Elements and Strategies for a Worldwide Genealogy System

At Least

One Complete Solution Is Available

For 3rd Annual Workshop on Technology
for Family History and Genealogy Research – FHT2003

Kent Huff

Presentation Summary

- Everyone in the world would like to know their genealogy

- We have the basic technology we need to create the necessary system.

- But we would need to introduce a new way of thinking about research methods and related computer support – a **new paradigm**

- Briefly, I will describe a **complete system design concept** (including a running **prototype** at www.genreg.com)

We need to get the right

- Data Quality
- Data Coverage
- Tools for users
- Incentives to do the work

The Criteria for a robust, World-class System:

1) High Quality Lineage-Linked, Online Data
Leaves no reason to do duplicate research.

2) Mass production data entry
Multiply by 100 times the current number of finished researched names.

3) 100% data coverage is achievable
Points out holes in the database so they can be filled

4) Manage duplication and error levels
Minimize duplication and errors while still allowing alternate versions of data when necessary

The Criteria for a Robust, World-class System:

1) Features that promote and enable good **quality standards** and measures so that, in most cases, when research work is done properly and recorded, there is **no reason to duplicate research**. This alone will drastically cut the duplication of ordinances and the unnecessary repeat of underlying research work.

2) Features to make available to all users the highly efficient process of **mass production data entry** and linking, easily multiplying by 100 times the current output of researched names.

3) Features and the incentives for users to actively **look for holes** in the database and to fill them with good data, until **100% coverage is achieved.**

4) Features to **manage duplication and error levels** within the database while still allowing users to present their alternate versions of data when the records are in conflict.

No Existing System Meets All Criteria

I am not aware of another system design that meets all the criteria needed for worldwide success.

If there were one already operating, I assume we would all know about it, and could relax.

Many Major Benefits

This new concept should
advance the pace of genealogy research
nearly as much as the use of **microfilm** has done in the past.

With this system, some remarkably
efficient data extraction
is possible.

Large But Reasonable Genealogy System Goals

- Collect the basic genealogy data for all deceased Americans (250 million). Do this within four years. **No increased workload. Just use different methods.**
- Collect the basic genealogy data for the 6 billion deceased people of the world for which we have records. Do this within 10 years.
- Do this using currently available Internet technology. **Create a highly successful system.**

What is the true scale of the effort that's necessary?

Probably less than you might think . . .

If it were done commercially:

U.S. genealogy database:

$50 million, disk storage 250 GB

World genealogy database: (20 times larger)

$1 billion, disk storage 5 terabytes

We don’t need a new technological “silver bullet” . . .

Current technology is good enough to build a highly successful system.

The **efficiency** of the new system will draw in millions of users. It can be a **hundred times** more efficient to use, in the aggregate, than current technology.

These expanded participants will add huge amounts of good data and do much of “our” work for us.

Large-Scale Cooperation is the Key

Large scale cooperation and **mass production** techniques means that we don't need further technical advances.

(Of course we would always be happy for more technology, but it's not required)

“Build it and they [,the names,] will come [in].”

The success of the new system is not dependent on automated conversion or indexing of old records. Manual methods are sufficient.

The conversion of old records into a more usable form is a one-time task and much of it will probably have to be done manually anyway (through cooperation).

We should **put our technical efforts into creating the new system** for accepting the data, and then re-assess whether other records conversion technology is needed. It will not be necessary, I believe.

To Illustrate the Powerful Effects of Cooperation:

Do all U.S. in 4 years, do world in 10 years

No greater workload

If all adult Church members in the U.S. (3 million) put in just 8 hours of work a year on a mass census data entry project – a super Name Extraction program – the entire US genealogy data in census records could be entered and lineage-linked in 4 years. Each member does 25 names a year, 100 names in all.

If all missionaries were to work on the U.S. project, two months would do it all – 250 million names.

After One-time Conversion, Keep Our Archives Current

Don’t wait 100 years to access today’s records.

- Once the old records are converted to a new format, we should be able to keep our genealogy data current thereafter using only the newer computer formats.

- We may still have to wait nearly 100 years to get census data, but we have many other sources of data today to compile genealogical data almost at the moment it happens.

- There should be no need to wait 100 years to get important genealogy records about today’s vital statistics events.

Some Mathematics of Genealogy
relating to the duplication problem

1. An ancestor 10 generations back could easily have one million descendants today. $4^{10} = 1{,}048{,}576$. All of them might be interested in researching him. This obviously could lead to much duplication of effort.

2. A researcher going back 10 generations would have about 1024 surnames to search. $2^{10} = 1024$. This is a difficult task for a single researcher, especially if all surnames are in different countries.

Scale of effort to do all United States genealogy:

Proposed Methods imply a

$50 million effort (= one temple)

250,000 times more efficient than:

Current Methods imply a

$12.5 trillion total cost (= one year U.S. GNP)

Cause of high cost: Duplication of effort

Implied duplication of research effort:

3. If everyone in the U.S. did his or her own genealogy back 8 generations, they might find 1,000 names in their pedigree families.

250 million U.S. people times 1,000 names equals 250 billion names. Times $50 per name
= **$12.5 trillion total cost** (one year U.S. GDP)

A huge price in effort or money. Done cooperatively, at **$50 million in effort, it is 250,000 times more efficient.**

“The Royal We” - Duplication of Ancestors

Mathematically, everyone here is descended from many historical figures, including:

1. Nefertiti and Confucius, Muhammad and Charlemagne.

2. The same particular (but unknown) person in Europe in 1400 A.D.

3. 80 percent of the adult Europeans alive in 1000 A.D.

“We're all descended from Charlemagne. But can you prove it? That's the game of genealogy.”[1]

Obviously, mechanisms for avoiding unnecessary duplication of research will have huge payoffs for all concerned.

"The Royal We" - Actual Quotes

"The mathematical study of genealogy indicates that everyone in the world is descended from Nefertiti and Confucius,
and everyone of European ancestry is descended from Muhammad and Charlemagne."

"All Europeans alive today have among their [common] ancestors the same man or woman who lived around 1400."

"20 percent of the adult Europeans alive in 1000 would turn out to be the ancestors of no one living today (that is, they had no children or all their descendants eventually died childless); each of the remaining 80 percent would turn out to be a direct ancestor of every European living today."

[1]Steve Olson, "The Royal We," *The Atlantic Monthly*, May 2002, pp. 62-64.

Shortcomings of Current Genealogy Systems To Be Solved by the New System

1. Extreme **duplication** of research, partly because
2. **Communication** is difficult among researchers, so
3. **Cooperation** is also difficult among researchers
4. **No suitable place** for everyone to put their finished work to make it viewable and shareable.

All factors are interrelated - These are 4 ways to say the same thing. "4 sides of the same coin."

Improve Communication of Researchers:

- Replace most E-mails (& Internet searches)
- Emails will be the exception, not the rule
- Get exact information instantly

With all the data in a central database, one can quickly see what any **researcher** has done and is doing now, or check the current status of a particular **family line**. One can get exact and complete information instantly rather than wait for days or weeks for the less complete data that might be sent in an email. This will have the effect of **replacing billions of emails** (actual and intended) each year among those trying to coordinate their efforts with other researchers, known and unknown.

Communication - Email Replacement

The **central database** allows everyone to know instantly and continuously what everyone else has done and is doing. It has the effect of providing or **replacing billions of emails** each year to those trying to coordinate their efforts with other researchers, known and unknown.

Mass Production Techniques

Reverse the usual research process. Instead of researchers each finding a few names using endlessly repetitive searches of millions of records, we

Enter the source records once and for all, and derive all the names and relationships. All names are thus automatically documented. Anyone can quickly learn to assist in this process.

Get help from millions of the world's genealogists. Give them a convenient way to do their work, which is also our work.

A Self-Documenting System

Extra documentation benefits from mass record entry:

1. Maintain direct links (or library references) from each individual name to all original records
—allows anyone to quickly review provenance of each record.

2. Indexing together by individual name all the various kinds of source records (census, land, court, etc.) also provides other valuable historical and sociological research information. Enter census records first, then other records.

This feature will help greatly with compiling interesting family and personal histories.

Illustrate Cooperation Possibilities

- Theoretically, if each of the 250 million living people of the U.S. did one name of the 250 million U.S. deceased, all the research would be done.

- Same for the 6 billion world residents and world deceased.

- This is difficult to do now, but:

- New system would help this cooperation happen.

Peer review will increase quality

A critical point epitomizes the changes I propose:

Which would you prefer to find in your searches?

1. **One name, checked by 20,000 people.** Use a centralized system to look up one name which has been reviewed by 20,000 people who could not find anything wrong with it (or who added their notes to that name if they had questions). This is the peer review process.

2. **20,000 versions of one name, each checked by one person.** As often happens today, one might look up 20,000 versions of the same name, all reviewed by only one person. Those versions could all be wrong, and the right version could still be missing. No mechanism is available for large-scale peer review.

Strategy:

Planning to do *all* the world's genealogy is the cheapest way for us to do our "own"

This approach to system design:

- **allows mass production methods to be used,** plus
- **allows millions of outsiders to help in the process.**

Church members could get *all* the world's genealogy work done more easily, quickly, and cheaply than if those same Church members focused all their resources on doing only their "own" genealogy work. We hope "they" and "us" will soon be the same. As the Church grows larger, the efficiencies increase. We should use our expertise to help the whole world do their genealogy work, while we are helping ourselves.

Let Us Act like One Big Family:

1. It is an **appropriate** thing that family history research can be done most efficiently **if done by everyone** together – the more, the better.

2. The Final Goal and Benefit of all this effort is to

Do Real Family *History*.

By using the most efficient **mass production clerical data handling** methods to quickly do all basic identification of people and relationships, we then leave the largest possible amount of time to actually get acquainted with our **ancestors' personal and social histories**, instead of exhausting ourselves on merely finding their vital statistics.

The universe is made of stories, not of atoms [or vital statistics].

-Muriel Rukeyser

Many Good Missionary Effects:

Everybody in the world cares about their genealogy, even if they don't yet care about the Church, the Book of Mormon, etc.

But if they learn about their genealogy through the Church, then they will look further, and learn more about the Church.

Special opportunities may occur in the many large countries with a history of ancestor worship. We can serve them and help them learn the true reason for their concern.

Other good effects on Welfare and Public Relations. A total value to the Church of up to $20 billion.

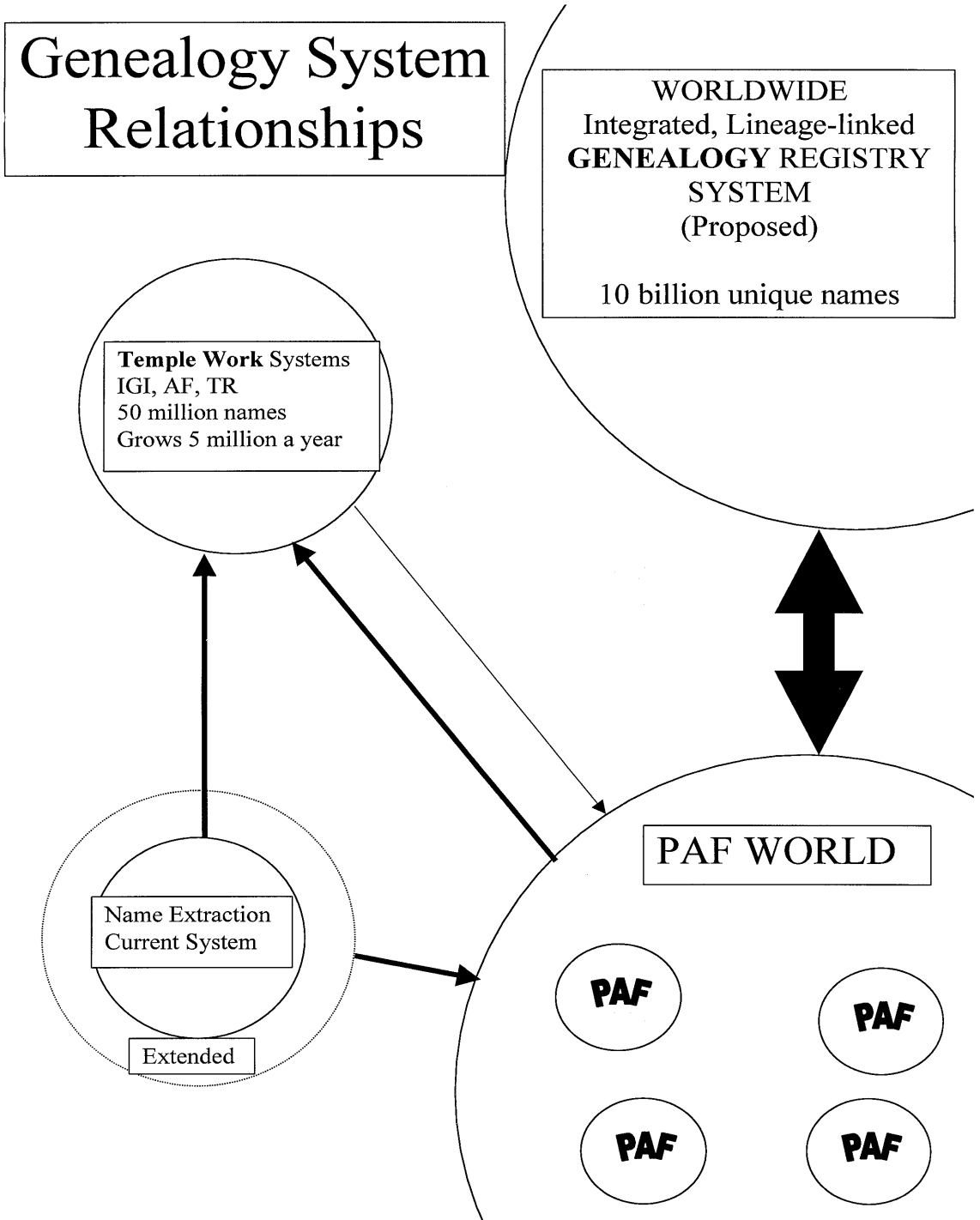

Constant Improvement
Capstone:
Consolidated Results
and Indexing
By
Others
By
Church
Records Processing:
Collection, Conversion, Indexing
The World's Records

Full-Data Genealogy System

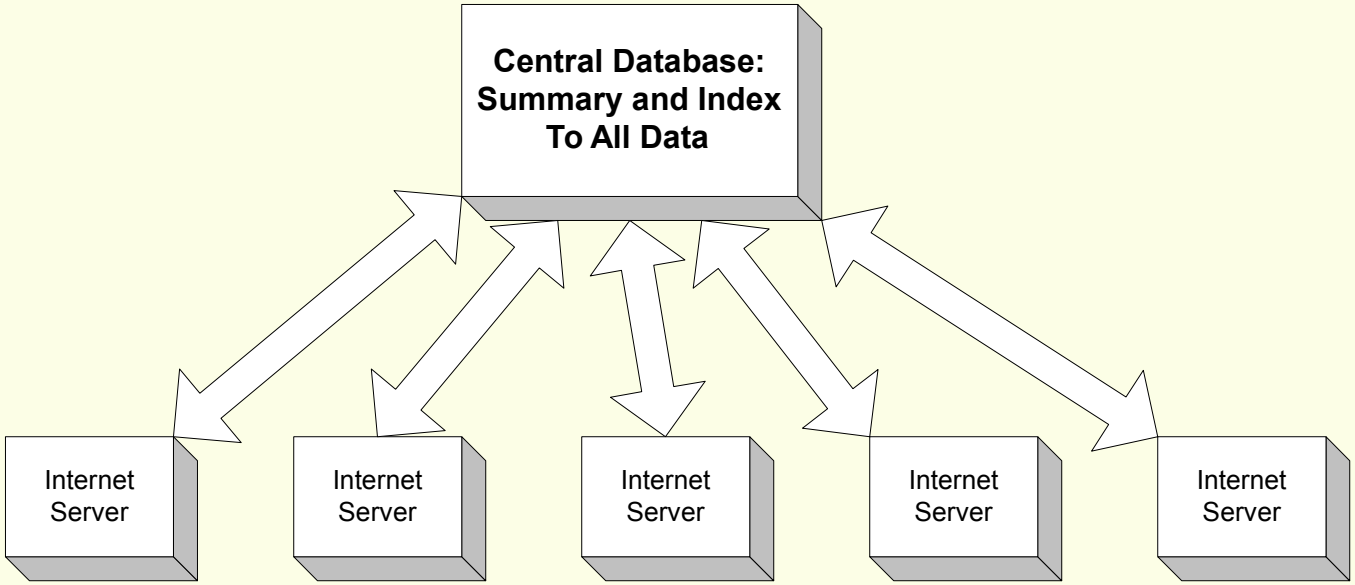
VS.

Limited Data Temple Work System

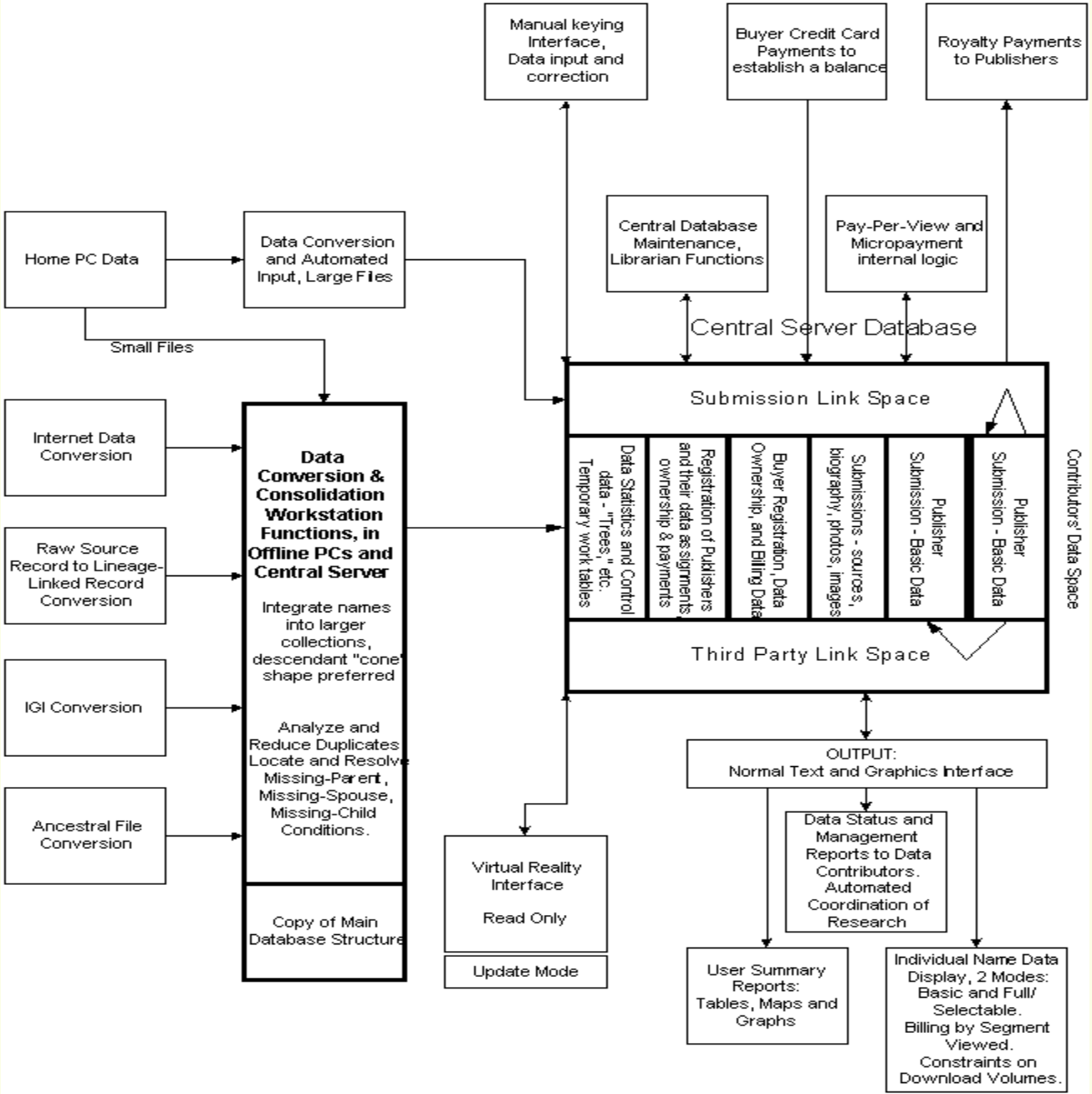
A complete genealogy system is needed to get accurate data to the temple work system. Someone needs to sponsor such a system.



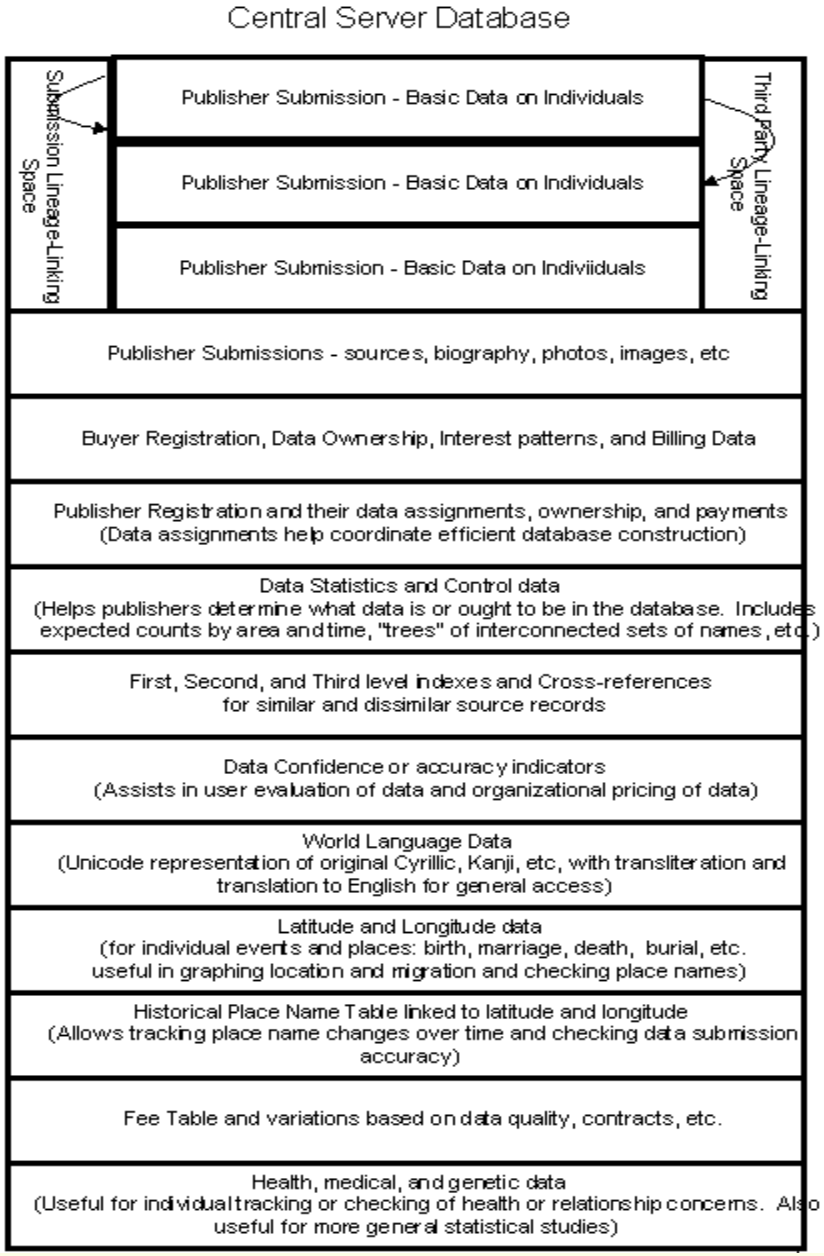
General Architecture:
Balance Centralized Summary and Control with Unlimited Storage for Enrichment Data



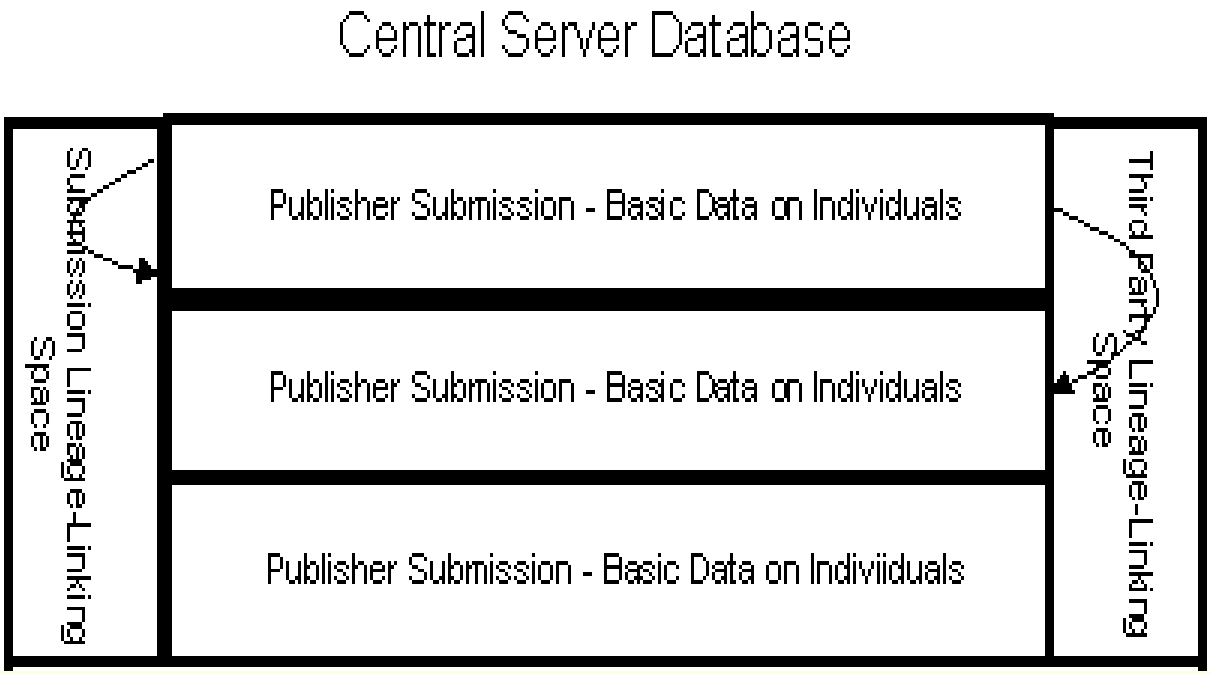
Genealogy Registry Main System, Simplified Overview



Genealogy Registry Main Database, More Detail



Genealogy Registry Main Database, More Detail

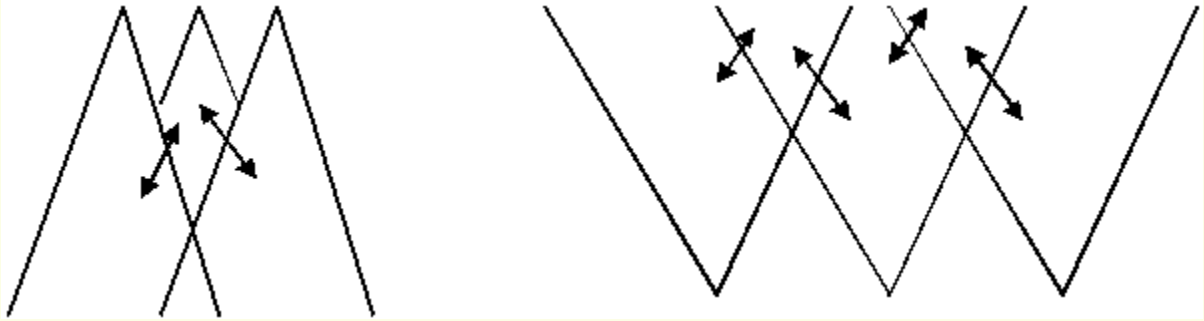


Virtual Merging

Arrow on left – represents links **within** data submissions

Arrow on right – represents links **among** data submissions

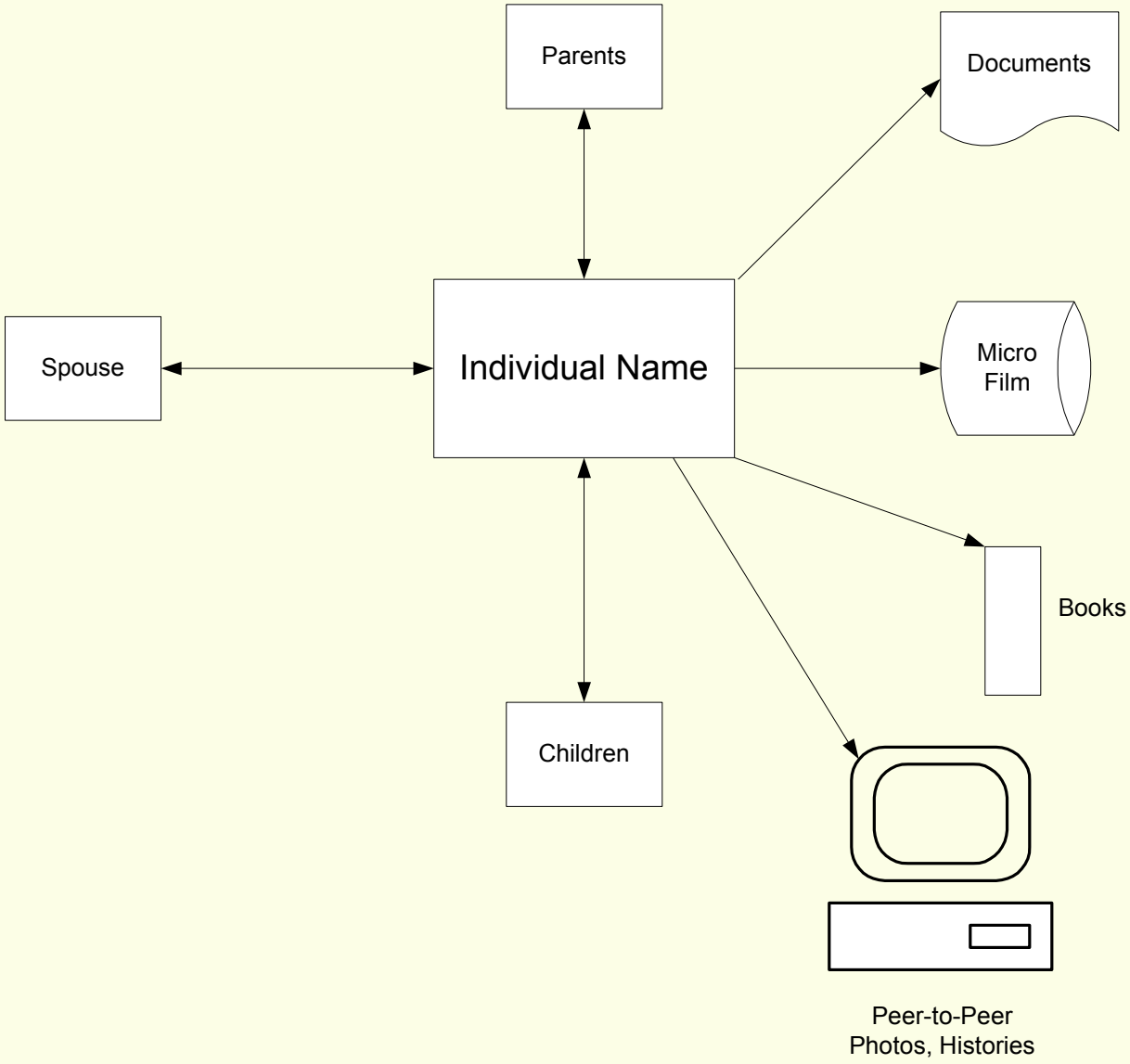
Virtual Merging Concept



- Link separate submissions
- Allow removing or hiding of duplicates
- Reduce redundant effort

See www.genreg.com for details

Database Pointers
From Individuals



The Best Way to Enter Data

Descendant Form—30 times better

There IS a best way to enter data to minimize the duplication of data and the effort of manually-controlled merging.

Data is prepared offline and entered in descendant form – top down, oldest to youngest. This can bring a 30 times improvement in merging efficiency over the usual ascendant form.

However, that method is not required. Any way it is done, millions of researchers can contribute online and eventually end up with a high quality combined file – stitched together by hand.

Cooperative Publishing System

This is a cooperative publishing system, where one name at a time can be added to the whole and is immediately available to all.

Paying for the System:

1. Who funds its development?
   The Church or some business?

2. To users,
   is it **Free** or **Pay-Per-View?**

On a system of such a grand scope,
I can safely predict that

The Politics

will be more difficult than

The Technology

- Getting the **consensus**, administration, and economics right

  will be far harder than

- Getting the **technology** right.

For more information, see **www.genreg.com**

1. Copy of this PowerPoint presentation

2. January 2003 presentation booklet (PDF, 64 pages), intended for use by Church to help on new genealogy project

3. To examine the **prototype** features and database, click on
"Add to the Database or Search it"

The End

How Could the Needed System Be Created?

1. A **merger** of several existing genealogy businesses? Use their current features, plus add the extra features needed.

2. The **Church** does the central part and leaves the distributed parts to the businesses?

3. Have **new business** investment build the central part? Perhaps get Church **endorsement** to help assure success?

The Innovators Dilemma: When New Technologies Cause Great Firms to Fail (Boston, Mass: Harvard Business School Press, 1997), Clayton M. Christensen.
Normally, existing businesses cannot adjust to the new parameters, and so are eventually replaced.
Focusing on the near-term, ignoring the long-term, and fear of cannibalizing existing products are causes of failure.

Internet Genealogy Business Worries

Will their data become obsolete?

Existing Internet genealogy businesses seem to fear that their huge offerings of "raw" data (non-lineage-linked) will become obsolete if a large, high quality, cooperative genealogy database is created, and duplication of research is vastly curtailed.

Their fears are unfounded.

I believe this business fear is unfounded, although changes will occur. Their data will tend to be used even more as the database is being constructed, and then will continue to be actively used as many people periodically review the source records supporting each name, gleaning historical data, etc.

Law and Public Relations
(consider all issues)

Transborder data flows. Only economic activities are of interest to governments. The dead are not economic actors and so genealogical data are of little interest to governments.

Billions of records are already here, but unprocessed. No legal concerns.

Holocaust-type concerns about use of names of the dead in religious services. Proper system will easily show links with LDS living, giving us equal claim. See “Royal We” article.

What is the State of the Art of Genealogy Technology?

Old technology:

Pro:
- Provides for a consolidated lineage-linked file (AF)
- Allows for shareable individual submissions (PRF)

Con:
- Offline submission of data, 2-year turnaround on updates

AF
- Expensive periodic data updates for 3,000 libraries
- Data submissions not kept intact in merge process; lose data
- Submission data is merged (by machine algorithm) into one combined file – causes errors. Experienced researchers could do better.
- Can't quickly correct errors
- No multiple interpretations of data in main pedigree file
- Difficult for outsiders to use system – no incentive

PRF
- PRF submissions are kept intact, but cannot be combined online, cooperatively. Must be done offline, individually.

What is the State of the Art of Genealogy Technology?

New technology: Take best of the past, plus add many new features

- Online, for immediate update and verification – easy error correction
- Allows for shareable individual submissions
- Provides for a consolidated lineage-linked file
- Data submissions are kept intact, no data loss– only virtual merging
- Submissions can be gradually, incrementally linked and combined
- Most duplicates can be removed (or at least removed from sight, and from most search options, to avoid confusion)
- No machine merging – there are enough experienced eyeballs to do this manually
- Cooperation, peer review facilitated; multiple data versions possible
- No data distribution expense
- Outsiders can easily use system and share research

Author's Biography

Spent 30 years in technical data processing work, including four mammoth "billion dollar" projects. A BYU graduate and also have two law degrees. Full resume available on website **www.genreg.com** at end of first document listed there (64 page pdf).

Prototype

I coded running prototype in Visual Basic Script (ASP) using the Microsoft ACCESS database engine. Only a few simple functions are directly visible on the prototype website without a narrator to demonstrate it, but the internals (40,000 lines of code) do all the important functions needed for a worldwide system. See www.genreg.com.

Footnote

Steve Olson, "The Royal We," *The Atlantic Monthly*, May 2002, pp. 62-64.

Contact

Kent W. Huff
1748 West 900 South, Spanish Fork, Utah 84660
801-798-8441, huffkw@juno.com, www.genreg.com

The Royal We

"The mathematical study of genealogy indicates that everyone in the world is descended from Nefertiti and Confucius,
and everyone of European ancestry is descended from Muhammad and Charlemagne."

"All Europeans alive today have among their [common] ancestors the same man or woman who lived around 1400."

"80 percent [of the adult Europeans alive in 1000] would turn out to be a direct ancestor of **every** European living today."

"We're all descended from Charlemagne. But can you prove it? That's the game of genealogy."[1]

Obviously, mechanisms for avoiding unnecessary duplication of research will have huge payoffs for all concerned.

The Royal We

Complete quotation:

"20 percent of the adult Europeans alive in 1000 would turn out to be the ancestors of no one living today (that is, they had no children or all their descendants eventually died childless); each of the remaining 80 percent would turn out to be a direct ancestor of every European living today."

Elements and Strategies for a Worldwide Genealogy System

Goals

My personal goal for several years has been to design and see implemented a complete worldwide genealogy system using current technology. I believe the task could be done now very efficiently without relying on or waiting for technology that is more exotic than what is commonly available on the Internet today. Reconceptualizing research methods to take full advantage of current technology could move genealogy research ahead as much as the use of microfilm has done in the past. In this paper I point out a few key points of the reconceptualization I think we need. See **www.genreg.com** for more technical background and discussion, and **a running prototype** of the proposed system.

The system would first aim to contain all of the names for the 250 million deceased residents of the U.S., with minimal duplication, and then go on to do the same for the 6 billion recorded names of the world's deceased. The U.S might be completed in a four-year period without stress, and the remainder of the world by the end of ten years. No higher level of work effort would be required than is going on now.

A practical business viewpoint

I believe that the constraints and inefficiencies of the historical and current methods of genealogy research have blinded people to the relative simplicity of the task if it were viewed as just another business data processing project. For example, Walmart Stores probably does more computer processing in one day than it would take to do the entire ten-year worldwide project. The results of the U.S. portion of the project could all be stored using about 250 gigabytes of disk space, a $500 cost on a single home PC with today's hardware costs.

The mathematics of genealogy

To reach my conclusions on various issues, and create the resulting design, I have explored several aspects of the simple mathematics of genealogy. I began with quantifying the processing assumptions contained within various approaches to genealogy research. I discovered that an explicitly cooperative plan, in the aggregate, could be up to 250,000 times more efficient than that possible from the viewpoint of a single lone researcher. That number comes from comparing the effort required if every U.S. resident were to do their complete 8-generation genealogy on their own, as opposed to the efforts of genealogists organized to do that same amount of finished research using a system explicitly designed to avoid duplication of effort and to encourage and assist the maximum cooperation.

The concern about unnecessary duplication of research can be illustrated in two simple mathematical facts: 1) An ancestor 10 generations back could easily have over 1 million descendants today (4^{10} = 1,048,576), all of them possibly interested in researching him. 2) A researcher going back 10 generations will typically have 1024 surnames to trace (2^{10} = 1024), an impossible task for one person. More rigorous statistical studies along this line have demonstrated that "all Europeans alive today have among their [common] ancestors the same man or woman who lived around 1400."[1] Mechanisms for avoiding unnecessary duplication of research will have huge payoffs for all concerned.

General architecture

A partially centralized, partially distributed system reaps the full advantages from each method of organization, without the disadvantages of only one method. What the world needs is a centralized summary and index of all the world's genealogy, with links to unlimited amounts of enrichment data residing on a distributed network of computers.

Database structure
A critical part of the system design is a database structure that is sufficiently flexible to accept the work of millions of researchers, and, through peer review processes, allow gradual integration and improvement until a nearly error-free, non-duplicative database is the end result. It avoids both the massive fragmentation and duplication of some of today's systems, while also allowing for legitimate alternate interpretations of original records.

Incorporating source records *is* the main process
In most cases the data presented would be from the original records themselves, with library references or direct links to those original records in their text or image form. This should make unnecessary any elaborate presentation of data to justify any particular data element on the final name record. Indexing together by individual name all the various kinds of source records also provides other valuable historical and sociological research options.
As one example, the 1.1 billion census entries for the 250 million deceased U.S. residents should mean there are about 4 entries for each person, giving good cross-checks of accuracy. Entering that data in a coordinated way is a far smaller task than most people would guess, and the payoff would be immense. For example, if 6 million Church members each spent one day a year for 4 years doing data entry, 32 hours in all, entering only 6 names per hour or 50 names a day, the entire task would be done. Everyone could contribute, without the steep learning curve and huge amount of time required for a person to become proficient with today's methods. Instead of mass production methods, we still use the apprentice and master craftsman techniques. These methods may be individually satisfying, but are hundreds of times less efficient than the modern alternatives.

Genealogy and temple work
The distinction between a temple work system and a genealogy system needs to be made clear, with the temple work system being viewed as a subset of the larger system needed for both efficiency and the accuracy of the data finally delivered to the temple work system. The current temple work system assumes the existence of the other more general system but does not now supply it.

Gaining the world's participation
Gaining the world's participation in a general system would be extremely valuable. Everyone in the world is interested in his or her own genealogy, but only a few of us are interested in the temple work part. At 3 million a year in temple ordinances, it would take 2,000 years to do the 6 billion dead for which we have records. We might wonder why we should ever need to assemble more than 3 million names a year, but there are many other reasons to collect the 6 billion names. As soon as one of our surname lines takes us to another country, the difficulty of doing further research may stop us, although it would be easy for a resident of that country who might share our ancestry, and who might be happy to help in the process if efficient means were available.

Conclusion – Doing all the world is the cheapest way to do our "own"
Remarkably enough, using the concepts just described, Church members could get *all* the world's genealogy work done more easily, quickly, and cheaply than if those same Church members focused all their resources on doing only their "own" genealogy work, which inevitable connects with the rest of the world. This will happen because mass production methods can be used, plus many outsiders will want to help in the process. These economic efficiency calculations will become even more true as the Church grows much larger.
It is a marvelous and appropriate thing that genealogy research or family history research can be done most efficiently if done by everyone together. The most pleasant way really is the best and most efficient way. This startling realization should show us that what we thought was practical self interest or Church interest in focusing on our "own" families is actually the least efficient way to proceed. This should encourage us to use our expertise to help the whole world do their genealogy work as we help ourselves.

Biography

I have spent 30 years in technical data processing work, including four mammoth “billion dollar” projects. I am a BYU graduate and also have two law degrees. Full resume available on website www.genreg.com at end of first document listed there (64 page pdf).

I coded the prototype in Visual Basic Script (ASP) using the Microsoft ACCESS database engine. Only a few simple functions are directly visible on the prototype website without a narrator to demonstrate it, but the internals (40,000 lines of code) do all the important functions needed for a worldwide system.

Footnotes

[1]Steve Olson, “The Royal We,” *The Atlantic Monthly*, May 2002, pp. 62-64.

Kent W. Huff
1748 West 900 South, Spanish Fork, Utah 84660
801-798-8441, huffkw@juno.com, www.genreg.com

To be successful,
we need all elements to be correct and complete
A partial list:

Theory
Goals
Policy
Legal
Efficiency

Public Relations
Economics
Doctrine
Quality
Data coverage

Online and offline methods and processes
Minimize duplication, maximize cooperation
Technology

The final System Design is determined by all these factors

All needed features are available.

Massive Cooperation vs. New Technology For a One-time Process

Some people may assume that a few bright people have to solve most of the genealogy technical problems for everyone concerning record conversion and accessibility.

But that isn't a necessary assumption. Well-organized cooperative methods are great solutions in some cases.

Converting old records to computer electronic format could be viewed as a one-time historical event. We just want to catch-up the old records technology to current technology. There may be no need to put a huge effort into creating special tools, if those specialized tools will not be of great value after that one-time catch-up process.

Communication - Email Replacement

The **central database** allows everyone to know instantly and continuously what everyone else has done and is doing. It has the effect of providing or **replacing billions of emails** each year to those trying to coordinate their efforts with other researchers, known and unknown.

Improve Communication of researchers:

- **ReplaceThe end of most E-mails (and most Internet searches)**
- **Replace most current laborious email traffic with the new system**
- **Get exact information instantly**

With all the data in a central database, one can quickly see what any researcher has done and is doing now, or check the current status of a particular family line. One can get exact and complete information instantly rather than wait for days or weeks for the less complete data that might be sent in an email. This will have the effect of **replacing billions of emails** (actual and intended) each year among those trying to coordinate their efforts with other researchers, known and unknown.

Emails will become the exception instead of the rule

Large Scale Cooperation is the Key

Large scale cooperation and **mass production** techniques means that we don't need further technical advances. (Of course we would always be happy for more technology, but it's not required)

One major technical concern is the **conversion of old records** into a more usable form. But this is a **one-time task** and much of it will probably have to be done **manually** anyway.

We should first put our technical efforts into **creating the new system** for holding the data, and then re-assess whether other records conversion technology is needed. I expect it to be unnecessary.

Don't let concern about the old records stop us from building the new system needed. The data will flow in, almost by itself.

"Build it and they will come."