Towards Effective, Real-Time Collaboration In Genealogy Research

Scott N. Woodfield
Department of Computer Science,
Brigham Young University
Provo, UT 84602 USA

1 Introduction

Collaboration in genealogical research provides two primary benefits. It reduces the amount of redundant work. It also provides investigators with current information which allows them to do their work more efficiently.

Unfortunately, genealogical collaboration is not as effective as it could be. There are several problems. The first is conceptual impedance. That is, prospective collaborators find it difficult to share information because their conceptual models differ. The second problem is lethargic information exchange. Once new information is found, it takes too long to send to other interested parties, and, once received it takes too long to incorporate it into the recipients own database. The third problem is lack of coverage. Many collaborative groups cover only a subset of all interested researchers. When information is broadcast it only goes to the connected members of the group. All others are excluded. The last problem is scalability. In many cases the technology supporting the efficient exchange of information does not scale to large groups.

2 Common Forms

Efficient and timely communication is the key to good collaboration. There are three common forms of communication used in genealogical research. First there is point-to-point communication, second, distribution list communication, and third central repository based communication. Each has its strengths and weaknesses with respect to the collaboration problems mentioned above.

Point-to-point communication supports pair-wise collaboration. Usually we have few impedance problems because we communicate with a single person who views genealogy as we do. We often do not communicate with those who view things differently. This often leads to coverage problems. Any information we may discover will only be shared with a few other researchers. All others are excluded. Point-to-point communication is often lethargic for one reason. The sending (writing) and receiving (reading) of information is manual. Although transmission may be nearly instantaneous, the writing and reading make communication slow. Point-to-point communication is not scalable. If there were a thousand researchers interested in the same person, we would find it difficult to keep in constant contact using point-to-point communication.

Distribution lists are another technology that supports collaboration. Email-based distribution lists have become especially popular. Distribution lists can reduce the coverage problem. If all those interested in a particular person register for a list, then, when anyone discovers new information and mails it, all interested parties will receive it. However, distribution lists do not improve lethargic communication and often increase impedance mismatches. They also do not scale well. A person with a few individuals in their genealogy can track a few mailing lists. However, when their database reaches thousands of people they soon find it impossible to track thousands of mailing lists.

Central repository based communication requires people to deposit their information in a central repository and then search that repository for any new information they may be interested in. As with distribution lists the coverage problem is reduced. Lethargic communication is still a problem because users must manually upload, search, and download information. Impedance mismatches are also an issue. If the information model of your database does not match that of the central repository it is difficult to exchange information. Last, central repositories are only linearly scalable. Double the number of people and you must double the size of the repository. Another problem with a central repository is control. If the the owners go out of business or start charging money, the users have no recourse.

3 A Peer-to-Peer Virtual Database

With advances in conceptual modeling, improvements in computer hardware, and always-on internet connections we can create a collaborative environment for genealogical research that will reduce or eliminate many of the mentioned problems. We propose a peer-to-peer virtual database as the technology for this collaborative environment.

Conceptually, the virtual database is a single, large database holding all genealogical information. It does not reside on a single centralized machine, but is distributed among all interested researchers. Each researcher's computer contains a local genealogical database that is a subset of the virtual database. It only contains information for those ancestors or descendants that the researcher is interested in. We assume that each person's computer (or proxy), is always connected to the internet. Also, all information is linked at the person level. That is, all researchers interested in person A form a collaborative group. If a researcher has a thousand people in their local database, they would be members of at least a thousand collaborative groups. Each group would be linked using peer-to-peer rather than client/server topologies. This provides faster, more reliable, and scalable communication. Information would be transmitted and received automatically. When a researcher records any new information about person A in their local database, it is automatically broadcast to all others in the collaborative group and it is used to semi-automatically update the recipients local database. In this manner communication becomes real-time.

This organization has several benefits. As the virtual database reaches critical mass, all researchers need only link to one source for current information. As more join, the coverage problem diminishes. Because of the automatic broadcast and semi-automatic update, we can reduce communication time to seconds. Because of the peer-to-peer topology scalability is not a problem. We suffer only logarithmic speed decreases and actually gain in reliability as we make linear increases in group sizes.

This technology does not solve the impedance mismatch problem. For that we propose that users not access their local database directly but instead manipulate the database through a view. The view presents the information in a form that the researcher prefers. The underlying software maps the view to the database. Thus, communication takes place using the database's conceptual model, but the user sees the information from their preferred perspective.

A peer-to-peer virtual database provides other advantages. Since information is highly replicated, there is a built-in backup mechanism. Since there is no central server, there are no ownership or control problems. Last of all, the view mechanism allows us to deal with conflicting information in a way that is natural.

4 Conclusion

We can do better genealogical research if we collaborate effectively and and in a timely manner. We believe the the proposed solution is more effective because it provides good coverage, is more reliable, is scalable, and provides a means for those with different perspectives to communicate. Because of the peer-to-peer network we can communicate in real-time.