20120906 Prodigy Link feature listV07-trim

Progeny Link System Features

The Big Picture

The Progeny Link system is intended to solve every important problem which the genealogy industry has today, which will lead to a high quality nationwide genealogy database containing no duplications. It is now feasible to reach these goals because the new features we are adding to the industry toolbox allow serious genealogists to assemble high-quality names at up to 1000 times faster than is typical using traditional methods.

As a simple illustration of what the "1000 times" factor means, if all the serious genealogists in United States were working in concert using the same set of rules and the same central database, the basic genealogy for the United States could be completed in two weeks time. (Calculation: 4 million genealogists each put in 18 high-quality names which would total the 70 million names of those who died in United States before 1930. If they spent about four hours on each name, that would mean they spent about 80 hours of work.)

I assume most readers will have trouble absorbing these broad new claims to begin with, but I invite you to go through the list of issues and problems we have found with the current genealogy industry, look at the features which this computer system offers to deal with those problems, and make your own judgment about which problems have been adequately solved and what is left over. We try to stay aware of any problems ourselves, but we welcome anyone pointing out new features that we might implement to help the genealogy community have a more efficient, successful, and satisfying experience.

Our goal is to offer some enormously powerful new features to the genealogy industry, but the individual needs and wants of millions of genealogists will probably cover many more things than we can ever understand and offer ourselves, so we plan to partner with other companies who have interesting solutions to some of the "smaller" but still very important problems in the industry. Some of these will be noted throughout our feature list.

The Industry Problem List

Some of the big issues addressed and solved here are these: Efficiency, Duplication, Quality, Cooperation, Uniqueness and Integration, Fairness. For more information, see "Doing Genealogy the Henry Ford Way," an 80-page document discussing these general concepts in detail.

Efficiency – A change in procedures can take advantage of the previously unexplored "mathematics of genealogy" that allow us to improve efficiencies first 30 times and then 1000 times through multiple kinds of cooperation.

Duplication – Duplication levels are so high in today's genealogy industry that they nearly paralyze any net forward motion by the industry.

Quality – Low quality, unverifiable, and untrustworthy data prevents serious cooperation and adds greatly to duplication of effort as every researcher finds it necessary to do the research again.

Cooperation – Almost no serious cooperation is possible among researchers today, causing staggering duplication rates as seen in the LDS Church ordinance database where 200 duplicate entries for a single name are not unusual.

Uniqueness and integration – It is important to identify each unique individual and then link all known public data to that person. For example, that will stop multiple entries in perhaps six different censuses being treated as six different people. And integrating all that data together will also allow the maximum information to be known about each unique person.

Fairness – Some of the most dedicated and productive researchers are often treated badly by the industry. They are often not given the recognition and rewards of they deserve, and that may cause them to keep their data to themselves rather than share it with the world.

Website Features – The Basics

Data content and sequence – You can put in any genealogical data in any sequence you want to, but we encourage people to assemble high-quality work in descendent sequence (top-down, oldest ancestors to youngest, most of them having the same surname). The system offers many special features that support that method and lead to higher efficiencies.

For example, achieving the 1000-times increase in productivity for large groups of researchers is dependent upon using the descendent-sequence entry method. Another powerful feature of using the descendent-sequence data entry method, is that it ensures that no names are left behind. In typical pedigree-sequence research work, about 20% or 25% of historical people are lost. They never make it into the genealogy books. Children who died young or didn't marry and have children, obviously will have no posterity who would be seeking them out through genealogy research. This "missing persons" problem is solved when genealogy research is done in descendent sequence, following the actual path of how families were formed.

Members have complete control of their data
Each member user has their own data space, and no other member may make any changes without access rights granted by the data owner. This gives the data owner complete control of what occurs in that data space. Other users can enter material into their own data spaces and link it to a particular user's space, but no actual changes are made to anyone else's data. This makes each user completely responsible for any problems, and makes him the one who is rewarded by any benefits.

As a point of contrast, we should note that the LDS Church, in their FamilySearch program, offers a *communal* database situation. Although there are a few constraints, one can generally say that anyone can change anything in that database, meaning that there is a high level of instability. People can get into "editing wars" where they change one set of data one way, and someone else comes back later and changes it back to where it was. People can make these changes without offering corroborating proof documents. This system was originally intended to help with checking on the performance of temple ordinances, and adding new names to receive temple ordinances. It does that job well, but is not suitable to be morphed into a worldwide generic genealogy recording system. Many people have asked for the feature of being able to add documents to verify materials, but that has not been made available.

One of the consequences of merging all of the ordinance databases for the LDS Church, plus other membership related information, means that there are enormous numbers of duplicate names in the resulting database. This excessive duplication can easily be tolerated when carrying out temple name processes, but it becomes intolerably confusing when trying to use that same database system as a generic genealogy system.

Searches – If users meet certain work contribution criteria, then they have the option to search the work

of all other member users. One purpose would be simply to find if there are others working on similar research projects. If they are, then it would be important for the two groups to find a way to cooperate or to at least ensure that there is no significant amount of duplicate research occurring.

Search results are categorized by quality level, with the highest quality appearing first. There are four quality levels, with the fourth one being low-quality, and it is not even shown unless requested. In this way, the system tries to protect the user from the garbage data which is so much a part of the Internet world.

High Quality -- We encourage and assist researchers striving for high quality data to add to our database all useful source document images and links to source documents. Most genealogy databases are full of low-quality data which means almost nothing can be trusted and most researchers feel that they must reverify all the data that they use. This causes enormous amounts of duplicate research and frustration.

We encourage users to construct and use only the highest quality data to avoid confusion and duplicate research. Your searches of everyone else's data in our database will bring back that data in quality sequence, with the highest level of quality and completeness first. If you are linking to other people's data, you obviously only want to use their best data. Linking to low quality data or data that is in the process of construction may just be a confusing waste of time for you and everyone else.

As an example of a simple quality enforcement mechanism, if a user has not entered at least 50 to 200 high-quality names into his data space, and made them accessible to other members, he will not be allowed to access the other names in the database. That should at least make the point that we are serious about quality

Safety – Users can treat this system like a safe deposit box for their data. Our data is stored using the Amazon triple redundancy storage system, or its equivalent or better.

Transferability -- We will offer the ability to transfer your data among perhaps 20 different personal and Internet genealogy data storage systems, such as PAF, Legacy, Roots Magic, Geni.com, Ancestry.com, familysearch.org, etc. We will offer some of these services directly ourselves, such as PAF to ProgenyLink, and we may partner with other companies for a more complete offering.

Outputs – All the typical genealogy reports will be available in at least one format – pedigrees, descendencies, family groups, detailed individual presentations, etc.

Data downloads will be available in common formats for such projects as creating charts, graphs, books, etc.

Special outputs for LDS Church members – LDS genealogists are a very important user group, because of their special enthusiasm and commitment to genealogy for religious reasons, but they probably represent only about 2% of the genealogists in United States. We will want to add special reports and data files that will assist LDS members in submitting names to their temples, for example. But since our basic reports should suffice for most LDS purposes, specialized LDS reports may not be available at the beginning.

Cooperation features

Reaching the goal of making genealogy researchers up to 1000 times more productive can only be accomplished through "industrial strength" cooperation. There are many different cooperation features which work together to accomplish this huge boost in productivity.

Surname registry

Use this facility to let everyone in the world know what you are doing or planning to do, and check to see

what they are doing. This is the first feature and the first step to getting everyone in the genealogy world synchronized so they can coordinate, cooperate, and collaborate efficiently. Users can easily communicate with each other through the internal messaging features of the ProdigyLink system. They can request View or Update access to other people's data, where appropriate, and allow the same kind of access to their own data, if they wish.

Many methods have been used in the past to notify the world of a researcher's areas of interest and expertise. For the purposes of this system, giving the name and other identifying details of the ancient ancestor which sits at the top of the surname group that a particular researcher or researcher group is constructing is a shorthand way of designating a large group of people quite succinctly. For example, Engelbert Huff, born in 1637, is the ancestor of about 5000 Huff-surname descendents. Providing his name is a simple way to refer to that well-defined group of people.

Allowing access rights to other users – trusted friend and family workgroups
Each member has his own data space, and completely controls who has access rights to view that data, update it immediately, or update it provisionally. All updates can be reviewed after-the-fact in date sequence, and for data submitted by those with "provisional update" authority, the data does not enter the main tables until the data owner has specifically approved it.

This feature is ideal for organizing multiple cousins with the same surnames and with common ancestors so that those cousins can work together to complete the descendent structure of names for a particular ancient ancestor. Several anecdotal studies indicate that simply by collecting data for a single surname descendent group, the number of names actually collected can be 30 times as great as the typical genealogist working alone in pedigree sequence. This is one important element of the overall efficiency goal.

Cooperation between workgroups/surname groups
This is where the biggest single productivity boost can be found. If users will enter their single-surname data to a high quality standard, they can then have access to the 1023 other surname groups which are needed to complete their full 10-generation pedigree identifying 2048 people.

After the single-surname groups are constructed, then the marriage connections are identified between these surname groups, connecting a woman where she appears as a daughter to that same woman where she appears as a wife. At that point the database is complete, and it can be read in pedigree sequence just as easily as it was previously read in descendent sequence. Since most people are curious about their pedigree, this procedure provides the previously unnoticed benefit of allowing someone to specialize in entering data for a particular surname in descendent sequence and then reading out data of interest to them in pedigree sequence.

Since these are different groups of workers which are expected to work together for everyone's advantage, there need to be some rules regarding quality and fairness. These are discussed throughout this list.

Semi-automatic use of online genealogy indexes
To improve data quality and speed up data aggregation, a specialized browser program can access the indexes of the various online offerings of "raw" unprocessed public source documents and pull off all index entries and digitized images for a particular surname. This collects into one place all the data that is available online so that the researcher can piece those materials together into completed families. For example, one might request that all of the historical data concerning the Huff family be located and placed on his PC for further processing.

It can also be used in the reverse mode where the researcher presents the program with a large set of names, and the program can locate all online materials related to those names. This version is especially useful where one has taken a large set of names in descendent sequence from a publication, and merely wishes to find corroborating source documents with a minimum amount of work

Accessing the Genealogy Library card catalog

When seeking out corroborating source documents to verify genealogy data, it is very helpful to have easy access to the Church's Genealogy Library catalog. I believe it can be said that only some things have an Internet URL, but *every* source record that the Church holds has a card catalog record. For all things which are not in digitized form, indexed, and online, this is an easy way of providing a standardized reference to those materials.

Synchronizing ProgenyLink data with FamilySearch data

Because of the generally low quality and high duplication rates of names found in the FamilySearch system, plus other confusing factors, it is questionable how much value there might be in offering a synchronization tool of this type. Misuse of such a tool could easily add large amounts of low-quality or even useless and confusing data to the ProgenyLink database which is intended to contain only high-quality data. After careful study, this feature might be added, but this might be another case where we would partner with someone who has already solved the problem and use their software.

A specialized numbering system – linking separate websites

The internal numbering system helps accomplish several valuable functions. It assigns a unique number to every name which is entered into the system. That number identifies the owner of the name, that is the person who entered the data and has assumed responsibility for its accuracy, and it also identifies the particular family surname group to which that name belongs. There are cases where it is useful to be able to address an entire family group rather than just an individual in that family group.

This number can be used to prevent duplicates and assist in hiding low-quality data where that would be useful.

Assigning a unique number to each name means that any data which may reside anywhere on the Internet concerning that name can be linked together. That makes it unnecessary to store all data concerning one name in one location. This offers an efficiency factor when families may have multimedia sites which can be referenced by the central database without actually transferring all those voluminous materials to that central Internet site. In other words, this unique numbering system can allow the entire Internet to be treated as a giant integrated genealogy database. The database can be stored in multiple segments, if that proves useful, without difficulty.

Special features for community databases

There are numerous projects in which people work together to construct a "community database." That refers to the process of choosing a particular community or geographical area and finding all historical records which are available for that community or area and linking them together to identify and link the historical families. I believe one of the famous test cases occurred using a county in Norway.

To work efficiently, such a project needs a unique set of rules for those who participate. Since it is impossible at the beginning to know where all the descendents of such a community might live in the world, and therefore who the interested participants might be, it is necessary to make it a "communal" project in the sense of allowing anyone to participate who wishes. However, certain controls are still necessary for this "communal" project lest some uninformed or malicious person do some damage. These controls are similar to those discussed above which allow View or Immediate Update or Provisional Update access rights. If the default is Provisional Update, then everyone can enter data, but only the person who is the administrator for the project can allow that data into the main data files permanently.

I believe the current technology in this area requires that all updates be done off-line by an administrative group, which then periodically rebuilds the static website database so that participants can view the status after the most recent update. This means we have a variation of the old "sneaker net" where participants can submit data to a central site, but it may be weeks before they can tell whether their data has been entered properly. Any corrections would have to go through another of these multi-week update cycles, etc. This is not a very efficient way for a large group of participants to stay synchronized, especially since

we can make a far better real-time update system available.

Special list-processing features
There are cases where a researcher might begin with a long list of individual names which have not been connected into families, and wish to gradually connect those names into family groups, or add other names which would result in assembling and incorporating the original list of names into family groups. This might occur, for example, if he used our semi-automatic data collector program to assemble in one place, probably on his PC, all of the names of every Huff who has ever lived in the United States, as indicated by index entries to census and other public records (ships' manifests, telephone books, street directories, etc.), and including the underlying digitized images. The names would be interconnected based on year ranges, locations, and data from other sources.

The system offers features which allow names to be connected while specifying relationships – husband to wife, father to son, etc. Gradually, complex family structures can be built up from the original list of raw names.

Other examples of use for this feature might include recording the names of the millions of Russians who were sent to prison camps. The would be a place to identify and document these individuals while other efforts went on to link them into Russian families.

Fairness considerations
Since so much work in the genealogy industry is done by hobbyists and volunteers, there can be a great disparity between the work which people do, and the recognition or rewards which they receive. Some industrious people feel they have been disrespected and abused when people use their data without any recompense in valuable data or money to the original workers. To add insult to injury, some of those who take the data damage it and misuse it, perhaps claiming it as their own work, etc.

One way to help resolve these problems, and encourage those with high quality data to make it available to others, is to increase the recognition and rewards. A first-level solution is to implement a "points" system whereby people accumulate "funny money" through their efforts, which they can then spend acquiring data from other people. If they supply highly valuable data which others are willing to pay their "points" to acquire, then the supplier accumulates a positive number of points. If that person has done little valuable work but wishes to copy data from others, then he will have to pay for that data in points. If he has no points, he can pay in dollars, and those dollars can be transferred to the more industrious people with the most high-quality data.

At some point, when the database has reached a commercial size, more of those points can be turned into dollars and net account balances can be paid out to the hard workers as royalties for the names which they publish. It will be possible for members to login in the future and do little or no work to add to the database, but merely take out the data they wish to use. It will be necessary for them to add money to their account which then can be transferred to the appropriate data owners. Since the fees for these names can be in the range of about $3 per name, rather than in the range of $50 or $60 a name as is typical today, it will still be a great bargain for those who are purchasing the names, and will provide some rewards to those who have supplied the data.

Industrywide efficiency

Completing the nation's genealogy

It appears that all of the online data subscription services, including the free LDS Church site, which mostly show unconnected, or "raw" public records of various kinds, have together amassed about 18 billion entries. (I assume there is much duplication among these websites, with all of them having the same census records, etc.) These many entries probably represent no more than 1 billion unique people, and may only represent 500 million unique individuals.

The serious participants in the nation's genealogy activity have basically two choices: 1) They can continue to work alone, doing much duplicate work, as they comb through this ocean of data endlessly for many more decades with there being almost no sign of any net industry progress nationwide, or 2) they can choose to join this cooperative project which has the goal of going through the world's records ONCE and extracting that 500 million or 1 billion unique names so that this endless and very expensive fishing expedition can stop after being completed, at least until records from new areas of the world are added. To begin with, it seems ideal to complete the basic genealogy for the 70 million people who died in United States before 1930. That is a much smaller task and would provide all the experience needed to make a larger project more efficient.

A side note: it would cost about $1 million a year to store a complete copy of the nation's basic genealogical data, connected into families, and complete with images of all relevant and connected source documents. This is what would be left after collapsing all of Ancestry.com's data into finished form. Ancestry currently receives about $450 million in annual subscription income.

No more emails, no more duplicate data

One of the interesting consequences of having everyone put their best data online for everyone else to see and link-to, is that the need to write e-mails to seek information will almost disappear. Even better, the laborious duty of answering those e-mails will mostly disappear. All the best data will be online, in high-quality form, and it will become easily accessible.

Everyone copying everyone else's data is one of the major causes of the massive duplication seen in genealogy activity. This continual copying is unnecessary if the data is of high quality and can reside in one fixed place. With the new system, in most cases there is no reason for a user to copy data from one location to another in the database. A simple link should be sufficient, for example in the case of linking a woman where she appears as a daughter to where she appears as a wife. By that simple means, both surname groups have added another line of surnames to their collection without having to duplicate any of that data. If someone wishes to reverify that data, in most cases the source documents will be easily available, either through an external link or a link to the internal database. The curious researcher can quickly step through those documents to satisfy his curiosity.

Supporting a "farmers market" of small genealogy websites

There are many genealogists who have done work of interest to others, but they find it difficult to market their data to receive some recompense for their many long hours of work. If they offer to sell their data in bulk to one of the large companies that offer raw data online through subscriptions, they are likely to receive only a pittance for their work, if anything at all. One example is a gentleman who digitized all the records in a county courthouse in a southern state. He wanted to be able to market his data so that he could continue to do similar work. He could make no satisfactory arrangement with one of the large data subscription sites and so was looking for an alternate method.

What is needed is a kind of "eBay for genealogy" which allows industrious genealogists to prepare data of interest to the public and then quickly and easily market it through a single portal. That allows them to control their own data and their own website, but have the benefit of centralized indexing and marketing for their data. It is possible that the owners of many of the nearly 300,000 sites which are indexed on Cyndislist.com would be willing to modify the formats of their websites so they could participate in this

"farmers market." A new site called Mocavo.com has the goal of making data from these many scattered sites available through one portal, but I am unaware of any arrangements to return royalties to the people who supply this data.

Linking all public records to all unique individuals
Whether it is called a "community genealogy project" by genealogists or a "records stripping project" by sociologists, it is a very valuable process to link all known records to the appropriate unique individuals. This greatly limits the amount of duplication, because a person whose name appears in six different census records over a 60-year period, is treated as only one purpose person for genealogical purposes, rather than six different people. But perhaps a more important consequence of a nationwide correlation of all historical records concerning the individuals referenced in those records is that the process will create an "individual level" national history. This is good for genealogists who wish to know all the data they can about a particular individual, and it is also good for those who want to take a broader view of history by studying many people similarly situated throughout specific historical time periods.

Acquiring all the world's source records images through the efforts of individuals
The LDS Church has been capturing genealogically significant source records for more than 70 years, first using microfilm and now digital cameras. This has been of great benefit to genealogists, but at current rates, it will take about another 220 years to capture all the existing source records, by which time many of them will have been lost. Even if the LDS Church could move more quickly, there are probably many areas which would not be willing to cooperate with the Church's efforts but would not restrict the efforts of their own citizens. There are also legal bottlenecks to making the previously acquired records available online to genealogy researchers. These practical and legal constraints can be at least partly overcome by allowing and encouraging individuals around the world to digitize records of interest to them and others and submitting them to a central site such as ProgenyLink. Today's inexpensive cameras and scanners make this relatively easy to do. This could result in having 1 million "camera crews" rather than a few hundred sponsored by the LDS Church. Ideally, these images sent in by individuals would be tied to family data structures as they were entered, but they could be entered into the database independently, without family connections, and later incorporated into family structures using a version of the "community genealogy project" concept.

A more accurate method for indexing source records
The LDS Church has a powerful system in place called Online Indexing which allows perhaps 150,000 volunteers to transcribe and index large numbers of documents such as the recently completed 1940 U.S. Census which contained about 140 million name entries. This provides a very valuable rough index to these masses of records, but probably contains many errors and omissions because of the relatively untrained efforts of its volunteer work force.

One of the necessary shortcomings of this massive cooperative effort is that the indexers have a very limited choice over the materials they are indexing. It is extremely unlikely that they would be indexing the names of people that relate to themselves. In contrast, if they could choose the exact records they were going to index because of the highly likelihood that their people were represented in those records, these volunteers might be able to do a much more accurate job because of their level of personal interest and because they could accumulate some expertise in interpreting those records.

So if we could use that reverse model, where researchers choose the records they wish to examine for their own genealogical purposes, and they do a careful and high-quality job of entering that data for their own reference purposes, then, with the necessary computer resorting process, the result would be high-quality entries for just a part of a record set. However, if there were 4 million people doing this high-quality index process, about 25 times as many people as are currently engaged in the Online Indexing process, then the quantitative results could be similar while the qualitative results could be much higher. This could also be viewed as simply a way to get many more people involved in the indexing process, without having them directly using the Church's Online Indexing system.

Gifts of genealogy data

With the highly efficient processes and low costs of the new system, giving people gifts of genealogical data could become a more common thing. As it is, only nationally known VIPs are important enough for the Church to spend tens of thousands of dollars to prepare genealogical products to give as gifts. But at a cost of perhaps three dollars per name, there would be a much larger market for data from a high-quality genealogy database.

The perpetual "Jewish names" problem
This is a small feature specifically designed for benefiting the LDS Church. Almost every year some anti-Mormon activist finds some Jewish activist person to complain to the Church that there are some Jewish names in the Church ordinance database which might possibly belong to people who were killed in the Holocaust. This allows the anti-Mormon activist to cause trouble, and allows the Jewish activist to pretend to be defending Jewish interests.

Under current rules of operation, there is no way for this problem to be permanently solved. Each year, the Church can offer an apology and delete some names, and the same thing can happen again the next year. The problem could be ended once and for all if every name which was submitted to the temples was shown to be tied to some living Church member by some acceptable degree of family relationship that would give the living church member the authority to authorize baptism of that person. Then the argument would have to move away from the Church in general to the question of whether that specific living person had a recognizable right to authorize the baptism.

This process of clearing each name obviously cannot be enforced as long as the overwhelming bulk of names which are sent to the temples are simply lists of names from census records and other such public documents. The Church system never takes trouble to establish the family relationships, so the only thing that could be done would be to delete every slightly Jewish sounding name from every census or other public records. That is obviously going too far in our self-censorship. A much more precise solution is possible.

Business Considerations

(Disclaimers and cautions)

Goal: Replace or absorb most of the current genealogy industry

This system will quickly replace or absorb most competitors by its sheer efficiency, but only if it is done full throttle. Half measures are cheaper, but far less effective. There needs to be several million dollars available initially, with the possibility of getting more if conditions indicate it.

(Any smaller version of the project may be unsuccessful because the full mathematical power of the new genealogy mathematics of cooperation only kicks in for a large-scale operation. Of course, the large-scale version could fail as well, but the pathway to get there is intended to be full of checkpoints to reverify assumptions as the project grows, allowing the project to be halted if some important piece does not work out as expected.)

Business Model Considerations

There are two basic business models that could apply to this project at different times – 1) the subscription or membership model, and 2) the pay-per-view model. I will try to show here that the real money is to be found in the pay-per-view scenario, not the membership scenario. If enough capital is available, the membership scenario could be skipped completely.

1) The fully capitalized model
In the case where plenty of capital is available, it would make sense to go straight to the pay-per-view version. In this scenario, depending on the quality standards to be set, somewhere between $70 million and $700 million would be invested in creating the first high-quality database containing the 70 million people who died before 1930 in the United States. The data could be processed very quickly and efficiently and to a very high quality standard. Assuming that each of the highest quality names, on average, would be worth $40 in sales over perhaps a 10-15 year period, then that completed database would be worth about $2.8 billion. The data could be assembled in from one to three years, with sales beginning perhaps after the first six months and continuing on to the 10 year or 15 year mark, with the bulk of the money collected within the first 10 years. If one could invest $70 million, and five or 10 years later collect $2.8 billion, 40 times as much, that would seem like a good investment. Obviously, the project could start small to be sure to test all concepts including the quality and value of the data that was assembled, but then it could be ramped up very quickly to full-scale.

2) The membership/bootstrap model
If sufficient capital is not available for a quick and thorough start up, and the project needs to start small and bootstrap itself up, then the subscription or membership model is required.

This startup model is hampered by many constraints. Most important is the fact that the overwhelming 1000-to-one efficiency advantages of "industrial strength" cooperation, and even the smaller but still important 30-to-one efficiency advantage of entering data in descendent sequence, do not come into play at the beginning. This tentative new project might be viewed by many as simply the 30th entry of a genealogy database system into a very crowded marketplace. Competing head-to-head with the free PAF

program which is adequate for many beginners, or the $20 one-time cost of the Roots Magic program, which also has a free version, etc., is not a very exciting marketing position to be in. The new system does have fantastic powers of cooperation built in, but most genealogy hobbyists would not grasp the power and would see no need. Typically, they are looking for a few generations of their closest family, and are not even thinking about improving the state of the genealogy industry nationwide.

However, if it is possible to get past this painful startup phase, then something magical could happen. As the database reaches a substantial size and contains mostly high-quality data, then the accumulated database could be worth far more than the membership dues, as is suggested in option one above. At that point we could promise a complete refund of the membership dues of most people, plus a substantial amount of royalties to be received back from their good work. However, at the beginning of it all, these are all very vague concepts and promises, and seem very unlikely to the typical user with only a very small personal vision for the future concerning genealogy work.

At the point that the database can be considered of a substantial size and quality, suddenly there would be a need to increase security considerations. At some point, many people would catch on to the fact that the database is of great value, and they would begin devising ways to drain off some part of that $2.8 billion in market value. Even for those who are legitimate members, that would suddenly become a temptation.

3) A hybrid model
It seems possible to create a hybrid model, where the first $1 million or $2 million are spent on building a significant demonstration database, potentially including enough names to begin to make legitimate sales. That might be a forceful enough demonstration to drive into the consciousness of many serious genealogists that this methodology could work. At that point, some of them might be willing to invest their own time and effort in publishing high-quality data which can then be sold to others. If this new concept can be accepted, then the remainder of the necessary $70 million to $700 million to finish the job properly can essentially be had on credit, with half the proceeds going to those enlightened genealogists who do the work.

Summary
In summary, a small and feeble bootstrapped startup is not very exciting unless there is a good possibility that the project can eventually go from the membership model to the much more lucrative pay-per-view model. The hybrid model at least has one of its goals as establishing the viability of the pay-per-view model early in the process.

More on the bootstrap/membership scenario
If we start at the bootstrap/membership level, and offer the typical free/premium or "free-mium" deal where they get many important features for free and then can pay for an upgrade, the Internet and Church culture concerning genealogy of getting everything for free, means they will go with the free version, and probably never get to the premium version.

The biggest problem at the beginning is that there is no huge reservoir of data to draw in the data-hungry seekers. That situation only happens later. Most genealogists cannot even comprehend what a serious cooperation effort would look like.

The users have to BUILD that reservoir of data they are instinctively looking for, and apparently no one wants to do that unless they are paid a wage to do so. Even getting free use of a website is probably not enough to draw them in.

Let's go through a few practical scenarios here. If we had 3000 members paying in $30 a year to use the system, that produces $90,000 in revenue. (We are probably need to have 30,000 members altogether to get 10% of them to pay the premium fee.) That is barely enough to pay one full-time person and a few part-time people. The question is whether the company could grow enough so that the 1000-times cooperation feature can kick in. There is a possibility that with 30,000 users they could accidentally discover that they are cooperating with people of importance to them, so that more of them would choose to go premium for the extra features..

Or let us try 30,000 members at $30 a year (probably translating into 300,000 members, most of them using the free version) which equals $900,000 per year. That is reaching the size where some of the system effects should start to be obvious as users find other users who have data which is valuable to them.

Professional genealogists' resistance to changes to their historic business model
Probably the largest impediment to getting this system started is that all of the professional genealogists, many of whom inhabit the largest organizations such as Ancestry and the LDS Church, are happy to have more raw data materials online, because that helps their private genealogy research business, but they have a powerful resistance to the idea that the process of aggregating high-quality genealogy data could be done more efficiently and more cheaply by less qualified people. It appears that there is no convincing these people to change their ways for the benefit of the genealogy hobbyists masses, so this project should be prepared to go ahead in the teeth of this professional resistance from probably all existing organizations. I can say this with some confidence because I have talked to hundreds of these people, and the more professional they are, the more certain they are that they do not want this cooperation project to succeed.

It should be noted that this powerful resistance to any improvement in genealogy industry methods also extends itself to the programmers who support the desires and projects of the genealogists. Anyone who wishes to work as a genealogy programmer realizes that they have to stay on good terms with the large organizations that have all the money – the LDS Church, Ancestry.com, etc. – if they want to work in that industry. If they have shown themselves to be traitors to the cause by working for some competing organization such as mine, they would irritate these single-minded genealogists, and typical comp politics would keep them from working at lucrative positions at these kinds of large organizations.

Richard,

I don't feel like I have communicated to you the true nature of my project idea, so here is another try. You thought it troublesome or pointless that I spoke about billions of dollars, but I need to show that that those kinds of estimates and computations are the _essence_ of the project, not a mere sideshow or affectation.

The commercial goal of the project is to quickly finish all the genealogy research for the United States and then sell it for about $3 billion.

1. Timing

The basic genealogy for the entire United States can be done in one year using my procedures (or in three years, at a much more leisurely pace). About 1000 to 5000 people need to be employed in this project, depending on whether it is the one-year version or the three-year version.

(It could theoretically be completed in two weeks by the entire body of serious US genealogists – but this illustration is impractical because it would take more than one year to inform the participants. They would only need to enter 18 names each, but they would need to know a great deal in order to do that correctly.

2. Cost – Church

This genealogy data project could be done completely for as little as $5 or $10 million if done by the Church. The Church could use volunteers and pay out very little money. Notice that the Church is spending about $100 million every year on genealogy software and databases. The cost for adding in my new system would be very minimal in comparison.

3. Cost – commercial

Proceeding commercially, the eventual cost for the completed database could be as low as $70 million. The project could begin for much less, and bootstrap itself to completion.

4. Data value

The value of the data is important to know. Unless we have a good idea of the final value of the US genealogy database, we cannot make good judgments on how to do the project.

I have done my own informal studies at many genealogy conferences, but perhaps we need some formal marketing studies to make our point to money men.

I estimate that the value of each high-quality name in the database will eventually be worth about $40, assuming names are sold for about three dollars each, and they are sold multiple times – about 15 to 20 times on average. Recent names may be worth only a total of $3 each, sold once, but names back multiple generations will be sold many more times, so they can be worth hundreds or even thousands of dollars each.

If the completed database of 70 million names, at $40 apiece, is worth a total of $2.8 billion, that means the project gets back 40 times (or more) than the amount that was invested. That sounds like the payoff ratio that venture capitalists are always seeking.

5. Church versus commercial

The reason this should be done by a private company, or by the Church in conjunction with a private

company, is that if the Church creates the database for "free," it will still have created an asset worth $2.8 billion. The Church would need to decide if it wishes to just donate that very large amount of money to the professional genealogists of the United States, and allow them to sell that data and collect the $2.4 billion for themselves, or whether the Church wants to use a method of collecting most of that money itself, so that it can use that money to greatly accelerate the creation of a truly worldwide database.

6. Focusing only on the tool

We have spent most of our time talking about the website tool, and that is important, but it pales in comparison with the concept for doing the larger project.

Here is an analogy to help make my point. If we were reviewing the D-Day invasion of Europe in operation "Overlord," we might get overly focused on the importance of the landing craft used to ferry soldiers from deepwater ships to the landing beach. It was indeed important to have the right kinds and numbers of landing craft, but without the intense preparation on hundreds of different levels of all aspects of the operation (all air, land, and sea assets, intelligence, sea state, disinformation campaigns, etc.), the landing operation could not have been launched and thus could not be successful, and the existence of the landing craft would have been a great deal less important. The landing craft were "necessary but not sufficient."

If we start out with just the tool, and ignore the larger plan, then we immediately lose most of the power behind this project idea. If data is collected and entered in descendent sequence, that offers a performance boost of up to 30 times, assuming that the goal is to get all relevant data into the database quickly. Once the data has been entered correctly and connected with the work of other people, the productivity multiplier can go up to another 1000 times. (There is a second-level productivity multiplier working here which could theoretically be 30x1000 equals 30,000. I'm sure there is some second-level multiplier effect, but it is probably not as large as 30,000. We will only be able to calculate that factor part way into the project, after accumulating significant data.)

In other words, starting out with just the tool, without the advantage of the applied mathematical power of this 1000-to-one productivity advantage, it is going to be a daunting and difficult task marketing it in an environment where there are perhaps 30 other free and commercial products for people to store relatively small amounts of personal genealogy data. If people were willing to accept the promise of future feature benefits, the sales process might be possible. But after talking to thousands of people at a dozen different genealogy conferences, it is clear to me that no one is going to comprehend this new concept, and change their behavior accordingly, until it is proven unequivocally to work (or they are commanded by their church leaders to get with the program).

Therefore, it appears that the minimum way to begin this project is to create a large demonstration database where people are paid to do the assembly work. If this partial database is large enough to begin serious commercial sales, then genealogists may finally come to understand that they can make money by contributing their best data, and that this is the wave of the future. At that point, it may be possible to accept their data on consignment, saving the project tens of millions of dollars in upfront costs, with the data contributors receiving royalties when their data is actually sold.

We can be sure that Ancestry.com will truly hate this project, because it threatens to use ALL of their data just ONCE, and that will greatly lessen the long term value of the masses of data it now holds. In the current situation, genealogists make almost no net progress each year, as 2 million of them each pay $300 a year to Ancestry.com, and Ancestry.com's data can be used over and over again, endlessly. Ancestry's current "infinity" of future sales would suddenly be cut to a much smaller level.

7. Is this project impossible, or merely pointless?

The very "impossibility" of this method is its big advantage, its magic. Patents are required to be "new, useful, and non-obvious" – and this procedure and plan is certainly "non-obvious." Otherwise it would have been discovered and completed years ago. Actually, this new project can only happen on top of all the work which is gone before. It is a harvesting operation, and its time has come.

8. Some data storage numbers
The entire finished US database would fit easily on a 1-terabyte disk drive costing about $100. Calculation: 70 million names times 5000 characters per name equals 350 GB, or about one third of a 1 TB drive.

If we add all of the images, which might average about 2 MB per person, that would add about 140 TB of data. 140 TB times $100 for each spinning drive drive equals $14,000 for one copy. For three copies for security, that would be $42,000.

Notice that it would take about 70 of the millennium DVDs to hold the basic data. At two dollars for each disk, that would be $140 to store this basic data.

9. The basic pay-per-view concept
There is a strong impulse to think in terms of subscriptions for genealogy data, since that is the way that most of the successful businesses are run today. But there's no reason to believe that the subscription model would work on this project. At Ancestry.com, and others, they are offering many billions of names in simple lists, often weakly identified, and unconnected together into families. To any particular subscription customer, each of those names is worth about 3 millionths of a penny. ($300 a year or 30,000 pennies/10 billion entries equal 0.000003 per entry). Obviously, there is not much incentive for users to drain off large amounts of this data which is of such low value to any particular customer.

For any particular subscriber, no more than about 100 names out of those 10 billion names will be of any use, and at the beginning no one knows which names they are. Thus it can make sense for someone to pay $300 a year to go on a fishing expedition to find this handful of names, and Ancestry need not have much fear of its customers taking all its data.

However, in the situation which I am describing for this new project, each name in the database is worth at least $3, and some of those names will be worth $3000 because of their situation in the database structure. In other words, with each entry in the database being worth 1 million times or 1 billion times more than any line of data in the Ancestry.com database, then there is an overwhelming incentive for users to try to benefit by draining off data and selling it elsewhere. In other words, it would be suicidal to have a subscription database which allowed users to look at and copy whatever data they wished. The nearly $3 billion value of the database would be gone in a fairly short time, drained off by professional genealogists for their own resale. This is why a carefully controlled pay-per-view sales method must be used.

10. Sales to the church or the public?
The main clientele is the 4 million genealogy hobbyists in United States. The Church genealogists may be as low as 2% of the total genealogists in United States.

The Church has decided that it's very low quality names are sufficient for its needs, so the Church would probably pay little or nothing for our data. However, individual Church members might have higher quality standards and so might pay $3 dollars per name to ensure that they were putting only high-quality data into Church systems. For example, we might find 500,000 church members willing to pay $3 dollars a name for 1000 names each, or $3000 over several years. That would be $3000x500,000 equals $1.5 billion – a nice portion of the total value of the database.

11. Size of the market
It is important to know the size of the market and how it can and will be changed by introducing our new product. I have estimated that there is $60 billion worth of volunteer activity in today's genealogy industry, with about $6 billion being spent on computers, Internet connections, books, conferences, and professional research assistance, totaling $66 billion. The big question is how much of that activity can be turned from unpaid volunteer or hobbyist activity into paid semi-professional work through data supplied by us.

In today's genealogy industry it appears that the cost per professionally researched name is in the $50-$60 range. I know of no estimates of the real cost for each non-professionally researched name, but it is probably quite a bit higher than when done by professionals, simply because the process must necessarily be much more inefficient, for example burning up enormous amounts of time for self-education before effective actual research can begin. (Many genealogy hobbyists are in it for the game and the hunt, and they enjoy the constant learning.)

With the new methods, the costs should be in the $1-$10 range. Almost everyone should benefit from this new price point. The genealogy hobbyists will be most benefited. Some of them will be able to earn money applying their expertise to help to extend the database. The Church will indirectly benefit as it begins to get a large amount of high-quality data, so that it can gradually raise its expectations about genealogy programs. Professional genealogists can get on board if they wish, and become data suppliers. Otherwise they may lose a few low income customers. Ancestry.com may eventually earn a little less, even though they will probably earn more at the beginning as activity increases. If they add new areas to their data offerings they can continue to function much like they do today. If they are willing to cooperate with the new methods, and offer access to their data using permanent URLs, they can stay completely current.