The Current Genealogy Industry and ProgenyLink's Prominent Future Place In It

Introduction

The following diagram shows the major components and processes of the current genealogy industry. There are parts of that industry which are very efficient, and there are others which are extremely inefficient. It is the goal of the ProgenyLink project to offer new services in the area which is most in need of improvement, which is that crucial step of the final assembly of historical names into complete family structures.

It would obviously be very helpful to understand the main components and processes of the current genealogy industry and have a common set of terms which makes discussion easy. A glossary, together with the following diagram should also help everyone to have the same understanding of this complex and interesting industry.

Current Genealogy Industry Processes

The attached diagram shows the main functions which make up the current genealogy industry activities. It is a fairly simple chart and should be relatively easy to understand. The main problem will probably come later in other documents when we are discussing concepts and terms which are not clearly displayed on this simple diagram.

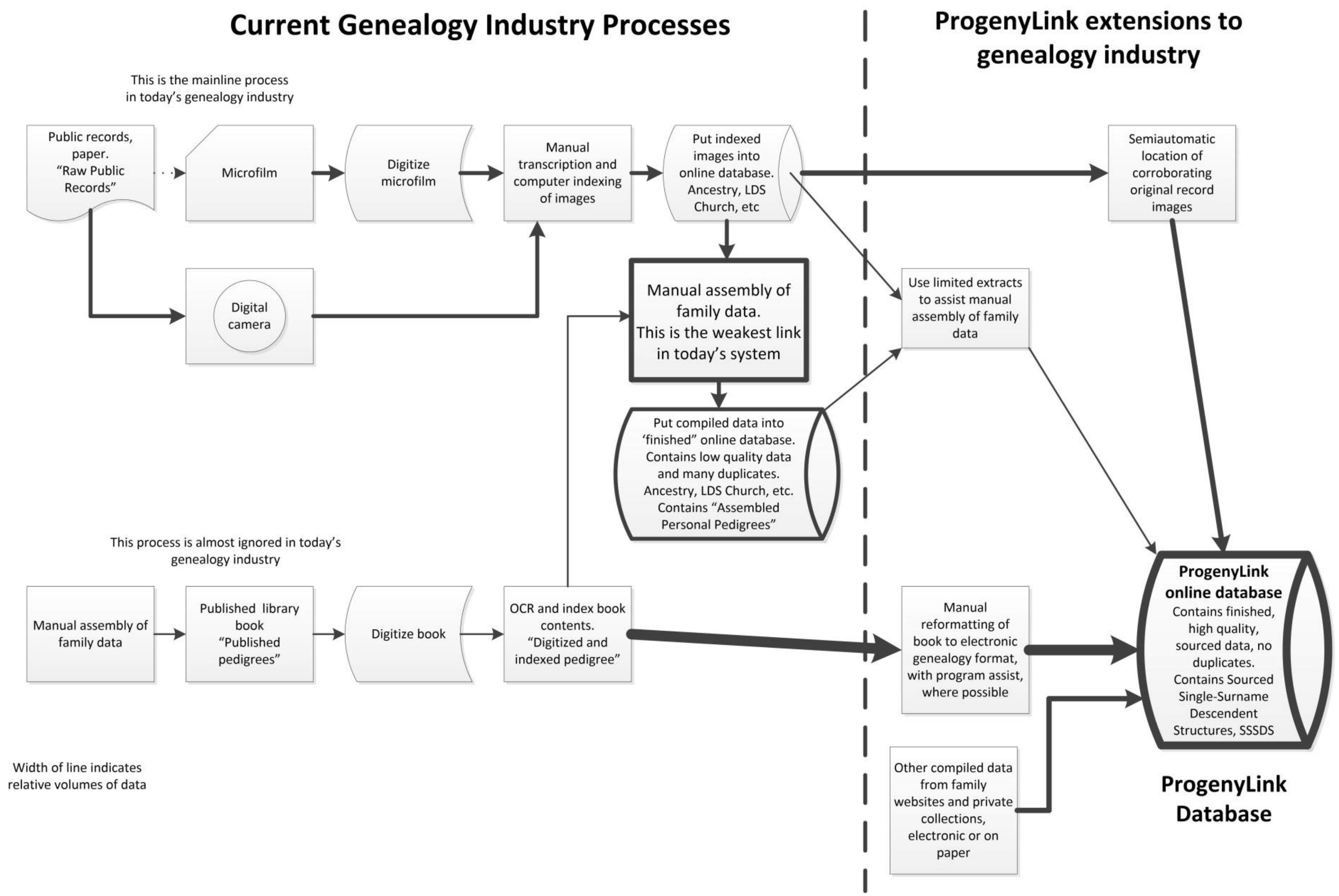
The diagram shows paper records being microfilmed and then digitized, as in the past, or simply being digitized as the first step when a digital camera is used these days. In either case, those ancient images are typically not computer readable, and so they must go through a manual step where people read the documents and transcribe them using a computer keyboard. That human-generated data can then be put into indexes for future use in conjunction with the source document images. The images and indexes are put into large online databases which are maintained by the LDS Church, Ancestry.com, and other such organizations. At that point they become the "raw data" which people then use in whatever manual processes they prefer to use to compile the data and then enter it into what everyone hopes is the "finished" version of those family structures, which typically are far from really being "finished" at that point.

It should be useful to quantify some of the current activities. It is estimated that about 10 billion people have lived on the earth for whom public records are still available. It is also estimated that the world has about 60 billion pages of genealogical records which are stored somewhere in the world, of which only about 10 billion have been imaged and indexed and are now available online. This situation should allow about one billion unique historical people to be identified from the 10 billion genealogy record images which are currently available. It is not clear how many names have in fact been placed into their proper family structures, but it is likely that the number is less than 100 million people, out of the one billion possible. We might notice that there are, on average, about six records for each one person who has lived on the earth. If true, that should provide a fair amount of information about each individual person.

There is another process which was very popular in the past, and which continues today, but which is mostly ignored by today's genealogy researchers. That is the process of families manually assembling family history data and then publishing a book which contains all of their findings. Typically, those books do not contain the underlying source records or any library references or computer links to those source records. This makes them less than ideal, although they do still contain a great deal of high-quality data.

A relatively new process has started to make these somewhat hidden and ignored books available for more general use. They are being digitized or imaged, so that they can be stored online and people can download the entire book and have it on their own computer. Most of them have also been put through an OCR (Optical Character Recognition) process which allows a computer to recognize the text in the book and make it searchable. That also means that researchers can copy and paste portions of the book into some new electronic format.

These published books are an important part of the ProgenyLink project plan, although many other sources will also be used. One of the major activities of the ProgenyLink project is to reformat the contents of these digitized books so that they can be entered directly into a new database. At that point, a new feature, which has never before been used, as far as I know, will be used to locate corroborating source documents in the giant online databases which can then be linked to these names in the ProgenyLink database. This will provide the corroboration and quality which is missing from nearly all genealogy databases today.


Current Genealogy Industry Processes
ProgenyLink extensions to genealogy industry
This is the mainline process in today's genealogy industry
Public records, paper. "Raw Public Records"
Microfilm
Digitize microfilm
Manual transcription and computer indexing of images
Put indexed images into online database. Ancestry, LDS Church, etc
Semiautomatic location of corroborating original record images
Digital camera
Manual assembly of family data. This is the weakest link in today's system
Use limited extracts to assist manual assembly of family data
Put compiled data into 'finished" online database. Contains low quality data and many duplicates. Ancestry, LDS Church, etc. Contains "Assembled Personal Pedigrees"
This process is almost ignored in today's genealogy industry
Manual assembly of family data
Published library book "Published pedigrees"
Digitize book
OCR and index book contents. "Digitized and indexed pedigree"
Manual reformatting of book to electronic genealogy format, with program assist, where possible
ProgenyLink online database
Contains finished, high quality, sourced data, no duplicates. Contains Sourced Single-Surname Descendent Structures, SSSDS
Other compiled data from family websites and private collections, electronic or on paper
ProgenyLink Database
Width of line indicates relative volumes of data

ProgenyLink extensions to the genealogy industry

In past centuries, people were determined to record their family genealogies, even in the face of the enormous difficulties of having to use only the original paper records as their sources. But many of them organized themselves and did the work anyway. When they had completed their labors, in these times before computers and the Internet, they used the only method available at the time to preserve and distribute that data to multiple family members, and that was by the physical printing of a book, multiple copies of which could then be distributed to family members and to libraries, making it an enduring effort. Typically, those books do not contain any of the source records or any of the citations to source records, simply because it was completely impossible to provide copies of the original records or even to provide detailed citations to those records.

For some reason, today's researchers tend to scorn these old books, these old paper records as not being very reliable. Certainly that is true in the sense that they do not have source records printed with them or even cited, in most cases. For some reason, these books have never been indexed in the same sense that census records have been indexed in recent years. Many of these books may have their own internal index, but that is of very limited use if one has to go through all the last pages of many of these books in search of some name. A consolidated index is what is really needed. But for some reason, those with the capital to have this transcription and indexing work done for published books have simply chosen not to do so.

So for researchers, who have become used to having everything indexed which they seek to use, it is a throwback to the old days of spinning through microfilm to have to look through these old paper books. There may be a little bit of technological snobbery going on here, but until those books are indexed, en masse, then they are likely to get very little use.

If these published books became the input to today's "online indexing" process, they could be transcribed and put into index form with a few years work. And perhaps that will happen later when all the easiest census records and other such records have been transcribed and indexed.

There is a certain excitement in being able to see those original historical records as they pop up on your screen at your request. Perhaps that is one reason why the published paper books are just not as interesting and exciting. If one is just looking for a handful of people, and the certainty of original records is also desired, then there is little reason to use these published books. However, if the goal is to complete all the genealogies for an entire nation, then these books become an extremely valuable compilation and condensation of vast amounts of historical records and research work.

One of the outputs of the ProgenyLink project would be, essentially, a transcription and indexing of this vast supply of published books. Those books with the highest quality information would be the first ones used, and others would be added later. In that sense, the ProgenyLink project might reasonably wait until the vast resources of the "online indexing" process could be brought to bear on those books. But, hopefully, it will prove to be economically viable to do this work on a for-profit basis.

I am highly confident that the ProgenyLink project will be successful, but even in the unlikely event that it fails for some reason, there is likely to be great good that comes from it because 1) it will introduce numerous ordinary church member genealogy researchers to a much better way to do their work, and create a demand that someone to continue to offer that facility, and 2) it will also provide great impetus for the central Church genealogy planners to start some of their "online indexing" volunteers working on reformatting and indexing the huge collection of published books.

One of the tests of the viability of the ProgenyLink concept is to see how well it is now possible to find appropriate source records and link them to the data found in these published books. If that can be done on a consistent basis, then these published books can become the genealogy treasures which their authors surely hoped they would be.

Luckily, the most difficult part of using these published books as a major source for genealogical data has already been accomplished. At least 150,000 books are digitized and available from books.familysearch.org. Google Books has digitized about 25 million books which have the term "descendents" or "descendants" in their titles. Surely there are many books of great value to the ProgenyLink process which need to be located and examined. Since such large numbers of these books can simply be downloaded to a researcher's computer and examined in that form, it is highly likely that many of these books can be used very effectively in the ProgenyLink process.

With these books in their digitized and OCRed form, the main steps remaining are simply to take the data which describes individuals and their family connections and rearrange it in the format required for entry into a genealogical database. Where the data in a book is high quality and the book is printed and digitized clearly, it is just a matter of copying and pasting the correct information into the electronic computer format. It will also be necessary to examine and proof-read the data to avoid any OCR errors. But at least it will not be necessary to rekey every character needed to be placed in the new electronic genealogical format.

With a little experience, it should be possible to use a computer program to take the digitized and OCRed books and rearrange the data within them semi-automatically. This should greatly speed up the reformatting process.

Glossary of terms

In order to discuss the ProgenyLink project easily and accurately it would be useful to establish some clear definitions. We need to be able to understand and distinguish between pedigree-sequence genealogy research and results, or family structures, and descendency-sequence genealogy research and results, or family structures. We need to know the difference between sourced and unsourced family structures, regardless of whether they are pedigrees or descendencies. And we need to know the difference between the "old raw data" and the "new raw data" that goes into assembling these family structures.

Family Structures

The term "family structure" is used as a generic way to refer to either a set of family names presented in pedigree sequence or a set of names presented in descendent sequence. In many cases, a researcher is happy to have either kind of information available to him, as long as he can pluck out the elements he is looking for. But just to make it clear, in the ProgenyLink project, almost every time the term "family structure" is used, it is intended to mean a descendent family structure, since that is at the heart of the entire ProgenyLink design. That past-to-present structuring method allows increases in clerical efficiency of at least 100 times and often much more.

Pedigree versus descendency genealogy research and results -- two basic kinds of family structures

I am going to have to invent a slightly new term here in order to help make sense of the ProgenyLink project. Almost everyone today does their research in pedigree sequence, that is, they start with themselves and go back to their parents, and then to their parents' parents, etc., and then they record the results of their research in pedigree sequence. I prefer to call the results of the research a pedigree, meaning it starts with them and goes back in time.

The ProgenyLink project is going to depend extensively on descendent-sequence research and the recording of the results of that research in descendent sequence. So here comes a new term: I want to call the result of that descendent-sequence research, recorded in descendent sequence, a "descendency." This makes it possible to easily distinguish between a "pedigree," which is a family structure which starts in the present and goes back in time, from a "descendency" which starts with an ancient ancestor and comes forward in time. There is a tendency for people to use the term "descendent pedigree," mixing the two concepts together. It is true that a "descendency" also includes some information about a person's pedigree, but mixing these terms together can quickly become confusing.

Sourced versus unsourced family structures

For hundreds of years it has been common and typical for people to record their genealogy information without bothering to give the library citation of their source records or to include a copy of the actual source record itself. Providing that corroborating evidence was extremely cumbersome and difficult to do in the past before computers.

The problem with that ancient method of presenting just the bare facts (conclusions, as genealogists like to say) with no corroborating public evidence, was that no one felt they could trust that data as being accurate, and, worst of all, there was no way for them to quickly verify whether that data was accurate or not. And re-doing the research from scratch was often out of the question and was prohibitively expensive. In today's world where we have almost instant access to the source records themselves, it is relatively easy to give the library citation or the Internet URL for that source document, and we may be able even to include a copy of that source document in our family history information. That makes it so someone else who wishes to use in our data can quickly verify the accuracy of the data presented. If there seem to be ambiguities, then the later user of this data can go back to the source documents and try to resolve that ambiguity himself.

The "old raw data" versus the "new raw data"

I sometimes use the term "raw data" to describe the input data to the "Manual assembly of family data" process in today's main-line genealogy industry activities. But today's "raw data" is not the same as yesterday's "raw data." Today's "raw data," which is the material used as the last step in the process before the manual assembly of family names is actually quite highly processed at this point. The rawest of data is the original public records, wherever they may be found, which are typically found on some paper storage medium in a library or archive or government or church office. A century ago, that was the only way to find out about one's ancestors. One had to actually travel to these record repositories to read them in their original form and piece together on paper your family history. But for everyone to do that much traveling was prohibitive and even when a researcher was in the archives or libraries, access to these materials was not easy or certain. There was naturally great concern about loss or damage to these records, and so the security arrangements may contribute to making it very difficult to use these records.

With the advent of the microfilm process in the 1930s, these widely scattered and fragile records could be preserved in a new form which could also be duplicated as many times as needed to allow people to have access to this data without needing to travel any further than to their local library. People could even buy their own inexpensive copies of these rolls of microfilm which contain about 1000 images each. Certainly, the research process was accelerated by at least 1000 times by removing the need for millions of people to crisscross the country and the world in search of the original records to examine and manually copy off whatever may be of use to them.

And in many cases, there was no index to these original paper records, so that researchers would find themselves having to look endlessly through the original books and papers in search of information of interest. Even after they were microfilmed and thus were saved from that potentially terrible wear and tear on the original records, people still had to scan through thousands of pages of microfilmed records searching for data of interest.

The LDS Church and other interested groups did begin an indexing process. People would scan through these rolls of microfilm and transcribe the data by keying it into a computer, and that could then be used to create a complete index to each of those microfilm rolls. The LDS Church had an incentive to do this kind of work simply so that it could have millions of names of real people that it could then process through its temples, so that everyone who chose to participate in a temple session would have a valid name for whose sake they could do their work.

The next major technological advance was to digitize these microfilm records so that people could sit in local libraries or even in their own homes, and view these genealogical records without the very significant time delays of having to wait for microfilm copies to be made and distributed for their use.

The next step was in adding an online process, now known as "online indexing," whereby people could transcribe these online records, very much like the original "records extraction" process, and online indexes could then be created to provide quick and direct access to these original records.

These highly processed records are now considered the new "raw data" for most practical genealogy research. Only in the rarest of cases would people consider traveling to a distant locality to review the actual original records. If it is done at all, it might be done as part of a family history tour, which might be as much a vacation trip as an attempt to make serious use of those original records for research purposes.

When someone subscribes to an online data provider such as Ancestry.com, they are looking for raw data for their research as part of their personal assembling of family names. And it is these highly processed records which then become the "raw data," the last step before the data is included in assembled family structures.

Single Surname Descendency Structure -- SSDS, or just "Surname group" for short
Just for the purpose of the Progeny Link invention and system, I have had to invent some completely new terms to describe what is going on. It is the basis for the Progeny Link database structure.

An SSDS is a family structure which begins with an ancient ancestor and comes forward in time, limiting its contents to those people who are descendents of that ancient ancestor, and who were born with his surname. That makes this structure only a very narrow slice of all the descendents of that particular person. All of the daughters and granddaughters who married into other families would adopt that new surname and all of their children would be born under that new surname. All of those children would be excluded from this SSDS. For example, Engelbert Huff was my grandfather back about 15 generations. An SSDS starting with him would include me, since my surname is Huff, but it would not include the children of any of his daughters or granddaughters, such as those of my sister Janene whose married name is Eller.

By restricting the names that go in each structure to a single surname, it makes it very easy to research, since the researcher is only looking for people with a single surname. It also reduces the workload to a level which most people can complete within a reasonable time. A very large SSDS such as that beginning with Englebert Huff has about 5000 names. That is a large number, but it is completely within the bounds of what can be done by one person or a small family group. In contrast, a person born in 1637, like Engelbert Huff, could now have approximately 1 billion descendents, assuming all children born to all daughters and granddaughters were included.

For use within the actual ProgenyLink software itself, we simply used the term "surname group" to denote the SSDS that any particular participant was supplying.

The next step is to look in more detail at the products and features which the Progeny Link project will add to this industry.

A Few Details of the ProgenyLink System

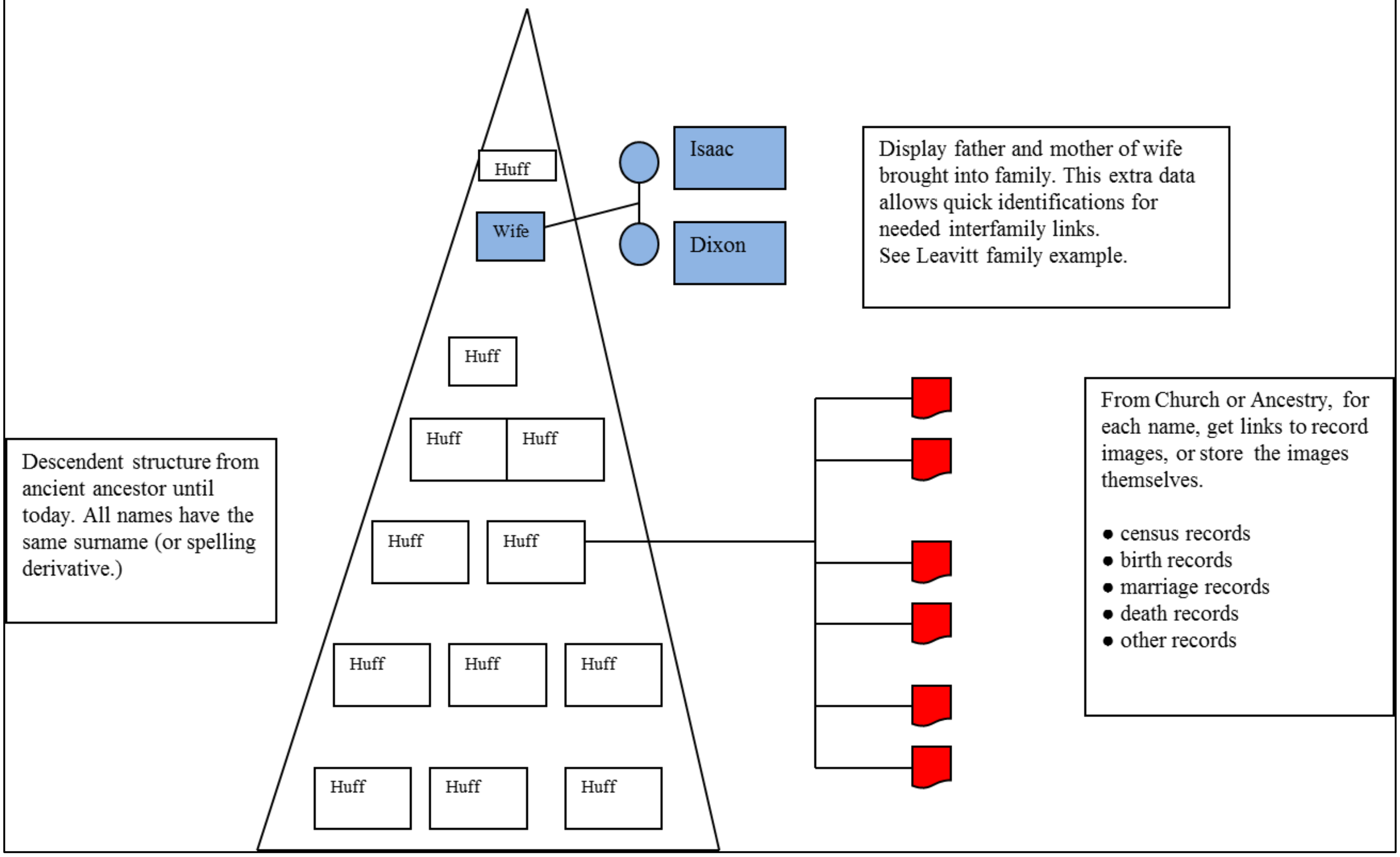
1. What are we building and selling?

The process I am proposing is very complex at the detail level, even though the overall concept is really very simple. Perhaps the first thing I need to present is a precise definition of what we are building and what we will sell. The following diagram might look like any number of strange things, including some kind of bug-eyed alien, or a levitating Egyptian pyramid, but it is one way to represent the basic unit of work and unit of data for sale in this new system. It shows a family structure beginning with an ancient ancestor and coming forward in time, in descendent sequence, being restricted to those who were born with the same surname as the ancient ancestor. This is what I call a same-surname family structure.

There are two kinds of appendages off to the side. One relates to women who marry into this family structure. They were born with a different surname and we need to at least show their parents as part of this structure so that we can easily make sure that the women that are imported into this family structure are also connected into their correct birth family. For the women who leave the same-surname structure to join another family structure, they will be treated the same way there.

The other kind of appendage shows the various kinds of source documents which tell us all the information about any particular individual -- birth, death, marriage, journals, land records, etc. Wherever it is feasible, we need at least one of those records to demonstrate that we have entered the information about that person correctly.

Assemble High Quality Names Using Single-Surname Descendent Structures As The Cooperation Mechanism



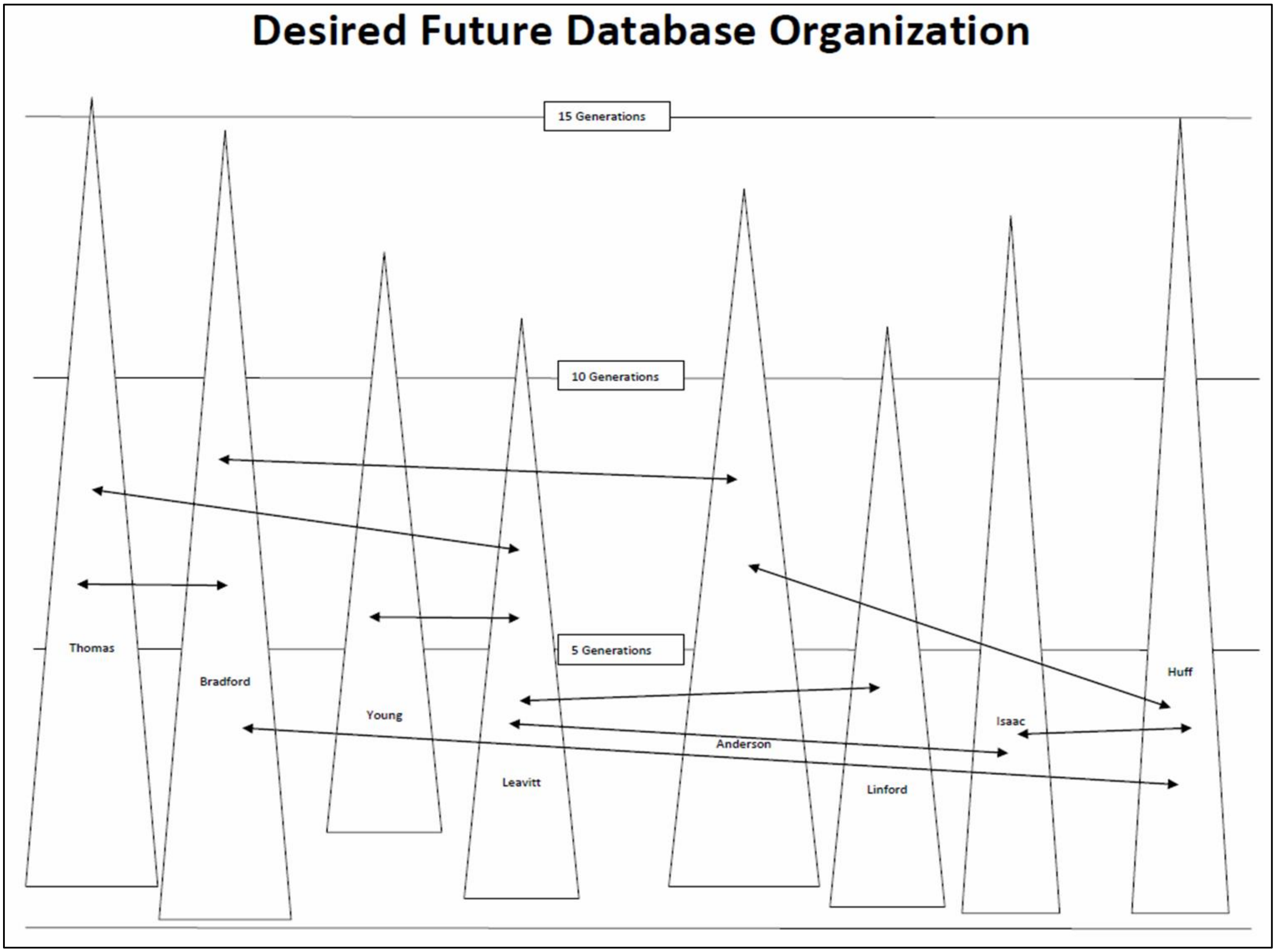
Display father and mother of wife brought into family. This extra data allows quick identifications for needed interfamily links.
See Leavitt family example.

Descendent structure from ancient ancestor until today. All names have the same surname (or spelling derivative.)

From Church or Ancestry, for each name, get links to record images, or store the images themselves.

- census records
- birth records
- marriage records
- death records
- other records

The next slide shows that data product in a larger context. The entire database will consist of these same-surname family pyramid structures, with an enormous number of marriage links, which are really same-person links, which link the imported women with their birth families. **With these links in place then it is possible to navigate the database in a pedigree or ascending sequence so that all possible pedigrees can be computed and displayed on demand.**



So that is a quick once-over of the basic data product, also seen in its larger context. We then need software which allows these structures to be built and interconnected and then sold, either one name at a time or one generation at a time, as the client prefers.

2. Why all the fuss about making such a radical change to traditional genealogy research?

Most people today, even those actively involved in genealogy research, have no conception of how catastrophically inefficient that traditional process is once you go back further in time than those who are still living -- perhaps about three generations. If one imagines and desires that an entire country can have their basic genealogy work done, using all of the public records, which go back about 15 generations, and in some cases further, then it would be extremely useful to figure out the mathematics of genealogy before you commit many billions of dollars to the process.

It is obvious that analysts working with the LDS Church genealogy database, the Ancestry.com database, etc., took no time whatsoever to study the larger process before they plunged into applying computers to the traditional paper-based methods. This "flying by the seat of their pants," and doing what seemed obvious, has actually made it so that it is completely, mathematically impossible to ever finish the task they set out to do, or even come close, using current methods and tools.

We might start out with the worst-case scenario and then compare it with what I am proposing. Let's say that everyone in our nation wanted to have their complete pedigree. That would mean 320 million people would each need to compile the names of their ancestors going back 15 generations, which would be 65,536 people. So 320,000,000 people times 65,536 names each = 20,971,520,000,000 or 20.9 trillion names. Considering all the typical time and cost inputs to research, $50 a name is actually a very reasonable price under current conditions. That means the total cost of the project would be $50 times 20.9 trillion names=$1,048,576,000,000,000 or $1,048 trillion or $1.048 quadrillion. That is 69.9 times the United States GDP. One might easily guess that it is mathematically and practically impossible to ever complete the project using that kind of brute force method. If we estimate that there are 170 million people who have died in the United States since it began, that means that, on average, we would have 123,361 duplicate entries for every person who actually lived and died in United States. In other words, if we have a database-building system which allowed for having each name only once or twice, with an average of 1.5, that new process would be 82,241 times more efficient than the old process.

In fact, if we could get each name in its proper place for $1 each, we would have a final cost of $170 million to build the entire database as opposed to $1.048 quadrillion. That makes the new process 6,168,094 times cheaper.

Obviously, families could work together to reduce the cost by a factor of five or six or even more, but the ending cost would still be astronomical. We should not be surprised that no one has done all of this nation's genealogy work on their own, or even their small part of it. They may not understand the mathematics, but as soon as they get into it a little ways, they will sense that what they're trying to do is impossible to accomplish in a single lifetime. Another concept and method is needed.

All of the current major systems, as operated by the LDS Church, Ancestry.com, etc., are using the concepts just discussed, which naturally lead to massive duplication, and there is nothing they can do about it without changing fundamental concepts. And many of their other operating techniques and rules make it even worse, as I will outline below.

Techniques have been invented to try to cut down on the duplication rates, without changing the underlying concepts, but they are fighting an inherently losing battle. The techniques devised so far are horribly complex, confusing, and expensive.

3. The problems and frustrations with the current Church system

The Church computerized genealogy system has been functioning since 1999, about 15 years. It has probably changed almost every week since that time, but there are several major aspects of that system which can be described as fairly enduring.

The methodologies used by other database operators such as Ancestry.com, MyHeritage.com, FindMyPast.com, etc., may vary a little bit from that used by the LDS Church, but not enough to justify describing them each separately.

Startup, including retaining the massive previous duplication

When the Church paper records were first digitized for inclusion in the first versions of the new electronic database, there were about 1.5 billion entries included from all the ordinances previously performed. However, all of those ordinance entries probably only represented the work on about 50 million people. In other words, the duplication rate was about an average of 30 times, at least in the years before 1999. There were some names which were in the database 10,000 times, meaning a man and his wife or wives and children would all be in there perhaps 10,000 times each. I believe it was one of the Pratt brothers who had the largest number of records. A special term was even invented, jokingly, for these situations. They were called ROUS's -- Records Of Unusual Size. The joke relates to the movie "Princess Bride" which featured a fight scene involving Rodents Of Unusual Size. When people looked at or downloaded such a name, they would get 10,000 copies of however many family names were in the bundle. This sometimes exceeded that ability of computers to accept such a mass of data, or to transfer it within a reasonable time.

The intent apparently was that Church members would spend time on the site and merge all of these names together so that there was really only one name left of the many slightly-varying copies. However, I don't think that merging process went very well. The basic problem is that unless someone had a really high-quality set of research, including source documents, for all those people, then they would be unable to merge them accurately, making the whole process nothing but often pointless guesswork. There was no way to tell which copy was the most accurate.

My thought was that if someone had that kind of high-quality data in their possession, they should just have the opportunity to enter it into a new clean database, perhaps using clues from the Church historical database, and then each name would be done right the first time, and all the confusing duplication would be gone. I think that after 15 years of ridiculous levels of effort, which might be quantified to be at least $22 billion in volunteer effort, the data quality has still not improved very much. In contrast, a single year of effort would have been enough to re-enter the entire database to create a clean copy.

Pedigree-sequence research and data entry versus descendant-sequence research and data entry

It may make a great deal of intuitive sense, at first, to begin doing one's genealogy research by proceeding in pedigree sequence, that is, going from you to your parents, and then to their parents, and so on. And that is probably true for the first two or three generations. However, as soon as one wishes to go back beyond the living to the dead, multiple generations back, all of those intuitively "obvious" methods quickly start to break down and become counterproductive. One of the major problems is the overwhelming duplication levels that occur when everyone works alone, without synchronization with the work of others.

Doing research and data entry in descendent sequence quickly solves all of those problems with massive duplication. For example, if I declare to the world that I am doing research on the Huff surname, then everyone can know that they should either avoid researching and entering data on the Huff line, or they should at least be sure that any Huff's they work on are part of a completely different family line, at least as far as the limited historical reach of the public records are concerned.

Communal access

It is true that it is a corporate responsibility as well as an individual responsibility to complete the temple work that the Gospel calls for. However, when it comes to actually handling data, carrying those communal attitudes and methods too far can be very counterproductive.

From a cooperative standpoint, it would be nice if all participants could see the entire database and add whatever data they had to the whole, but since this database was essentially open to all the Church members and to the entire world, there were many concerns about privacy, for every name in the database, but especially for the living.

Unexpected access limits

One of the earlier implementations contained a set of rules which were intended to maintain privacy in this otherwise very communal world of the Church genealogy database. A researcher was limited to looking at, and modifying, only his direct progenitors. What this meant in practice, for example, is that if a father had a child who then married, the father of that child could not see any of the information about that child's spouse or children -- his grandchildren -- since they were not part of his direct family or direct ancestors. This struck me as absurd at the time, since, as a parent, I probably had the most data to add to the database about my children and their families, and the greatest interest in making sure that it was done correctly. But the database was telling me that it was none of my business. Similarly, if I had cousins or in-laws about which I had data which I wanted to enter, I could neither see nor modify nor add to any of those names or families.

Editing wars

One of the major problems of a communal database can be discussed under the topic of "editing wars." These happen all the time on websites like Wikipedia were everyone can change whatever they like. If someone is unhappy with some webpage content, he can simply change it. But then someone else who has an interest in that webpage might notice the changes and change it back. This can go back and forth several times, perhaps until some arbitrator decides that they should simply lock the webpage since the editing wars are not getting anyone anywhere.

When this is someone's family that is being tossed around like a football, people can get very emotional. I attended one genealogy conference at BYU where a group of six or eight women from Arizona had come up to attend the conference. They accosted of one of the speakers, a Church genealogy program manager, and were in tears as they explained their vast frustration with the unending editing wars which they had experienced.

Lack of transparency

One might think that for a communal project in which essentially all the labor is volunteer, the process would be made very open so that people can measure their performance and progress, for example, including general quality. But, strangely enough, the Church genealogy computer people treat the statistics concerning the database and the research process as trade secrets, things they cannot reveal. One is usually justified in assuming that things that are hidden are things that are embarrassing in such a situation. If they can hide their poor performance from the Church volunteer populace, and especially from the Church leaders who supply the budgets needed to operate the computer facilities and library facilities, perhaps they can avoid responsibility for their continuing poor performance.

For example, one might expect that learning the number of new names which are completed to the necessary standards in the database each year would be very interesting and useful information, and would be encouraging to the volunteer people who do nearly all the work. However, that is a deep dark secret. My best guess is that the number of new, unduplicated names with adequate research quality is probably less than one million per year. I am guessing that the temples could process 10 million names a year, leaving perhaps a shortfall of 9 million in the acquisition of new names, in spite of the vast amount of effort put into this project.

Other technical processes related to genealogy are done much more efficiently. The number of new paper archive records which are photographed each year is probably in the 100 million range, and the number of those images which are transcribed and indexed could be as high as 300 million a year. The 1940 census, with its 140 million name entries, was finished in about four months, for example. So it is clear that although the raw materials are being produced at a very respectable rate, the final steps of putting these names into family structures at the proper level of quality is probably less than 1% of the names prepared by the other preliminary processes, and probably less than 10% of the number needed to keep the temples busy.

This brings up another embarrassing secret: because Church members have historically never been able to prepare more than about 5% of the names needed each year for temple work, that is, names that are fully researched and fully interconnected within their families, that means that as many as 95% of the names which are processed through the temples are either individuals who have no family connections identified, limiting the ordinances that can be performed for them, or they are people who have had their temple work done before.

There are many Church members who simply go to the genealogy centers and go through the lists of names for which work has been done, because that is nearly the only names they will find in these particular Church records, and they will copy down a handful, or even a few hundred of those names and just do them again "to make sure." That also means that the effort they need to put into original genealogical research is nearly zero. Once the database contains a large number of names, perhaps 50 million unique names as we have today, those names could be used over and over again, and no one at the temples would notice. This is a long-suppressed "secret" which apparently the Church technologists feel would be extremely damaging to their credibility if it was generally known.

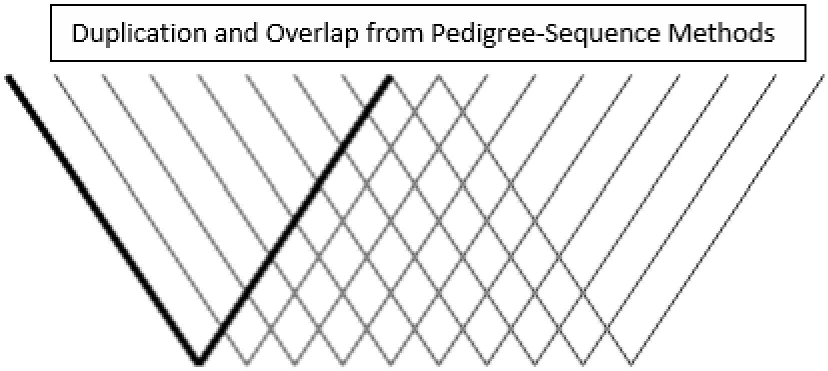
The Church technology people claim to have a program for encouraging professionals outside of their offices to invent solutions and implement them to further help the process along. However, the seriousness of that program is greatly in doubt. I once asked some questions to see if I could get some statistics about the content of the Church's main genealogy database so that I could check my assumptions against reality. I was not only told, No, that I could not have that information, but it was insinuated that I must have some evil intent for even asking. I don't know why someone would want to invest potentially years of work and millions of dollars in a process which they don't actually understand before they begin.

On the other hand, there are two or three organizations who have established mechanisms for accessing the Church genealogy database for use by individual people doing searches of various kinds. I wondered about getting involved in that program, but it appears to me that after a year or two of work trying to understand the complicated and constantly changing system they have, that it would never do me any good. The quality of the data is so bad, and the duplication so terrible, that I don't see any reason to invent extra ways to get access to it. The only thing I can see to do with that data is to examine it and take the best of it and put it in a separate database.

I suspect if someone asked the Church about the assumptions I have used here that they would deny everything and probably say that I was crazy. But I can also be pretty sure that they wouldn't give you the right answers either.

Endless duplication

Here is one simple way to look at the most obvious form of duplication: As each researcher creates his own pedigree, there will be an inevitable overlap with other researchers as he goes back in time, and this overlap increases until there can be many thousands of ancestral names which he shares with millions of other researchers.



Duplication and Overlap from Pedigree-Sequence Methods

Each of those ancestors might then be entered thousands of times by thousands of different researchers, and this is a deplorable waste of time and energy, which can be avoided through cooperation and coordination.

To show the astronomical levels of duplication that are possible, at 15 generations back, a person would share 32,768 ancestors, with about 1,073,741,824 (1 billion) descendants for each ancestor, or about 35,184,372,088,832 people in all. At potentially 35 trillion people who share the same ancestors as yourself, you would obviously be related to every person on the planet, and if any significant number of them were also active researchers, you would be working with and duplicating all the same data they are.

Actually, there are so many kinds and layers of duplication of effort and results found in the current Church genealogy database system that it is hard to even identify and describe them all. As described above, it is clear that it is mathematically impossible for members of the Church to ever finish the genealogy for the United States because of the multiple kinds of overwhelming levels of duplication which would need to be done to finish the project using current concepts. Of course, all these kinds of duplication were done in the past when all the work was done on paper, and many people optimistically hoped that these kinds and levels of duplication would end when the whole process was placed on computers. However, since the underlying problems were never understood and solved as part of the process of putting this material on computers, there are still massive levels of duplication going on, and even more massive levels of duplication which would be required to finish the project using current concepts. In fact, computerizing this chaotic mess probably means that the total amount of duplication has actually increased since computers were applied across the board.

No active way to plan and cooperate

As a practical matter, the only way people can cooperate, or plan any kind of future cooperation when using the Church genealogy system, is to do their own research on their pedigree until they happen by chance to run into someone else who has researched one of the same names. If that other person is still actively involved, and has provided the means to make contact, and chooses to respond to an e-mail request, then some limited cooperation might be planned on an ad hoc basis. Otherwise, everyone is left to work alone. What this really means is that people can do years of research by themselves and only later find out that someone else, or perhaps several other people, have already done that research. In other words, many people can waste many years of effort and never know it until after they have done that work and noticed each other at the end by happenstance.

I know someone who wrote an entire book about a family, and only after publishing the book did he find out that one of his cousins had done essentially the same work and was also publishing a book. Most of the work of at least one of these people was mostly wasted, and it was an unpleasant shock to them to realize what they had duplicated.

Intentional massive duplication as a business model

In the case of the Ancestry.com system, their "leaf" or hint system is a very clever and interesting feature, which draws people into their system and helps them achieve their personal goals, but what it does NOT do is encourage their 2 million subscribers to cooperate with each other to solve problems they might have in common. The most lucrative business model for all of these kinds of companies is to encourage people to work alone and to engage in the largest possible amount of duplicate effort, so that no one can ever finish any large amount of genealogy research or share any high quality results.

If nothing else changes, that ensures the business a continual source of subscribers indefinitely, perhaps at least 30 to 100 years. One of the goals of the new system is to stop that intentional infinite duplication and actually get some finished results.

I had hoped that the LDS Church, with its unique assignment to complete the genealogy for whole nations, would want to spearhead a change in concept and operation. However, it appears that instead of the Church controlling the genealogy thought processes, Ancestry.com is the actual thought leader. Many of the people who now work for the LDS Church were previously employees of Ancestry.com. Apparently, the LDS Church is unwilling to consider any change which might damage the current business model of Ancestry.

4. The solutions to all these problems

All the problems I have described above have been solved in the new system. I am guessing that it is far more important to describe the problems which I have solved, rather than describe the actual ways I have solved them. Certainly, if you don't know the problems, you cannot appreciate the power of the solutions. I will spend a few minutes on the solutions, but that is clearly not my focus in this paper. I have other documents, including two patents, which describe the solution in great detail, but contain only a small bit information about the problems to be solved. I assume it is much more important for you to know which problems I have solved rather than how I solved them, although we can go into that as well, if you have the time.

In my book *Doing Genealogy the Henry Ford Way*, I summarize the problems of the genealogy industry under 6 different issues: Cooperation, Integration, Uniqueness, Duplication, Quality, and Fairness. The book is one of the documents which can be accessed at www.ProgenyLink.com. I will use a slightly different set of topics to present the same ideas here much more briefly and simply. What follows are two of the most important changes.

Individual ownership of data instead of communal access

To answer the problems of the current communal access technique, the ProgenyLink system will provide individual ownership of data, and owner control of access to that data. This means that someone who owns valuable data can become its "owner" and control what data is presented to whom. No one can change that data, or even view it, without permission of the owner.

That owner can invite other collaborators to view and comment, or to actually change or add data, if the owner chooses to provide those access rights. This completely ends the "editing wars" and assigns personal responsibility for that family data. If others wish to offer another view or opinion of that data, they can place their version of the data in their own controlled space and simply indicate through links that this alternate opinion is available for viewing. If the data owner is willing to offer their data to other users, he can indicate that choice, and the ProgenyLink software will only offer that data for viewing under agreed-upon terms, usually including sale.

Preventing and hiding duplicates

There are several ways for people to plan their work together so that they can cooperate and collaborate efficiently. For example, someone can indicate their intent to enter the data for a same-surname descendent structure, so that anyone else with an interest in that surname can either to choose to join in and work with that other person, or at least refrain from filling the database with duplicate data. For example, someone could notify everyone else of their intention to enter all the descendants of Englebert Huff who were born with the surname of Huff. This obviously should prevent most duplicates in the first place. Any researcher can periodically check to see if someone is inadvertently duplicating their work.

In cases where duplicate work has been entered without all the proper checking beforehand, then that duplicate data can simply be hidden from everyone else so that it will not be a continuing source of confusion.

5. Other more general issues

Will Church members spend up to $15 million a year on our services? Do we want them to?

LDS Church volunteers are currently spending about 150 million hours a year on genealogy projects. That is about 300,000 participants, each spending an average of about 500 hours a year, we might guess, as volunteers or as called missionaries. About one half of that time is spent doing "online indexing" which involves translating the text of genealogy images so that indexing can be done. The other half of the time, about 75 million hours a year, is mostly used in assembling names into family structures. If we imputed a wage of $10 per hour for that second half of volunteer activity, it would be worth $750 million in volunteer labor each year. Since this is truly volunteer labor, and not missionary labor, members might be interested in considering more efficient and less time-consuming ways to do that work. If Church members were willing to translate just 2% of this labor into paying money for someone else's labor and product, which had been done a great deal more efficiently than anything they could do, that alone would give us the $15 million to make this project very successful. There are probably between 13 times that many nonmember genealogists (4 million) and 40 times that many nonmember genealogists (12 million) who would be willing to consider using our more efficient service, again trading their money to save them time by purchasing our product. Manufacturers or researchers often have the option to "make or buy," and we want to make it so that many people see that "buying" is the more efficient and higher quality option.

Someone made the suggestion that this new ProgenyLink database ought to be made available to Church members for free. There is some plausibility for doing that since it is mostly the Church organization and the Church member volunteers who have accumulated this vast amount of raw genealogical data in the form of microfilm images and now digital images, and indexed it and pieced some of it together into family structures. That is the material that we will be harvesting as part of this project, and that is why it can become profitable at this point.

However, I think there is a better way to look at it than offering more "free" stuff. In summary, even though we are seeking a profit through this work, we will be saving individual Church members, and the Church organization in general, enormous amounts of time, effort, and money by making the system available. It is a great deal cheaper for members to use this system to solve the problem quickly rather than to go on forever with the wasteful path of doing all this work and getting hardly anything for it. It is cheaper to finish the project this new way than to go on forever in the extremely expensive and wasteful "church way."

In other words, Church members will get a huge amount of benefit from this new system, and there is no reason to jeopardize the success of the project by being overly anxious to give the results for free to Church members. It is true that there are somewhere between 13 and 40 times more non-Church members who could benefit from using the system and who would add money to the project, but I'm not sure it is a good idea to have these people subsidizing Church member genealogists based on what has been done in past years by the Church and its members.

With the new system, Church members can do nearly all the work to get this project under way, and they can receive the income that will come to them as their work is sold to both members and nonmembers. Matching levels of work with levels of income is a powerful way to introduce overall fairness into the process. At the same time, Church members will themselves gain all the genealogical data which they seek so that they can actually finish these otherwise infinitely long projects which can never be completed using current methods. I'm going to say that that is such a gigantic benefit to members, that I would not want to risk the success of the system by trying to make special allowances for members.

For example, what would happen if one third of the work of Church volunteers, outside of online indexing, was made unnecessary because of this new methodology? That would mean that $250 million in effort could be saved every year, and this could go on forever. Let's say we just take the effort that is saved over a four-year period, which would be $1 billion. It could eventually be a great deal more than that, but a $1 billion savings to Church members seems like a nice enough gift, without going even further and potentially compromising the data marketing system of the ProgenyLink project.

Protecting the data and the proprietary interest in it

One of the obvious risks of putting this high-quality genealogy database together, is that, if it is not done properly, there are many thousands of people who would happily steal all the high-quality data and mash it together and sell it to someone else, so that they make the profit instead of those who put the data together. Partly this means that the first time a name is sold, it should bring in enough money so that we don't care excessively if we ever sell that name again.

At the same time, we need to make sure that it is really extremely difficult for someone to crack through the various barriers and controls and steal data without paying for it. I calculate that, on average, the names in this database will be worth about $40 a piece. That seems like that is certainly enough value to make these names the object of theft.

Even for those using the system, who are putting in their work, if they use someone else's work, they should be required to pay for it. Mere membership within the system is not enough to give them free access to all other data. So then we provide a kind of internal money or "funny money" situation, where people can put in high-quality data and sell it to other people who are also either members of the operating system or who are outsiders. And those who sell data can be given credit for those sales and they can use those sales to buy data from other people.

That keeps all the transactions under one roof and should provide a fair way for people to be engaged in the process but still pay and receive their fair share. If Church members are doing all the work, and there are from 13 to 40 times as many people outside the Church who are buying their data, then they can easily get paid fairly for the work they have done, and there's no reason to differentiate between members and nonmembers within the system. People can do work for which they get paid and then they can spend that internal money to get data from other people who have put in quality data.

There are already many complaints in the genealogy industry about free riders who take the work of those who have gone to great trouble to assemble family structures. Often they claim that data as their own or even make changes to it which can only lead to the disgust of the people who created the data in the first place. We need to make sure that this does not happen in the new system, and that will be part of the security features.

These kinds of unethical behavior are what keep many people out of any of the public genealogy systems. They could contribute some valuable data, but if they do so they may suffer embarrassment and loss rather than be rewarded in any way for that work. So we need to make sure that there is a way for these people to receive that appropriate award in money and information, so that they will be happy to join in the project actively.

Viewing the entire ProgenyLInk process

The following chart presents the next level of detail down on describing the complexities of the ProgenyLink project, as described in the maximum detail in the patent. It would obviously be a several-day project to go through and explain every technological aspect of the invention. The main point to realize is that this process of reengineering the entire genealogy industry requires multiple technology breakthroughs, all at once. None of them are particularly difficult to do, but they all need to be done at once or else the overall goal cannot be accomplished.

Sequential Process View of Invention – Reengineering the Entire Genealogy Industry

This requires multiple technology breakthroughs, all at once

Theoretical efficiency improvements of up to 2000 times are possible, as in any typical mass production/industrialization process, in contrast to the inefficiency of typical cottage industry methods. Actual improvement in the range of 30 to 100 times should be easy to achieve. (Mass production techniques have never been applied before to the crucial genealogy process of name assembly. Adam Smith, *Wealth of Nations*, put 4800 as the top industrial improvement multiplier observed.)

Register/setup	Enter Data	Assign numbers	Store names in descendent-sequence structures	Data owners control access and updates to data.	Continual data improvement process to reach required quality	Connect Surname Groups through Women	Data Quality Filter/Barrier	Pay-Per-View Data	Sale of Names/Final State of Database
• (1) Register (identify) user, register surname and associated ancient ancestor, allocate workspace, pay membership dues. • (2) Establish multiple workspaces for ambitious users or workgroups, each workspace used to store a different surname group. • (3) Users receive royalties to offset membership dues and may receive net positive cash flow.	• (4) Direct entry from prior manual research. • (5) Bulk GEDCOM input • (6) Employ specialized semi-automated assembly of all index entries, names, and related documents of potential interest for single surname. (Only possible using descendent-sequence system). • (7) Use "process of elimination" separate subsidiary database to show which public record images have been used in the main database. • (8) Broker and coordinate outside research.	• (9) Assign worldwide unique number for all dead and living, as data is entered. Turn the Internet into one integrated genealogy database with a unique ID for each possible person. See 4-level number, which identifies data owner, and keeps descendency groups separately addressable throughout the database.	• (10) Store names in descendent-sequence surname groups. – avoids nearly all duplication (data owners enter data for only one surname -- "Descendents of….") and notify other participants of each user's research intentions – allows industrial-strength cooperation across surname lines.	• (11) Specialized workgroup networking features for genealogists including multi-level access rights. • (12) Provide special provisional update methods. • (13) provide special temporary workspace or "shadow database" transition and transformation processing space to support numerous special transactions. • (14) Data ownership is recorded at the name and data element level. (optional) • (15) Provide "Everyone can update" feature for "community data" projects.	• (16) Improve data in normal ways. • (17) Semi-automatic mechanism to find source records to link to previously assembled names, using . "screen scraping" and other techniques • (18) Unique document and image upload process by individuals. • (19) Use public catalogs as input to source-identifying entries. • (20) Assess current data quality levels as needed. • (21) Special transactions, part of improvement process. • (22) Link names to source records which are uploaded to GenReg. • (23) Link names to source records on major sites with stable URLs. • (24) Link names to various websites cross-indexed to GenReg, containing videos and other voluminous data.	• (25) Provide internal email system • (26) Process for connecting surname groups must use high-quality data to avoid confusion and wasted effort from unstable data. – users receive 1000-to-1 return on data entered and connected as other users provide 10 generations of data on 1024 surname lines.	• (27) Categorize data by quality, and search and list results by quality category: Up to six categories of quality. High, medium, low individual quality, plus size of network of related names.	• (28) Confirm data has reached Pay-per-view quality levels, the final step in the quality improvement and recognition process. Diligent users receive: – 1000-to-1 data return on data entered. – Eligible to receive royalties on marketed data.	• (29) Sale of finished data. Internal financial system tracks all user-related transactions. • (30) Royalties go back to data supplier – adds big incentives to finish whole nations by filling in all the data gaps. Users pay for data they download, and receive payments for data others buy. System pays net royalties periodically. • (31) Collect and remit payments for "on consignment" data on other sites (see item 23). • (32) Record stripping -- The final state, an Historical "Facebook" for all historical people. • (33) History-based social networking. • (34) A more accurate method for indexing source records. (30, 31) Royalties go back to data supplier (profit-sharing)

(30, 31)

(3)

Extended Explanations

(6) Workstation software and hardware uses sophisticated "screen scraping" and other techniques to find and assemble all index entries and names from numerous online "raw data" databases and related source documents for a specific surname.

This process greatly accelerates the manual review and assembly of family structures at local PC document retrieval speeds which may be up to 400 times faster than unpredictable Internet speeds. Users can view dozens of documents simultaneously, while making comparisons among them, where useful.

This process can only be done using descendent-sequence (single-surname) method, and puts results in database in descendent sequence. (There is no practical way to do this with pedigree-sequence research, since new, usually unknown, surnames are introduced at every step backwards in time. For example, there are 1024 surnames needed at 10 generations back in time.)

--See conceptually related process under "continual data improvement," item (16), where source records are semi-automatically located for previously assembled name structures.

(9) Using a unique "tree-level" number, also allows entry of lists of names which are not connected into family groups. These names can later be assembled into family groups using the "descendency-level" number. This might include such things as the lists of Russian prisoners sent to death camps, where Russian genealogists have preserved those names, but have not yet included them in pedigree structures.

(11) Specialized workgroup networking features for genealogists including multi-level access rights.
Access rights of View, Provisional Update, and Immediate Update are granted to family and friends and provide numerous workgroup networking features, specifically for genealogists.
Same-surname cousins (who all have a common ancestor) should be the first group to invite to assist.
(12) Provide special provisional update methods to assure shared responsibility and control to achieve high quality. Includes option to review of all updates after-the-fact, by date and operator.
(13) provide special "shadow database" transition and transformation processing space to support numerous special transactions, including various provisional update transactions, as in item 12 above, item 15 below, item 21 below, etc.
(14) Data ownership and responsibility is recorded at the name level and at the individual data element level.
(15) Unique, carefully controlled "Everyone can update" feature for international "community data" projects.

(17) Semi-automatic and manual location of sources – link names to online sources or to sources uploaded to Genealogy Registry. This is similar to process used in item (6).
(18) Participants can upload personally acquired documents. (This should end most institutional, contractual, archival, media, and structural barriers to location and use of the world's genealogy source documents.) Scanners, cameras, downloads from various scattered sites, etc., may all be inputs to this process.
(19) Use public catalogs as input to source-identifying entries.
(20) Assess current data quality levels as needed
(21) Special transactions, part of improvement process. Collect data fragments from throughout database into one work space. Connect hierarchical workspaces with each other.

(26) Women usually appear twice in the database. A woman in the role of a daughter is connected to that same woman in the role of a wife in a separate surname group using "same person" links. This ties all portions of the database together so that all possible pedigrees can be read out at the end of the database construction process. Databases are constructed in descendent sequence, simply because that process can be hundreds of times faster than the traditional pedigree-sequence methods.

Some general benefits and consequences.

- Cooperating across surname lines is the most powerful benefit of all. (26)
- (35) Engaging genealogists worldwide will maximize the cooperation and achievable gains. General enthusiasm from expecting quick completion.
- Turn the Internet into one integrated genealogy database. (9)
- Choose which data should be visible to public searches. Avoid most confusion. (27)
- Semi-auto assembly of family structures. (6)
- (36) Encourage organizing family organizations, usually of same-surname cousins.
- Centralizing the indexing and marketing of data on 3rd-party websites. (24)
- "Records stripping" -- a nationwide correlation of all historical records for individuals, creating an "individual level" national history. (32)
- Acquiring all the world's genealogically significant source records images through individuals. Improve legal access to more genealogical data. (18)
- Researching in descendent sequence is very efficient. Avoids duplication of research and duplication of names in database. (4)
- (37) Reclaiming the 20% to 25% of people lost to pedigree-sequence research.
- End of email. Users rarely need email any more, and if they do, it is internal. (25)
- (38) Database entries can be reverified hundreds of times to improve accuracy.
- Financial subsystem -- gives revenue, increases fairness, adds new incentives to finish nation. (30, 31)
- (39) Achieve industry business integration for another layer of efficiency.
- Workstation for semi-automatic finding of source documentation. (17)
- (40) End most "brick walls" in genealogy research. No ad hoc "reverse gen."
- (41) Solve problems for LDS Church, and they put more money into the project.
- (42) Solves all industry technical problems at one time: duplication, cooperation, integration, fairness, quality, uniqueness -- end weak incremental improvements.