Global Genealogy Network
Extended Abstract

Eric Jarvi (eric@jarvi.com)
Nicholas Leippe (nick@byu.edu)

March 27, 2001

Abstract

The mission statement of our working group and the proposed architecture of the network are presented. The full paper should include a draft version of GML, as well as an enumeration of unsolved problems. This is a work in progress. If you would like to participate in our discussion of these problems, please contact us by email.

1 Mission Statement

Build a distributed network capable of storing all known genealogical data and suggesting links.

2 Network Architecture

While we are proposing a new platform, we do not intend to reinvent the wheel. This system is designed to leverage existing W3C standards such as HTTP and XML, the Unicode standard, and emerging standards such as that proposed by the XML-Signature working group. Existing data formats such as GEDCOM will also be able to be represented and stored by this system. By building on existing standards, we hope to provide equal access to participants regardless of their geographical location or native language.

The global genealogy network is composed of four conceptual layers: the source catalog, fact catalog, individuals and family relations catalog, and parallelized autocompletion. The later layers build on the foundation provided by the earlier layers. The rest of the paper describes the details of each of these layers.

3 Source Catalog

The source catalog contains digitized versions of source documents, such as photos, scanned documents, and audio/video clips. Personal compilations such as those represented in GEDCOM and PAF formats may be present. Web pages might also be stored at this level. For example, a web database of microfiche images, similar in concept to terraserver.com, which returns a specific image given a specific URL might be represented at this level. In fact, any digitized source accessible through a URI (such as HTTP, HTTPS, FTP, or future protocols) may be included in the source catalog.

While this may be conceptually thought of as one catalog, it is not stored on one single server. It is a cloud of many participants which each store a subset of the entire source catalog.

4 Fact Catalog

The fact catalog contains files in an XML format named GML (Genealogy Markup Language). These files provide semantic metadata about the facts contained in sources in the source catalog or sources which are stored offline in a digital or analog format. A GML file is broken into two sections: source and facts, similar to the way an HTML file is split into head and body elements.

The *source section* contains the following descriptive information about the source:

Categorization The type of source. For example, birth record, sickness or injury record, death and burial record, marriage record, divorce record, biography, family history and genealogy, or GEDCOM compilation.

Retrieval Instructions Information on how to retrieve a copy of the source. Multiple retrieval instructions may be associated with each source. For example, the source may be available online in the source catalog, through traditional methods such as postal mail or a site visit, by personally contacting the owner of the source by phone or email, or any combination of the above.

Copyright If access to the source is restricted or limited due to copyright, contact information and copyright notice would be presented here.

Authentication A digital signature of this GML file, and where possible, a digitally signed hash of the source. This will be based on work currently being done by the XML-Signature group. This makes it possible for users to self-moderate submissions and build trusted user profiles which can be used to make assertions or warnings about the quality of data submitted.

The *facts section* contains only the genealogical facts (such as names, dates, places, and events) present in the source document. By separating facts from creative works, users may have some degree of legal protection, as long as they are not in violation of any license agreement they may hold with the provider of the original source. In a 1991 case, Feist Publications, Inc. v. Rural Tel. Service Co., 499 U.S. 340, the U.S. Supreme Court unanimously ruled that copyright protection did not extend to compilations of facts, and emphasized that "... no one may claim originality as to facts."

These facts may be linked to each other with different types of links. For example, lineage links to represent lineage-linked data models such as GEDCOM. This linked, semantic metadata facilitates automated indexing of facts. For current related research in this area, see the Semantic Web Activity going on at the W3C. These GML files are URI-accessible in the same way as the source catalog.

5 Individuals and Family Relations Catalog

This can be thought of as an advanced search engine designed for genealogical data. It indexes and caches the information contained in the source and fact catalog. This makes it possible to search for individuals and investigate compilations of family relationships that have been submitted by other participants. The goal of the Individual and Family Relations Catalog is not to form one single family tree. The goal is to provide efficient access to data compiled by other users and stored in the source and fact catalogs.

This area should be capable of accepting maintenance notifications. For example, notification that a GML file has been deleted, moved, or updated. In this way, maintenance of the catalog can also be distributed.

6 Parallelized Autocompletion

This can be thought of as the application layer. Once the three catalogs are in place, automated computer indexing of the information can begin. One way of doing this is to have a distributed network of client applications (similar to distributed.net or SETI@home) retrieving GML files, traversing links, proposing potential lineage links, or submitting maintenance notifications.