A Peer-To-Peer Network Protocol for Genealogical Data

Conan C. Albrecht, Douglas Dean, Robert B. Jackson,
Stephen W. Liddle, and Raymond D. Meservy

E-Business Center
Marriott School of Management
Brigham Young University

Genealogy Network Transfer Protocol (GNTP) is a new protocol that allows searching and transfer of genealogical information through a peer-to-peer network. It provides a common interface that allows genealogists to publish information directly and search information being shared by fellow researchers. This model overcomes the traditional time delays and management problems associated with today's publication methods. The LDS Church's Family History Department has committed to help by providing a link on the FamilySearch.org site as well as opening the Church's large genealogical database to the network.

Information sharing offers great economy to genealogy researchers. Genealogists enjoy working together and sharing common lines with each other. Since all benefit from the sharing of genealogical information, researchers are normally more than willing to post data, whether that be to the web, to the LDS Church, or to another recognized repository.

GNTP is a natural development in such an environment. It is an enabling technology that allows genealogists to share data with each other in an automated, real-time architecture. GNTP allows exact searching for names, publishes GEDCOM directly, supports researcher control of data, automates publication, and potentially opens the boundaries between previously-closed organizations.

Limitations of Current Techniques

Despite the desire and need for information sharing, the disadvantages associated with current methods and technologies limit the amount of sharing that can be done. The following paragraphs describe the limitations of three current methods: submitting to the LDS Church repository; submitting to other, smaller repositories; and publishing via the World Wide Web.

First, the LDS Church manages the largest repository of genealogical information in the world. However, much of this information remains unsearchable because it is simply too voluminous to house in today's index-centric databases; it remains sealed in the granite vault in Salt Lake City. Family History Centers and FamilySearch.org help to alleviate this problem, but much of the data still remains unsearchable. The Church is currently employing groups (including BYU's Computer Science department) to solve this problem with new indexing algorithms and data media. Hopefully the entire data warehouse will be searchable in the near future.

However, even if the technological indexing problems are solved, a larger and more inherent problem exists with the central repository model. It concerns the management of such a large collection of data. The Church has limited ability to verify the accuracy and validity of submitted data. The people best informed (those who discovered and researched the names) and able to resolve conflicts and problems immediately lose control over their data when they submit it!

For example, consider two researcher relatives, Person A and Person B, who each find an ancestor named Craig Jones. Person A submits Craig's name with a birth date of 1750. Person B also submits Craig's name, but with a birth date of 1753. Is this the same Craig Jones? Or two Craig Jones with different birth dates? The only people qualified to make such a decision just lost control of their data as they submitted it! A simple solution is for the two researchers to discover each other, collaborate, and correct the data or make helpful notes. This discovery often happens, but only after the data is published and several years have passed. This management problem will only increase in the future and the volume of data and submissions increases.

Finally, the sheer number of names submitted to the central repository slows the process of publication. Therefore, several months or even years pass before submitted data is visible to the public. While it is true that FamilySearch.org has shortened this time required for publication, it still remains an issue in a world that provides immediate availability in so many other fields.

A second current method is submission to other repositories such as Ancestry.com. These repositories provide much quicker submission-to-publication times and often give the researcher control over the data. However, the existence of multiple, smaller repositories forces genealogists to search many different locations for names. Researchers must know about these other repositories and must understand each one's different interface. Many genealogists are not advanced enough technically to make effective use of these multiple repositories.

Finally, many genealogists publish their records to the World Wide Web (WWW). This worldwide Internet allows their genealogy to be indexed by web search engines. However, this method is limited because researchers must visit each site to determine applicability to the ancestry lines they are searching. This manually-intensive process limits the effectiveness of genealogy published on WWW sites.

Some may argue that specialized search engines can and have been created to scour the WWW for genealogical information. While this is true, current techniques for "page scraping" do not allow correct gathering of information on many Internet sites. Genealogy is published in unique ways and formats on each web site, so that even advanced artificial intelligence techniques produce only marginal results. Page scraping is an exciting research area with great potential future benefits, but it is not a viable solution today.

These limitations in current methods for genealogy publication adversely affect the effectiveness of genealogists. Most researchers still rely on decades-old techniques of manual library research, microfilm extraction, and person-to-person e-mail and other interaction.

GNTP: A Common Protocol for Genealogical Publication

GNTP is a new protocol and related architecture that enables real-time, ad-hoc sharing and transfer of genealogical information via the Internet. It allows the transfer of GEDCOM (or other formats) directly between peer-to-peer computers and enables efficient searching of the same. The GNTP design has been called by some to be the "Napster of Genealogy" for this new decade.

GNTP creates a peer-to-peer network of genealogical nodes, indexed by a second, smaller peer-to-peer network of directory servers. Figure 1 describes the proposed layout of this network:

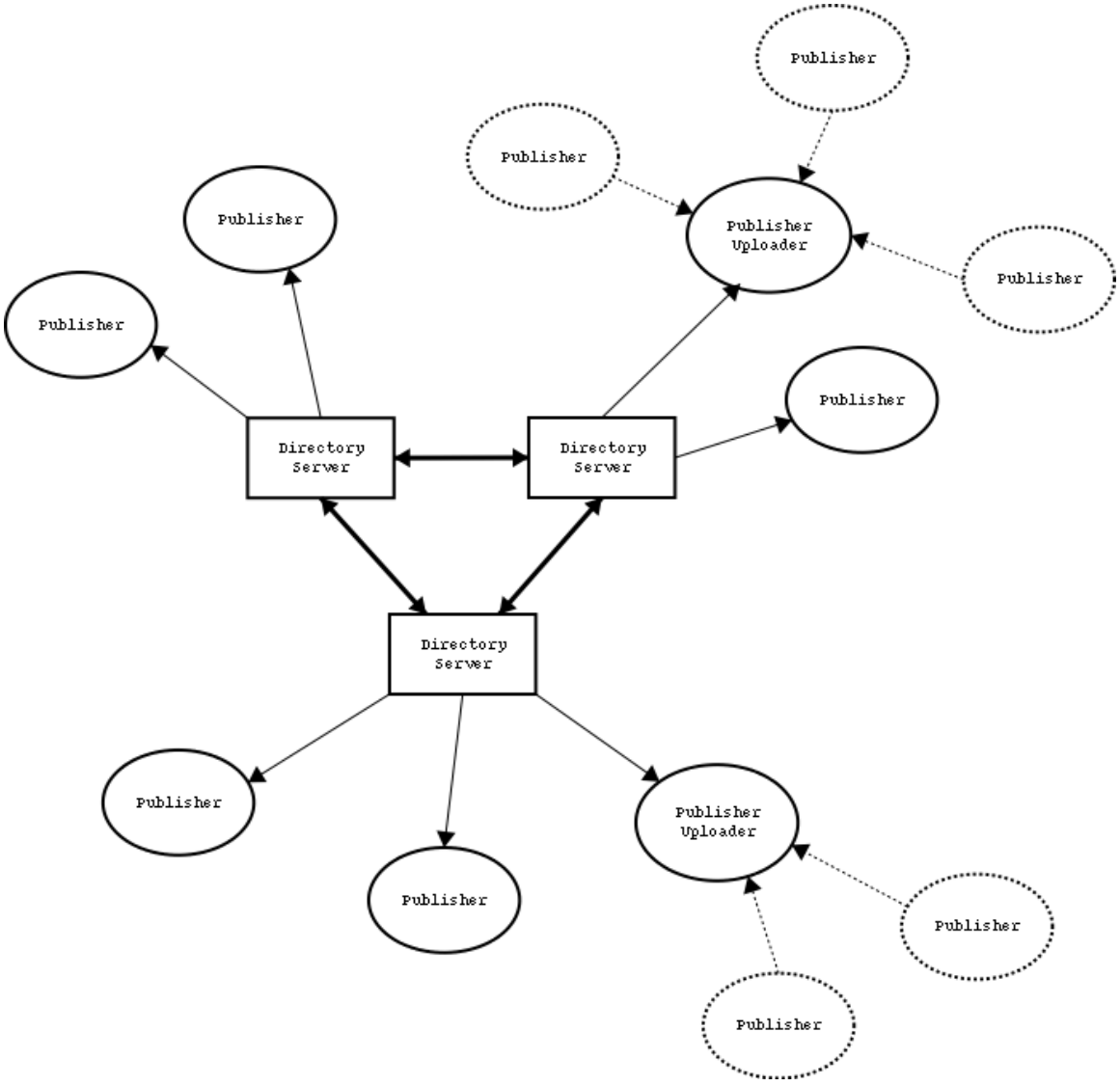

Figure 1: Network Architecture

As shown in Figure 1, GNTP defines two related peer-to-peer networks: an inner ring of directory servers and an outer network of publishers. The directory servers cache meta-information about the data contained on the publishers connected to it. This is accomplished by publishers registration at startup. Initial clients prototypes will register automatically with BYU's directory server (although this can be changed in a properties dialog) to allow less-advanced users publish information without needing to understand the network architecture. After the client registers, the directory server responds by querying the publisher (using GNTP) and caching meta-information such as names and/or locations. Publishers contact their directory server periodically to keep the network updated.

GNTP remains architecture-independent as to the types and models of connections between directory servers. It also does not specify the amount of information cached at the directory server level. This allows future participants to innovate and add value to the central directory ring.

Publishers represent individual nodes or users on the network. Users load their GEDCOM (or other formats) into their client through a File | Open... dialog. Clients then listen in peer-to-peer fashion for queries submitted from other clients. The "Publisher/Uploader" circle in Figure 1 represents a publisher that is also willing to accept uploads and publish data from clients without full-time connections to the Internet. The protocol includes commands for automatic uploading and updating of data to Publisher/Uploaders to allow less-advanced genealogists to publish via proxy. These users are represented in the figure with dotted lines.

When a user submits a query to the network, the client first connects to a known directory server (such as the one that will be hosted by BYU). The directory server responds with the IP addresses and ports of all publishers it knows about that may be able to answer the user's query. The initial directory server also returns other known directory servers, with whom the process is restarted. The client then connects to the returned publishers directly and queries each one in turn. The client-to-client connections make the network follow a true peer-to-peer model. The directory servers exist to make the peer-to-peer connections more efficient by directing clients to peers that are most able to answer their queries.

Since peer-to-peer networks experience network effects–the total value of the network is proportional to the number of nodes on the network–the LDS Church database will be one of the initial nodes. Therefore, its node will be a specialized publisher that serves as an interface to the genealogical information published on CDs at Family History Centers around the world. We expect the availability of this data will help the network gain usefulness quickly.

The GNTP protocol and related architecture provide many benefits above and beyond today's publication methods. These are described in the following list:

- Since GNTP defines a searching query language, it allows for exact searching of names, locations, events, and dates. It does not require any scouring, so the searching algorithms are simple and less intensive.

- While GNTP uses TCP (Internet) connections, it is not web-based. Therefore, there is less worry about illicit information being posted to the network than there is with the WWW. The only information transferred is raw genealogical formats such as GEDCOM.

- GNTP retains control with the users who discovered the information. These people are able to correct, modify, and add to the data with very little overhead. The source information transferred with the raw data helps researchers find each other and resolve conflicts.

- GNTP allows researchers to work together in real time, with no time delays between submissions.

- The inclusion of automatic publication and searching in GNTP allows less-technically-advanced users to participate and even publish their information through other known publishers with very little effort or understanding.

- GNTP potentially opens boundaries between organizations. Once the network gains usefulness through the publication of many sets of personal data and the inclusion of the LDS Church's name database, other organizations such as the Vatican and others (who house large unique data sets) may be pressured to share their data as well on the network. The open and free nature of the protocol ensures it remains free for public searching and publication.

Project Status

The GNTP 1.0 protocol its related searching query language were designed early in 2001. State machines and protocol definitions are available from BYU's E-Business Center. The reference implementation, include a directory server and publisher client, is currently being developed. The reference implementation and its related web site will become live July 1, 2001. The LDS Church's Family History Department will provide a link to the experimental site on FamilySearch.org, so we expect a considerable initial user base starting in July. BYU's E-Business Center will host the site and initial directory server for one year to allow for research and testing of the experimental network. If the project succeeds, the LDS Church will assume responsibility for hosting and development in July, 2002.

Involvement In The Project

Several areas exist for new participants to render help. First, the protocol has remained architecture independent with respect to the way the directory servers are connected. Research and implementation of architectures that will support efficient searching between these top-level nodes is needed.

Further, several initial test clients (i.e. end-user publishers) need to be created, including clients for Windows, Mac, and Unix/Linux. Some clients may be more advanced than others,

allowing different levels of users to participate in the most effective manner.