Genealogical Data Mining

by Ben Franklin

Definition of Data Mining

The expression "data mining" is now widely used in the Information Systems arena. However, the exact meaning of the expression is still not widely known. It implies "digging through tons of data" to uncover patterns and relationships. Data mining:

- Is a tool that supports research and allows new assertions to be made by disclosing previously undisclosed details in large amounts of data
- Integrates the results of research in database management, statistics and artificial intelligence
- Opens new horizons where the traditional methods are not adequate for efficient data analyses

Data mining in a genealogical context is the technique of rapidly acquiring new information, without stopping to categorize, analyze and make assertions. Data mining is used to gather, collate, and organize information so that it can be more effectively searched.

Throughout this presentation, I will be mentioning the successes that I've experienced in pursuing a single-surname database of Franklins. I do not assume that you are interested in this lineage, but I will use it as an example.

"Bad" data mining

Another type of illicit data mining is used to surreptitiously gather e-mail addresses and other personal information for malignant use by spammers, etc. This is not the sense of the term as used in this presentation.

Acquiring "Data"

In order to leverage the analysis of our data, first we must have something to analyze. Historically, the way we acquired data in our computers was that we would either manually type it into the computer ourselves, copy it from a CD, or find and download it from the Internet. These methods are OK to start. Allow me to suggest a few more ways of acquiring data:

Convert Hardcopies to Data Via OCR

Optical Character Recognition (OCR) software can be a very useful tool for turning mounds of paper into searchable information in our computers. This topic will be discussed further later in this presentation.

NEHGS

Periodically the New England Historical and Genealogical Society (NEHGS) gives access to some of its on-line databases. For instance, during the 2004 Thanksgiving Holiday the NEHGS offered free access to their database. I searched their on-line index for Franklins, and downloaded 1,600 GIF images representing pages of old issues of the NEHGS Journal, that pertain to Franklins.

JSTOR

JSTOR (Journal Storage Project) is an ongoing project to develop a digital library in support of the arts and sciences. It initially consisted of about fifteen journal titles in the areas of economics and history and contained approximately 750K journal page images. These journals are fully searchable. Those of you who are members of an affiliating institution, such as students or faculty of a university, can take advantage of this collection of scholarly journals.

I first became interested in this when I found that digital copies of the *William and Mary Quarterly* can be downloaded there. This is a very useful journal for early Virginia history and genealogy. I was able to download 90 pages that pertain to Franklins in *W&M*, and from the remainder of JSTOR I found about another 200 pages of various biographies and other journals with articles about Franklins.

Heritage Quest Books

Heritage Quest has a large number (25k+) of books that are available to download on-line. These can be downloaded 50 pages-at-a-time and are in (graphic) PDF format. Thus far I have downloaded about 25K pages of Franklin material from this resource. Heritage Quest can be accessed for free by anyone with a valid Durham County or Orange County Library card. See **`http://www.durhamcountylibrary.org/dclhqo.htm`**.

BYU's Digital Archives

BYU's Family History Collection is growing rapidly. The currently (15 Sep 2005) have 3944 books that are available to download on-line. These can be downloaded one page-at-a-time and are in (graphic) PDF format.

Copy Data from the Census Index

The census index can be copied and saved to your system for later searches, however unless you are going to use it as a framework for more information or in ways that cannot be used on-line, then it might be better to just access it from the website.

You can easily reformat this data, for example:

- To use it as a framework for census abstracts
- To create a spreadsheet of census data
- To build a GEDCOM.

To do this, you will need to use your word processing software or text editor (such as ***TextPad***). This is quite difficult to explain in detail, but will be demonstrated during the presentation.

Acquiring Hardcopies

Identify the books that interest you, using PERSI, the FHLC, references in others researchers' data, etc. Then you can order the film at the FHC, go to a local library, borrow the material via interlibrary loan, etc. [Getting books and films is a beginner topic...]

However, in a "data mining" mode, you do NOT stop to read the material. Merely copy it for later analysis. Too many people travel thousands of miles to get to the FHL in Salt Lake City, and then spend their time reading books. No. Focus on copying. Thorough analysis requires a lot of time. It

is time that you can ill afford at the FHL, unless you live very close to it.

Dealing with Images

Irfanview

This freeware application can be used to view convert, or print almost any type of graphic images. It is very, very useful. Download it at:
`http://www.irfanview.com`.

Print it and OCR it

When you have information in digital form, such as GIF images of pages of the NEHGS, for instance, OCR accuracy may improved dramatically if you print a hardcopy of the page and then scan it and OCR it. This is despite the fact that the OCR program can read GIF files directly. The problem is that in order to save disk space and improve download times, the original images are scanned at a very low resolution and the step of printing the image actually improves it.

Popular OCR applications include Scansoft's *Omnipage Pro*, and Scansoft's *TextBridge*.

More About PDFs

You will find a number of on-line sources are available in Adobe's Portable Document Format (PDF). This can be viewed and printed using the Adobe Acrobat freeware application that can be downloaded from the Adobe website. There are two basic forms of data in a PDF - graphic and textual. Most of the PDFs that you can download, such as from BYU or Heritage Quest, are in purely graphical form. That means that the text cannot be searched. You will need to OCR this data to convert it to searchable data.

Finding Stuff in Your Data Mine

Windows Explorer Search

For those who use Windows, using the Search capability of Windows Explorer provides a rudimentary search function. This will enable you to find specific text strings within your data mine.

GREP

The GREP command comes from the UNIX world where it is used to search within files for data in a flexible way. This is much more powerful and faster than Windows Explorer. There are several version of this software that can be downloaded from various site on the Internet. My favorite is ***wingrep***. It may be downloaded from:

`http://www.wingrep.com/`

Google Desktop

Like the popular web searching site, "Google", there is an application that you can download to your personal computer called Google ***Desktop*** is one of the most powerful and flexible search applications for your the data on your own computer. It gives convenient access to information on

your computer and from the web. It is a desktop search application that provides full text search for your email, computer files, music, photos, chats and web pages that you've viewed. By making the data on your computer searchable, Google Desktop puts your information easily within your reach and frees you from having to manually organize your files, emails and bookmarks. It makes searching your computer as easy as searching the web with Google.

- Email in various client formats, including: Gmail, Outlook, Outlook Express, Netscape Mail, Mozilla Mail and Thunderbird
- Files on your computer, including text, Word, Excel, PowerPoint, PDF, MP3, image, audio, and video files. You can even search your media files by meta-tag: for instance, by artist name and song title, not just the file name.
- Web pages you've viewed using Internet Explorer, Netscape, Mozilla and Firefox.
- Chats from AOL 7, AOL Instant Messenger, and MSN Messenger