM it would have to be done through a birth event substructure added to the father's record, and in most cases if the father already had a more complete birth event from another source the information from the child's birth record would be lost.

Solving GEDCOM Problems with Better GEDCOM

Listed below are some ideas on steps Better GEDCOM could take to solve the short comings of GEDCOM 5.5 list above. First a note about XML.

GEDCOM and XML

GEDCOM uses a specific syntactic structure that defines its format as a sequence of numbered, tagged lines that form hierarchical trees of tagged, text-based information. Representing information in this hierarchical manner is a natural and popular way to represent data throughout a wide array of computer applications. This has lead to the popularity of XML as a general-purpose syntax for holding tagged, text-based trees for any type of computer application. GEDCOM and XML are two different syntaxes for expressing the same kinds of information; they are isomorphic to one another. GEDCOM files can be easily transformed into XML files and vice versa. The decision to express Better GEDCOM in GEDCOM or XML syntax is not a decision that impacts other issues. It is entirely possible for genealogical applications to generate both GEDCOM and XML based transport files.

Better GEDCOM should design an XML schema for its transport file format as it is almost definite that XML is the way of the future and GEDCOM syntax is headed for the dustbin of history.

Notes on Ordering Steps

I don't see a natural ordering between some of the steps listed below.

Extend and Formalize the Sets of Tags Allowed

Two complaints about GEDCOM are that the set of tags (referring to the 5.5 lineage-linking standard) is not large enough for holding all the information one would like to transport between genealogical systems, and that the tags that do exist are often interpreted and used differently by different genealogical systems.

Better GEDCOM should address this problem for all record types (taking into consideration that points below recommend more record types).

Evidence Based Person Records

Person and Family records in 5.5 GEDCOM are used to hold a researcher's conclusions about the persons and families he/she is researching. Many genealogical databases also hold records for the evidence information that has been extracted from sources. GEDCOM cannot transport this information. The DeadEnds model extends its Person record to be able to hold both evidence and conclusion information, and conclusion Person records can be built up from trees of evidence and lower level conclusion Person records.

Better GEDCOM should support a model that allows it to transport both conclusion and evidence information, while maintaining the relationships between types of information.

Multi-Role Event Records as Both Evidence and Conclusion

GEDCOM does not support multi-role events, though a proposal from CommSoft in 1994, "Event GEDCOM," provided an excellent method for handling them. Event records hold the date and place of an event and contain a number of role sub-structures that are references to the Person records of the persons who play roles in the event. Event records can be both evidence, taking directly from a source, or conclusion, built up from a number of evidence events.

Better GEDCOM should include multi-role Events as a supported record type.

Place Records

GEDCOM uses PLAC lines in event substructures to hold the names of the places where events occurred. Because many places tend to be repeated in genealogical databases this leads to a waste of space and the potential problem of multiple updates if place names are modified. For these reasons many genealogical programs use separate records for places,

often arranging those places into hierarchical structures. For example, Place records for the Canadian provinces would refer to the Place record for Canada. Place records for towns in the United States would refer to county Place records which would refer to state Place records.

Better GEDCOM should include hierarchical Place records as a supported record type.

Other Record Types

A number of other record types are often discussed.

One ongoing issue is the need for the Family record and its possible replacement by a Group record that could be used for any kind of group of persons. Better GEDCOM must still transport the information from existing programs so there must be some mechanism for handling families.

Another record type sometimes mentioned is an Evidence record, which holds information about an item of evidence that Person and Event records are derived from. If an Evidence record were used, the Person and Event records would refer to them, while the Evidence record would refer to the Source record that the evidence is extracted from.

Another discussion point concern the Source record and the possibility of a Citation record. In reality a citation is a string of text used to represent the location of evidence, the string to be used in footnotes or bibliographies. Citations are generated from the contents of Source records when they are needed; they don't need to be represented by their own record type.

A difficult point concerning Source records involves that vast number of things that can be used as sources during genealogical research. The Source record must be rich enough to handle all of them. The most common thinking about this is to look to Elizabeth Shown Mill's work in describing how to reference all types of items, and then build into the Source record all the attributes needed to support the range of items. Then all the information for generating citations will be available and templates can be used to define to generate the strings from the contents of the Source records.

Finally there is sometimes discussion about a general purpose record type that can be used to record information about any arbitrary entity discovered during research that the user would like to record in his/her database.

Better GEDCOM should evaluate the need for these possible record types. If major software programs support the underlying concepts, then Better GEDCOM must evaluate the benefits of being able to transport that information.