SFGS to Access to GEDCOM to RM – the details

In August 2011 I attended my first SFA family reunion. Jeanne Jeffries made a plea to have someone convert the Sheldon Family Genealogical System to a normal program so a book could be printed. I was not interested until she said Microsoft Access, which I knew how to use and wanted to learn more.

I received a computer file copy of the CD file that was the SFGS Access database the next day. I made a few calls to Rose Newton as I couldn't seem to open it. This was my fault in not understanding the newest versions of Access. I needed to change some display options in Access so I could see all the tables, queries, forms and reports that made up the data, and I was used to seeing.

I quickly realized there was a few tables that had links that were missing. Trying to open some of the tables gave an error message of "C:\Documents and Settings\bret\Desktop\SFGS Dec 2003.mdb is not a valid path". This told me that the file I had wanted to find Bret's hard drive and it wasn't there. I hoped this linked in file wouldn't be needed as Bret was long gone from the project.

I was able to see the SFGS table structure with a few more tries and found most all of the data to be in 3 main tables. They were the Master, Detail and Marriage tables. There was a fourth table called Sheldon Source Table that contained the places the data had been found and quoted from.

The SFGS file was a very old Microsoft Database (.mdb) file format that was changed to .accdb in 2007. That meant the database hadn't been upgraded for at least 4 years to a newer modern version of Access. I was able to convert the old .mdb to .accdb in a few minutes using normal user commands.

To understand a database you need to see what fields are created to save information and how the tables fit together. I found a single field called COMPNO that was the only link between the main 3 tables. It was also the only field that had been indexed so there couldn't be duplicate values in it. This concerned me as duplicate names or sheldon numbers could easily be entered with no warning or validation.

In the next week I learned about Sheldon Numbers and Computer Numbers and how the 3 tables fit together. Access was great as it could sort and filter any of the tables by any of the fields in the blink of an eye. There were about 70,000 Master records with a few less Detail records and about 36,000 marriage records. I could sort any field in any of these tables and find hundreds of problems. Things like date of birth that were in the future, or before 1500. It looked like there were LOTS of typo problems in the data.

After seeing the CompNo was the link between the three tables I examined all the other fields of the three tables to decide what information they were keeping. I soon discovered by sorting each of the fields individually there many blanks in the data. Lots of the data that wasn't blank was of the wrong kind of information. Text instead of numbers in a "Year of Birth" field was this type of error. I also discovered in the Detail Table some stings of garbage characters that indicated the file had crashed due to a power outage or not being closed correctly. This concerned me as this kind of corruption could invalidate lots of good links.

I decided to do an easy problem first in dealing with the database. There was a field called REFER that was a pointer to the sources that were used to type in the record's information. It looked like "G001,G012,J058" which was the Sheldon Source table number of where the data was found. To have

multiple pointers in a single field was bad practice but I could make a new table and fix this problem. I soon realized that some of the number zero had been typed with the letter O. Computers know this is different and don't match up while people reading or typing it can't. There were also about 300 records where the REFER had notes instead of the 4 digit pointer notation in them. It took me a few days of cleaning to correct the random text and set the pointers correctly for this one field.

My next cleanup step was to remove all the data that had "?" and "Unknown" typed in the data. If it wasn't known then a blank was a much better way to leave the data. I soon found there were more than 64 ways that "Unknown" was spelled in the data. I also realized I needed to do the searching in every field of all 3 tables to clean this up. This seemed like an easy problem but took a few days of searching and replacing these words with blanks.

I next found a huge amount of data that seemed to be missing or redundant. An example of this is, there was a "Year of Birth" field and another field asking for "Date of Birth". I could use Access queries and easily find records that had DOB and no YOB. There were about 300 of these where I could add the YOB based off the exact DOB date. I also could check to see if the existing year in the DOB was the correct YOB number.

This first round of cleanup took about 3 weeks to complete and I started wondering what other genealogy programs had that SFGS didn't and why. I found that other software collect a lot more information that needed to be broken out. SFGS needed a lot more fields to save nicknames, name prefix and name suffix information. I created these extra fields in the SFGS and started finding all those Jr., Sr., General, words that were combined into the name fields and breaking them out to where they needed to go. I spend a few weeks finding and fixing these types of name separation problems.

I knew the data conversion project needed to result in a book of some kind and started looking for a good standardized Genealogical program to convert the data into. I had heard of GEDCOM and how it was a standard between programs and started learning as much as I could about it. I soon discovered that the standard wasn't really standard. Each different program had small differences in them that kept them from all working together seamlessly. One might have a nickname field and the other not. Some might allow notes in some places and others none. I researched and decided to standardize on RootsMagic, based on many reviews and features available.

In the first few months I had found and fixed a few thousand single field problems. I really hadn't tackled many of the harder problems. I had yet to learn GEDCOM or how to get this Access data into it. I was very good with Excel VBA and felt I could learn enough about Access VBA to see if it was possible. I still hadn't resolved how to merge duplicate people records together yet but knew RM had that feature and was thinking it could help with the chore.

At about the 7 month mark it became clear I needed some more ways of dealing with the SFGS data. I tied the 3 tables together in a single form so I could see a single person and their details and spouses all on a single screen/form. This was much easier to do with newer version of Access and let me see things I hadn't seen before in the data.

From the above view I discovered a way to see a single extended family. Using the COMPNO field I could filter a person's Parents, Siblings and Spouses. Using this filtering method I found many typos in

Sheldon Numbers and Computer numbers. I fixed all these by hand as there was no rule I could follow to fix them.

Because I could now create my own form views of the data and build some easy filtering tools this lead me to a discovery to continue cleaning the data. Each Sheldon had the names of his parents listed in their record. If there were 5 siblings from the same parents, the parents' names were typed in for each child. I created a view of the data where I could look at the parents' names from the master table and what the child's records showed as the parents' names. I created VBA code that would run through all 70,000 records and stop when things didn't match. This helped tremendously in name cleanup as some children would show Bob as the dad while others would show Robert or Rob. Making all parents of this siblings names spelled the same was a great help.

Access is wonderfully fast and has more features than I know how to use but one that I used frequently was "find unmatched records". This query would show me things that were very hard to pick out. The first was to find detail information that had no master data connected to it. This should NEVER happen and showed me frequently how poorly constructed the SFGS data was. I had no way to find or link these orphan records back to where it might belong, but knew people had spent thousands or hours collecting and entering data and I forged on knowing the data wasn't perfect. I was making it better and better with each fix I was performing. I told Rose to stop entering new records into the database in early 2012, knowing I could never reproduce the thousands of manual fixes I had performed.

After studying GEDCOM more it had a record for each person. In the SFGS the foundation was a SHELDON that then connected to spouses and then to spouses parents. The two structures were like Roman Numerals vs our normal numbering system. I somehow needed to give all spouses a number and also their parents.

I started studying the marriage table of 36,000 spouses. This was a nightmare as they might be the first or second up to 5th spouse tied back to a single Sheldon person. Worse is some of the spouses may have married multiple sheldon's and really need two (or more) valid numbers. Then I discovered Sheldon's who married Sheldon's in the marriage table and the real duplicate problem became clearer. I also realized that some of the marriage duplicate records were missing. Marriage dates needed to be checked and the multiple marriages made this much too hard.

It was somewhere in the 9 months dealing with the data that I decided to give each person their unique Computer Number. I would give the first spouse a 01 at the end of the attached Sheldon Number and the second a 02. So if Sheldon with CompNo of G123 married a second spouse that spouse would get their own computer number of G12302. The parents of that spouse would get G123020M for Mother and/or G123020F for Father. This new Computer numbering system was really needed as a start to get to GEDCOM.

To this point I was telling everyone who asked that the problem might not be solvable. I was finding more and more problems and not getting to the final goal in any organize fashion. I went back and started dealing with Access and GEDCOM.

I spent at least a month with Access trying to export the data to a GEDCOM text file. I had built small RM family trees and exported them to GEDCOM so I would know what they looked like. I had learned

enough of the GEDCOM tags to understand a lot of the SFGS data wasn't needed or redundant. I finally discovered that I needed to write the VBA code behind the Access FORM to accomplish this step.

At about month 10 I had written VBA code to export GEDCOM text lines for the normal information in the SFGS data. It took about another week to figure out how to export BIONOTE or memo fields to GEDCOM. I slowly build up this VBA code to do a single person. After the VBA was good enough to do a single person I could do a look and do all PEOPLE. My first few tries with this took the computer more than 12 hours of running time to build the People GEDCOM table.

In GEDCOM the People Table was only half of what is needed. Each person needed to come from a FAMILY and in the Family table it kept track of who the parents and children were in that family. Here was a structure that SFGS didn't really have. SFGS had a marriage table but didn't keep children and Husbands and Wives like the GEDCOM needed. In SFGS a person could be male or female but in GEDCOM this was a huge part of the solution. I spent a few weeks dealing with this problem.

I needed to build a new tables I called Families in the SFGS Access database. It needed Husbands, Wives and all the Children's computer numbers. This lead back to Access where I built another tool to show the children going to which family. This was a nightmare problem when it came to multiple spouses. It was much easier if the Sheldon was a woman as the children might keep their fathers last name. I build and ran many runs for each spouse number to build this connection. In some cases it was impossible to tell which spouse the children belonged to and I needed to guess. I hoped they were close enough to figure out later. The GEDCOM build of families only took a few hours of computer time.

The last piece of a "good" GEDCOM conversion was to include all the Sheldon Source pointers for each person. I had worked on this table in the beginning and was able to build this in a day or two for a very big GEDCOM build.

To assemble a GEDCOM import file I needed a Header that told it was coming from RootsMagic. I could use a RM export for this section. Then I appended the People section of the GEDCOM followed by the Family and Sources sections. Lastly I needed a footer section in the GEDCOM that RM uses for how it produces it narrative reports. This HUGE text file ended up at about 36 megabytes size and took about 20 hours of computer time and about 30 minutes of manual time to build.

UNBELIEVEABLY!!! This GEDCOM file could import into RM and it took less than a minute to do it!! I was overjoyed this had worked but it was about a month before the next SFA annual meeting and I'd worked on the problem for almost a year.

The next step was to let RM use its merge feature and see if it could reduce multiple records of the same person to a single person. Sue Sheldon and I worked on this problem for a few weeks before the Northampton reunion where I presented my "Beta8" conversion of the data, claiming that it was now ready for the publishing house but needed some work to clean up duplicates. I believe the board was in disbelief that the data had been converted and looked much better than SFGS in Access (that only Rose and I had) and could be viewed using RM.

A few weeks after the 2012 SFA reunion Sue Sheldon found a missing person in my conversion. Benjamin Franklin Luther was missing from the RM conversion. We could see him in the old SFGS but he was gone in RM. This pissed me off!!! I had done all this work and we could prove there were errors in the conversion somewhere. I never found how he got lost but realized having RM do merges of people

wasn't going to work. There were so many people with similar names and not enough linking information Beta8 wasn't good enough to meet my standards.

A few months later I went to a Genealogical convention that was a few miles from my home. The Mormon Church had a few hundred people show up for seminars and lectures about the topic. I wandered into a genealogical database class and asked about merging duplicates. The answers I got from everyone was not good. There seemed to be no good answer on how to eliminate them. I did learn from one class that we can calculate birth years from other family members. Siblings should be 2 years apart as should Husband and Wife and kids should start when dad is 25.

At that time there were huge holes in birth years, in the data that was based on name connections. I was able to use existing YOB and DOB numbers and Calculate or Estimate every person's year of birth. This was fabulous for those 650 "Mary Sheldon" that were in the SFGS. It also put everyone within 20 years of their correct birth year. This took a few weeks to accomplish in the Access data that I knew I could GEDCOM over to RM.

Now that RM had failed to merge duplicate people together I needed to create something in Access that would work. I decided to create a second field for Computer Number. I called it DupCompNo for duplicate computer number. I filled it in with everyone's original computer number. Now all I needed to do was to change the DupCompNo to the number of the person who was their duplicate. At this time my mind would go in spins wondering what rule I needed to perform this magic trick and keep it all straight.

I discovered a method that might be useful for others in cleaning up dups. I started by claiming that people with the same names and sheldon numbers were duplicates. I felt pretty good about this assumption. I could also find families that had the exact same named Husband and Wife, which would mean the families were duplicates. I had created a DupFamNo field for each family and started fixing those also. Then children of duplicate families that had the same names might be duplicate people. This worked until a "died young" was found with different birth years and I didn't make them duplicates. I followed this logic in as many ways as I could discover. The most impressive was to discover parents of 5 different children who married into the Sheldon lines. This allowed me to merge 10 people into 2.

At the Northampton SFA reunion Frank Sheldon and I sat and looked at what I had found and I still had no idea how to use this duplicate field to build a new GEDCOM. The problem was that I could have 5 records with different information about people that were the same person. One record could say they died and another didn't know. One could have a dob that was different than another. Two records could have BIONOTES in them that could not be lost and somehow needed to be merged or saved. After some pep talk Frank made me try things that seemed to work.

It turned out that RM allowed two of the same kind of facts for individuals. That means I could put in a birth date twice, each one being different. Using this method I didn't lose any of the original SFA data even though I knew some of it had to be wrong.

Because I had added the Duplicate Comp Number fields for this new attack on the problem, my VBA code to produce the GEDCOM needed updating. It took me a few weeks to revise the code and debug it enough to get another GEDCOM. I had learned that I could make the 20 hour of computer work less by not displaying what was happening in the process. I cut the build time down to about 8 hours using

some VBA tricks and thing I'd learned. Finally on 29 October 2013 I did my final Access to GEDCOM conversion. YEAH!!!

Sue Sheldon and I started immediately looking at the RM file and seeing if it matched the original SFGS data. We knew thousands of fixes in information had been corrected but needed to know the conversion was healthy.

RootsMagic has a feature called Problem Search. This allows it to produce a report that can check date information for correctness. It will find problems like "Died before Born" or "Married before age 14". We ran this report on the converted data and RM found 85 pages of problems. Counting about 47 problems per page this total was 4000 problems with dates only. We knew these couldn't be correct and I spent many days checking the original SFGS to confirm my conversion was correct and I hadn't created these problems. It turns out all of these problems were in the original SFGS database.

We also worked on finding duplicate people using the RM tools and many were found and merged together. These were duplicate people that my Access tools didn't find. All this work and fixing in RM made it our Master as the Access conversion was done and over. Access had played its part to get the data to RM and my work should now be done. I wish that was the case but it wasn't.

A focused effort was directed to fixing the 4000 date problems in the RM data. Sue spearheaded this effort and sent out groups of names to people who were good with genealogy and computers. They called themselves the cleanup crew. They sent back their findings and Sue and I were able to correct lots of errors. Many of these error were typo problem by dyslexic typist. Something like 1986 should have been 1968 and we would confirm and fix these date problems. The number of pages of problems slowly decreased to about 35 (1500 problems) at the time of this writing (July 2014).

Some of the problems in the RM file were due to blank information for duplicate people. In one record it would have a dob and a blank in another. When RM pulled in the GEDCOM file it might use the blank as the dob instead of the real date. These could show up as a problem in the RM report. I decided to attach this problem using Windows Notepad at the GEDCOM text file level. Notepad would take a few minutes to load and/or find the next problem in a search dialog box. This took much too long so I found a text editing program called VEdit that did loading and searching much faster. I spent a few weeks learning VEdit and trying to reduce the blank facts at the GEDCOM text level. Sue had a problem with this as importing a new GEDCOM file messed up the existing RM record numbers, so this method was never implemented to clean the data.

In the early stages of the cleanup crew we tried to allow others to use RM and correct their own families' information. RM has many ways to drag and drop full branches of trees from a personal file onto another. We found this process was very risky and not completely understood. With over 123,000 people in the database we felt allowing others to use the master was too risky.

Sue has a full time job and I'm unemployed, retired and have lots of time on my hands. Early in the cleanup process I spent many hours finding and fixing date problems. My reasoning was that if the SFGS and now RM dates can't be correct, let's fix them. I'd use email sent from the cleanup crew and tried to verify or reject their information. If I fixed something, I would put a web tag to where I'd found the corrected date. I believe someone would need an Ancestry or Find A Grave account to follow some of these paths. I knew that anyone with these accounts would be able to confirm the data. I was

concerned that Rose may not have subscribed to these services and may not be able to confirm our findings by following the web tags.

Just before Christmas 2013 I was helping Rose understand all the things that RM had to offer. I was using Team Viewer and remote controlling her computer. I was concerned that she had multiple versions of the SFGS database on her machine and I wanted to see if any records were entered into different files. This process is called version control and is done in different ways at different companies. I downloaded to my computer her latest SFGS file. I was able to do a "Find Unmatched" query from the SFGS file I had worked on for 2 years and the one she was currently using. I found 970 new people that needed to be added to our RM data. I spent about 6 solid days manually entering these names into RM. I was really the only one who could do this as I'm the only guy who has Access and RM and knows how the whole puzzle fits together.

Since the 970 new people event happened, Sue has been driving the conversion project. She has realized the problems with the data were more than names and dates. There is a huge problem with "Place" information in the database. Places should be City, County, State and Country. We had a huge number of places that were none of this kind of information. Text like "d. young" or cemetery names or other notes were in this field. Also many of the place names are spelled differently for the same place. This would be like "Los Angeles, CA" vs "LA, CA" vs "LosAngeles, Calif". To improve the data these should all be corrected to be spelled the same. Computers don't know that LA is the same as Los Angeles.

Now the Data Conversion Project is complete. That means it has been converted from SFGS to GEDCOM to RootsMagic and my original task is complete. To check myself I tried to do a narrative report using RM of everyone in the database. Using some simple calculations it looked like the Sheldon Book would be about 4 feet thick, if all the RM data was printed onto 8 ½ by 11 inch paper. I pondered what a silly request it was in the beginning to have a single book of the SFA family(s). I also realized people thought the data was clean and healthy from the beginning which it was not.

Now that the Data Conversion process was complete, Sue and I felt a new job of Data Integrity needs to be performed. She laid out specific steps to do more cleaning of the data before it was released to the world for scrutiny. She outlined her desires in her Data Integrity Report to the SFA board.

While examining a posting of the SFA data on Ancestry Sue realized we can keep the data completely private and only allow "invited" people to view it. She used this process to have the Cleanup crew see what was in the master database. They would then send her email of their findings and she would then correct the master data in RM. After a cleaner master was created she would DELETE the Ancestry.com file from the internet and post the newer one with all the corrections. This lead to some confusion from the cleanup people. Some had added links to the Ancestry.com file thinking they would be saved to the master. This was not the case and any links they had done were lost. To solve this problem, Sue evaluated Family Tree Maker (FTM) which was easier to coordinate with Ancestry and see changes. FTM didn't have a Nickname field as does RM so the question has been put on hold for now. RM also has a much better method for merging duplicates and tools to find problems.

My hope is the Master RM Database will be updated by only a few people who completely understand how RM works and can keep from creating linking errors. I'd like to see the data published on Ancestry.com for all to see and use. My hope is that others will find problems with our data compared

to their own. I hope this will bring the two conflicting parties together to “show your proof”. If and when this happens I hope the Allen County Public Library will have all the SFA files and our proof can be found and shown. I hope our RM database gets better through this process and is updated and reposted on Ancestry as it improves.

There were many dead ends on trying to convert the SFGS database to RM. I’ve haven’t told the stories above of the many things that didn’t work. One such story was with Ohana Software that produced a product called FamilyInsight that promised they could find and merge duplicates in our data. After a few month helping them as a beta tester, Sue and I found our database was simply too large for them to deal with. This is but one of the many things that didn’t work to get to where we are today.

I hope the above gives the reader a better perspective on what was done in the 2 plus years I spent converting the SFGS database to GEDCOM and then to RootsMagic. I did this project for free, not knowing it could be accomplished. I knew that the only way it might be done was to make the problem my own and learn, study and never give up. I did this in memory of my father in law, Loren Edward Sheldon who taught me much about computers, databases and pride in a job well done.

Sincerely

Marvin Parsons (Spouse of Elizabeth Jeanne Sheldon)