Digital Representation in Genealogy

We live in a digital age. Almost everyone carries around at least one electronic device, such as a smart-phone or tablet, and many family historians have a laptop or home computer. Although we live in a golden age of access to information, including information of genealogical interest, this information is frequently available only in inconvenient forms. This article is an effort to explain both to the users of digital databases, but even more importantly to the publishers of these databases, the advantages and disadvantages of various representations of information.

Whether distributed as traditional paper books, as digitized transcriptions on CDs or DVDs, or on web sites as either displayed pages or downloadable datasets, most sources organize data collected from original sources. Some are pure textual sources, for example correspondence and books. Others are tabular documents that represent answers to a specific set of questions. This includes surveys, censuses, birth certificates, marriage licenses, death certificates, municipal directories, phone books, cemetery registers or transcriptions, and lists of electors among many others.

Traditionally these datasets were recorded on paper, whether hand-written, type-written, or type-set. Paper documents are durable, we have many paper documents from thousands of years ago such as the Dead Sea Scrolls or Egyptian hieroglyphic scrolls. As long as paper is protected from damp, insects, and fire it will last essentially forever. However paper documents are difficult for humans to search, and impossible for computers to read, except for those documents that have been typed or type-set. Even if literally converted to electronic documents by text recognition software or human transcription the documents remain of only limited usability because the original documents were not designed to be searched. An interesting example is the correspondence of Canadian Prime Minister William Lyon McKenzie-King. Because his handwriting was idiosyncratic his secretary typed out a copy of everything he wrote. This typed version has been read by text recognition software and used to create a searchable index of the hand-written correspondence.

Despite the wonderful results that current search engines, such as Google and Yahoo, can achieve they are limited by the fact that the documents were never designed to be searched efficiently.

Most documents created during the last 30 years are the output of word processor programs, such as Microsoft Word. This is largely a consequence of the familiarity which almost everyone has these days with creating documents using a word processor as part of their work day. Word processors are specifically designed to represent traditional paper documents with the final objective of actually creating such a paper document as the final output. All current word processors use a user interface model called What You See Is What You Get (WYSIWYG). That is the document appears on your computer screen as an image of the final printed paper document, complete to the use of black print on a white background, by contrast with the green print on a black background that was typical of the first generation of computer word processors.

In particular one of the fundamental features of all current word processors is that the displayed information is organized as pages. A sophisticated user knows how to set paragraph characteristics such as the widow and orphan settings which prevent a

paragraph being split between two pages in such a way that only a single line appears on one of the pages.

However in the transcription of historical documents this pagination may be counter-productive. For example in the digital representation of a historical document what is important is the pagination of the original document, not the pagination of the transcription.

This WYSIWYG feature also carries over into one of the most popular formats for representing documents on the web: Portable Document Format (PDF). PDF is a file format developed by Adobe Corporation to ensure that the electronic representation of a document looks on a computer screen exactly like the paper representation. PDF is necessary because different word processors, or even the same word processor running on different computers, represent a document slightly differently. As a specific example I asked a professional printing company to print a book for me and gave them a Microsoft Word .doc file. They loaded that file into their copy of Microsoft Word, and attempted to print the book, but the result had a different pagination from the one on my computer because I was running Microsoft Word on Windows, while they, like most graphics companies, were running the version on Apple iOS. One of the typical reasons for this pagination mismatch is that a word processor program uses the set of fonts that are shared by all of the programs on the computer, and the exact metrics (character dimensions) of a font depend upon exactly where it was obtained from. PDF avoids this by including the font definitions in the document.

PDF permits the distribution of accurate electronic versions of documents whose final representation is on paper. For example the layout of a brochure, or a magazine, or a slide presentation, must be preserved in the electronic version. However for the digital representation of transcriptions of historical documents PDF is an inappropriate form because it imposes a layout, including pagination, which obscures the structure of the original document. Another issue is that PDF files are large, precisely because they must contain the accurate information about layout, and the programs that permit viewing them are also large and complex, because real-world paper documents come in a bewildering range of formats. In particular the standard viewer for PDF, Adobe Acrobat, is much larger than the browser you use to access other information on the web. Just loading this viewer application can slow the operation of your computer. It is also a licensed software program which you must electronically sign an agreement for before you can use it. That is why Adobe Acrobat is never included along with the other software when you buy a computer, but must be downloaded and installed the first time you try to view a file in PDF.

All current word processors can also export a document as a web page, using the Hypertext Markup Language (HTML). This permits the contents of the document to be viewed without requiring the installation of the licensed PDF viewer application. However without PDF the presentation of the document loses the fine precision of presentation. In particular HTML does not even implement the concept of pages. Illustrations embedded in a document lose their physical relationship to the related text, although references to figures and other document entities become click-able links. A big advantage of HTML as a representation of document transcriptions is that it is indexed by the big search engines, and an organization can exploit search engine technology to index all of the HTML documents on their web-site. The biggest problem with HTML is that

images cannot be embedded as part of the web-page file itself, but must be stored in a separate web addressable location which is referenced by the main web-page.

While both PDF and HTML documents can be searched, neither of them provides a structure for the information. Both PDF and HTML are concerned primarily with the presentation of the information, not with its meaning. So relationships between related pieces of information are lost whenever a document transcription is presented using these popular methods. What do I mean by information relationships? For example in the text "John Smith (1848-1912)" each of the words is related in a specific way to each of the other words. That relationship is intrinsic to understanding the meaning of the text, but is not represented when the information is presented on the web either in a PDF or an HTML document. If you use a search engine and search for "John Smith 1848 1912" you will get back a list of all documents that contain each of those words, no matter where those words appear in the body of the document, not just documents about "John Smith (1848-1912)".

To publish a document such that the meaning and relationships of the separate pieces of the information is represented you need to use a format that represents a table of information. Each column of a table represents a particular attribute, for example a surname or a birth date. That is the answer to a specific question. Each row of a table represents a single instance, for example a single birth certificate, or a single row of a census page.

There are a bewildering number of ways to represent a table of information. This very wealth of choice may partially explain the unfortunate lack of exploitation of this capability by most web-sites and genealogical organizations.

For those who have worked with spreadsheets as part of their daily work, they may provide a familiar choice. You can easily create very large tables of information with a spreadsheet program, such as Microsoft Excel. A spreadsheet can be a excellent way to collect and organize information. With the collaboration features of Google Sheets, for example, a team of volunteers could collectively create a spreadsheet that organized a transcription of a large municipal cemetery. If the team of volunteers includes a technically sophisticated contributor the spreadsheet input can even validate the input, for example insisting that some fields must be numeric.

Spreadsheets have a couple of minor disadvantages.

- Organizing data is not the primary design goal of a spreadsheet program. A spreadsheet is primarily designed to implement book-keeping and financial analysis functions. Most cells in a typical business spreadsheet contain not values but expressions which act on information contained in other cells. For example a cell may contain a column total or a percentage of another cell.
- Spreadsheet programs are also concerned with what the final output looks like, such as what the borders around a cell look like, what font is used, what highlighting is used, and so on, so the final paper output will be more easily read. This is because one of the objectives of these programs is creating business reports.
- Spreadsheets do not implement the concept of columns of data which have a particular meaning very well. You can tell the spreadsheet program that the first line of the spreadsheet contains the column names, and it will then permit you to,

for example, sort the spreadsheet based upon the values in a particular column by name, but essentially the entire spreadsheet is treated by the program as a set of individual cells.

- A spreadsheet program reads the entire spreadsheet file into your computer and works on it in memory. This works well for typical business uses, but is impractical if, for example, you want to work on the 1920 Census of the United States of America which contains over a hundred million lines!
- The search capabilities of spreadsheet programs are limited. You can search a single column at a time for a specific value. You cannot, without a lot of sophisticated programming, search a spreadsheet for "John Smith born in 1848".

The alternative tabular representation is as a database table. An example of this is the Access component of Microsoft Office. Unlike a spreadsheet, which is normally modified as often as it is viewed, databases are expressly designed with the primary purpose of facilitating searches and updates are performed only very rarely. All popular database implementations implement a standard method of requesting a search through Structured Query Language (SQL),. The abbreviation is pronounced "sequel". SQL is an extremely powerful language for expressing even very complex search requests and retrieving a response in a remarkably short time. Unlike spreadsheets which can be no larger than the main memory of your computer, databases can be of any size because only the data you are actually using at a moment is loaded into the computer. Because of their potentially great size databases are not normally distributed as files to end user computers. Rather a user interface, frequently a web page on a browser, is all that runs in the end user device, permitting the user to issue requests for subsets of the contents of the database. This makes such a database accessible on any device with a browser, such as a smart phone, a tablet, a laptop computer, or even an e-reader. In the field of genealogy the largest such databases are those run by Ancestry and FamilySearch.org, but anyone can establish an account with a web-hosting company to run such a service for a few dollars a month. That is a lot less than what your Internet Service Provider (ISP) charges to connect you to the Internet.

If you do wish to distribute the contents of a spreadsheet or a database, you have the issue of choosing a file format to represent the information. Actual data files, such as .xls, .xlsx, .mdb, or .ods are undesirable because they contain too much irrelevant information, such as expression formulas and formatting instructions, as well as actual executable code which can carry viruses. The application can export all or part of the data as a PDF or HTML document, but then the information cannot be searched except as unorganized text. Most commonly spreadsheets and databases are exported as text files that contain only the actual textual values of the cells. The most common such format is Comma Separated Values (CSV) in which each line of the text file consists of values separated by commas. To permit a value to contain a comma, the value is enclosed in quotes. Another representation is Tab Separated Values (TSV) which is easier because it simply disallows the use of a tab character within a value, which is something that spreadsheet programs already do, since tab is used by the user interface to get to the next cell. A CSV or TSV formatted file can then be loaded into a spreadsheet or database application, including Google Sheets through your browser, and searched column by column. If distributed on a digital medium as a CSV file the file will automatically be opened in the default spreadsheet application when you click on it.

Frequently non-profit genealogical organizations are reluctant to put the results of their volunteers' effort up on the web for free. They have a number of concerns. Loss of control of the information if it is distributed electronically is one. However most conventional publications I have obtained from volunteer genealogical organizations have not properly asserted copyright, which means that they are already vulnerable to theft of knowledge. Reputable organizations will not knowingly violate even an implied copyright on any electronic data they acquire. Also family history societies look at the revenue the organization obtains through publication sales. However that revenue should be balanced against the costs of supporting the publication sales, such as duplicating costs, maintaining inventory, accounts with credit card companies for payment, and so on. As well there is the loss of potential advertising revenue that could be obtained from a web site, or the exchange of rights to access information between organizations. Most importantly access to on-line databases created by a genealogical organization can be made a privilege of membership, increasing membership revenue and advertising revenue from other association publications.