Spreadsheet Programs in Genealogy

A common use of computer spreadsheet programs is to organize data collected from original sources. A spreadsheet is a natural representation of many forms of data. In particfular any data set that represents answers to a specific set of questions can naturally be organized using a spreadsheet. This includes surveys, censuses, birth certificates, marriage licenses, death certificates, directories, and lists of electors in particular. Spreadsheets have also been used effectively to represent cemetery transcriptions, although the inconsistency in how information about individuals is represented creates some limitations.

Generally when you purchase a new computer it comes with a copy of an office suite. Since marketers are trying to minimize the price of their products this is generally the cheapest suite that they can obtain a license for. Typically this means that it is one of:

- a back-level release of Corel® WordPerfect Office. This is provided in order to entice you to later buy a more current release.
- The Microsoft® Works suite. Despite the marketing logo this product was not developed by Microsoft, and it is not compatible with Microsoft® Office, with the exception of the copy of Microsoft® Word that is included. Again the purpose is to entice the customer into later purchasing the much more expensive Office suite.
- Lotus Office. This is included only on computers bought through an IBM retailer.

All of these applications provide similar basic capability, which is all that you are typically going to use in genealogy. However they are all, to some extent, deliberately frustrating. So sooner or later you are likely to be considering an upgrade to a more comprehensive package. In this article I discuss three specific options:

1. Quattro Pro; part of the Corel® WordPerfect Office. The latest version, incorporated into WordPerfect Office X7, can be downloaded from www.wordperfect.com for Cdn$249.
2. Microsoft® Excel; part of the Microsoft® Office. The latest version, incorporated into Microsoft® Office 365, can be downloaded from www.Microsoft.com for US$99 a year (about Cdn$120).
3. OpenOffice, currently represented by OpenOffice 4.1.1 from Apache. The latest version can be downloaded from www.OpenOffice.org for free.
4. LibreOffice, currently represented by LibreOffice 4.4 from The Document Foundation. The latest version can be downloaded from www.LibreOffice.org for free. This is a community supported split from the an earlier version of OpenOffice.

History

To understand some of the idiosyncrasies of these applications a little history is useful.

The first popular computer spreadsheet program, introduced in 1978, was VisiCalc. It was conceived by Daniel Bricklin of the Harvard Business School. Although VisCalc introduced the idea of an electronic spreadsheet, it failed to take advantage of the opportunities of the swiftly changing marketplace of the early 1980s. One artifact of this program continues to be supported by all of its successors. If you start a field value with

an apostophe (single quote) it will disappear when you leave the cell. This is because VisiCalc used the apostrophe as a signal to left align the value in a cell. This can be annoying if you are trying to record a value that actually starts with a single quote.

In 1982 Lotus Coporation was founded to produce and market a spreadsheet program called 1-2-3 to emphasize its ease of use. Lotus 1-2-3 was the most popular spreadsheet application on the old DOS based computers. It provided a lot of features that made it an effective business tool, such as integrated plotting. The rights to Lotus 1-2-3 were acquired by IBM in 1995.

When Microsoft moved to focusing on the graphical Windows interface they did not have a large installed base of office suite users to worry about, and their primary concern in developing the components of their new office suite was to sell the new Windows operating system. Therefore Microsoft® Excel concentrated on getting the best possible graphical user interface on the new product, even if this resulted in a less convenient keyboard interface. This proved to be a wise decision, as Excel is, by far, the most successful spreadsheet program.

Quattro Pro is the direct lineal descendant of the Quattro spreadsheet program that was developed for MS DOS about 25 years ago. The name implied that it was allegedly superior to Lotus 1-2-3, which was the most popular spreadsheet application at the time. When it was adapted to run on Microsoft Windows a marketing decision was made to maintain as much compatibility in keystroke sequences as possible. Quattro Pro, and WordPerfect Office in general, are therefore particularly popular with users who like to do as much work with their hands on the keyboard as possible, rather than taking a hand off to work the mouse.

OpenOffice was developed by Sun Microsystems to provide capabilities similar to Microsoft® Office, but on a operating system independent platform. If Microsoft had been willing to adapt their Office suite to the UNIX environment they might have made OpenOffice irrelevant. OpenOffice is now supported by Apache, the same organization that supports the Apache web server that is the basis of almost every web site on the planet. Its objective is to implement a complete office suite that can run on any platform. If a user finds that it is missing some desirable capability, they can add it themselves, and then contribute the new capability to the market.

LibreOffice is an essentially identical product to OpenOffice, starting with the same shared source code. It was split off from OpenOffice mostly because the license conditions imposed by Apache when it took control of OpenOffice were problematic for some distributors.

All three products can open and save spreadsheets in the industry standard Excel format. The native file format for OpenOffice uses eXtended Markup Language (XML). The advantages of this decision have been recognized by Microsoft, which has announced that the next version of Microsoft® Office will also move to XML file formats. Using XML file formats reduces the size of files while making the delivery of new functionality easier. More significantly, perhaps, it means that 3rd party tools can be easily created to perform functionality that is not available from the office suite itself.

Auto-Formatting

Spreadsheet programs provide you with a lot of control over the appearance of cells. If

you do not explicitly customize a cell the program tries to guess from the data that you enter into the cell what formatting you would like. Sometimes this works well, and other times it does not.

Frequently when entering ages a census enumerator will express the age of young children as a fraction, for example "3/12" to indicate a child that is 3 months old. By default, however, all spreadsheet programs interpret this as a date! So as soon as you enter the value it is changed to "03/12/06", for example! In Microsoft® Excel you may specify an age field to have a custom formatting of "# ?/12". Doing so has the advantage that you can perform arithmetic and comparisons on the values. However this option is not currently supported by either Quattro Pro or OpenOffice. An alternative that does not require fiddling with the cell formatting is to express ages in months as, for example "3m".

The following table shows the results of entering particular values into cells that have been explicitly formatted as "text" or as "custom" with a format string of "# ?/12". Note that any fractional value is displayed as the closest fraction with a denominator of 12.

text	general	custom # ?/12
1.5	1.5	1 6/12
5	5	5
3/12	12-Mar	3/12
1 1/2	1 1/2	1 6/12
11/12	12-Nov	11/12
3.95	3.95	3 11/12

Many genealogical data sources contain dates. These cause a problem for all spreadsheet programs because, by default, they are recognized as dates and internally encoded so they can be compared, sorted, and subtracted from each other. However the internal format adopted for dates is the number of days since 1 Jan 1900. When you enter an earlier date, for example "15 July 1874" it is treated as a string and left as you entered it, but any later date is translated and then displayed in the format chosen for the cell. For example by default "15 July 1974" is translated and displayed as "15-Jul-74". Dates that do not include a year are assumed to be for the current year, which is not displayed. I generally explicitly flag the entire column as containing "text" to defeat this automatic conversion.

Note that OpenOffice differs from Excel and Quattro Pro in that it does support dates prior to 1 Jan 1900 and stores them internally as the appropriate negative number of days. That means that all dates subsequent to the switch to the Gregorian Calendar can be compared and sorted correctly in OpenOffice. However if you open such a file using Microsoft® Excel it will not be able to interpret the value.

Adding Comments to Cells

Frequently when transcribing data into a spreadsheet there is additional information that

cannot be represented as part of the value of the cell. After all the normal rule is that, as far as possible, a transcription should represent the document being transcribed. For example if the original document gives someone's name as “Jhon Simht”, then that is what should be entered into the spreadsheet, even if you are absolutely convinced that the original author intended to write “John Smith”.

Most spreadsheet applications provide a way to add comments to a cell. These comments will appear in a floating bubble any time the cursor hovers over the cell in question. In Quattro Pro and Excel these are called “Comments”, and are added by “Insert/Comment”. In OpenOffice these are called “Notes”, and are added by “Insert/Note”. Regardless they all inter work.

I have used these in a number of places to explain cells:

- Where the original says “Mrs. John Smith” I sometimes add a note “should be Jane Smith” or even “née Jane Jones”
- Where I know, from evidence outside the document in question, that the enumerator or clerk made a mistake, I enter the value as supplied by the author, but add a note with the suggested correction.
- Where the original is illegible I add a note “covered by ink blot”, “erased by enumerator”, or just “bad handwriting makes it unintelligible”.
- However where the author has scratched out an entry I indicate this by using the text formatting features of the spreadsheet to draw a line through the value.
- Where the author ignored published instructions on how to fill in the form, and leaving the author's erroneous value in place would impair the use of the data, I put a “corrected” value in the call, and use the comment to record what the author actually wrote. For example in an ethnicity column I might enter “English” as the value with a comment “actually 'England'”
- Where the value entered is now considered pejorative I “correct” the value while retaining the original in a comment. For example I adjust “Welch” to “Welsh”.
- Where the cell should only have certain values, but the enumerator has inserted something abnormal or additional. For example marital status should be “S”, “M”, “W”, or (rarely) “D”. In one case the enumerator entered the letter “M” but added, in fine print “Separated from wife”.

Auto-Completion

One feature of spreadsheet programs that can save a lot of effort is auto-completion. Auto-completion tries to “guess” the complete value of a cell based upon the first few keystrokes you enter. The algorithm for completing the value is different in each application. Indeed if a new spreadsheet program came on the market that exactly duplicated the behavior of an existing program it would probably result in a lawsuit.

To compare the effectiveness and intuitiveness of the algorithms in the various implementations, imagine that you have a column that represents a month. For simplicity we will assume that the only values that may occur in this column are: blank, “Jan”, “Feb”, “March”, “April”, “May”, “June”, “July”, “Aug”, “Sept”, “Oct”, “Nov”, and “Dec”.

For effectiveness let us consider the number of keystrokes required to enter each of the months into subsequent cells of the column.

On **Quattro Pro** auto-completion is case sensitive. That means, for example, that if you start a new cell with a lower-case "j" it will not match to any of the already entered values. The Quattro Pro algorithm matches to the most recent matching entry. This works well when, for example, entering surnames in a census, but not so well when entering given names or months, which are essentially random.

For this kind of data the Quattro Pro algorithm is less intuitive than some of the other products. The user cannot depend upon the exact result of a series of keystrokes, because the algorithm depends upon the history of what has been entered into the previous cells in the column, not just the set of values in the column. Starting a cell with an uppercase "J" will result in one of "Jan", "June", or "July". Therefore if you are typing while not looking at your computer screen, for example you are either looking at the keyboard or at the original source you are copying from, then you must type 4 keystrokes for all three of those months to ensure that you get the right month. Furthermore if you accidentally forget to hold down the Shift key while typing the first character, auto-completion will not work.

Microsoft® Excel requires you to type enough characters to uniquely recognize an auto-completion value. Excel is not case sensitive in recognizing partial values.

Excel's support for multiple spreadsheets within a file means that it does not use cell values that are separated by blank rows from the cell that you are typing into as part of the auto-completion algorithm. This can cause problems with data that has blank rows, for example many censuses. In this case I sometimes have temporarily entered non-blank values into a blank row just to get Excel to recognize that the entire worksheet is a single table. Although Excel is not case sensitive in auto-completion if it does not auto-complete a value, even if the complete value does match, then it does not apply the capitalization that you entered previously. In our example this applies to the month "May". You must type all 3 letters to distinguish it from "Mar[ch]", but having entered all 3 letters, if you forgot to hold down shift you get "may" in the cell.

OpenOffice has a possibly unfair advantage. It was developed more than a decade after either of the others and therefore could take advantage of a lot of user experience. Microsoft, for example, could not change its auto-completion algorithm without upsetting its hundreds of millions of loyal customers. The OpenOffice auto-completion algorithm is simply to match to the value that has the lowest sorting order. In addition the Tab key can be used to sequence in alphabetical order through all of the matching values. This is also the most intuitive algorithm because you always know exactly what result you will get for a particular keystroke sequence.

The overall effectiveness of the three algorithms are compared in the following table.

	Quattro Pro	**Excel**	**OpenOffice**
Jan	3 (shift, “Ja”)	3 (“Jan”)	1 (“J”)
Feb	2 (shift, “F”)	1 (“F”)	1 (“F”)
Mar	4 (shift, “Mar”)	3 (“Mar”)	1 (“M”)
Apr	3 (shift, “Ap”)	2 (“Ap”)	1 (“A”)
May	2 (shift, “M”)	4 (shift, “May”)	2 (“M”, Tab)
June	4 (shift, “Jun”)	3 (“Jun”)	3 (“Jun”)
July	2 (shift, “J”)	3 (“Jul”)	2 (“Ju”)
Aug	2 (shift, “A”)	2 (“Au”)	2 (“Au”)
Sept	2 (shift, “S”)	1 (“S”)	1 (“S”)
Oct	2 (shift, “O”)	1 (“O”)	1 (“O”)
Nov	2 (shift, “N”)	1 (“N”)	1 (“N”)
Dec	2 (shift, “D”)	1 (“D”)	1 (“D”)
Total	30 keystrokes	25 keystrokes	17 keystrokes

Organizing Data for Effective Auto-Completion

Auto-completion works best when there are a relatively small number of distinct values that are possible in the column, with few distinct values that start with the same characters. The set of month names used in the previous comparison is an example of such a set.

In a field containing a date within a year, for example the birth date field in the 1901 Census of Canada, you can run into problems even with the text format setting because of auto-completion. For example if you have previously entered a value “March 12” into such a column and you then enter “March 1” it will be auto-completed to “March 12”.

You need to pay close attention to see if this is happening to you. To avoid this I recommend always putting the day before the month, even if the original source has it the other way around. Since there are 366 different valid values, not counting blank and "don't know" values, auto-completion does not work well in such a column until you have entered a lot of values, so you will generally end up typing every value completely, although Quattro Pro and OpenOffice may correctly expand the month.

As an experiment I have tried dividing such a date column into separate month and day columns. This permits autocompletion to work on the month names, as described above. However the additional cursor movements to move through two columns means there is no saving in total effort.

In the 1901 census of Canada the birthplace column contains two pieces of information, the location and, for places within Canada, whether the birthplace was rural or urban. If both pieces of information are entered into the same column of the spreadsheet extra work is created with both Quattro Pro and Excel.

In the case of Quattro Pro as soon as you type the initial upper case "O", it will guess that you want to repeat the last value starting with "O", for example "Ontario r.". But if you actually want to enter "Ontario u." you must continue typing until you reach the "u". That is you must type 9 keystrokes (including the shift) every time the value changes. Alternatively you may type "O" then tab, which takes you to the end of the auto-completed value, and then backspace to replace the "r". That is 4 (or 6 if you include the period) keystrokes to change entries.

Excel will not do any autocompletion until you get to either the "r" or the "u". That is you must type at least 8 keystrokes on every entry starting with "Ontario".

In the case of OpenOffice it takes one keystroke, "o" to get "Ontario r." and two keystrokes, "o" and tab, to get "Ontario u."

If this single column is divided into two columns, one for the location and one for the urban-rural indicator, then Excel and Quattro only need a single keystroke in each of these columns to fill in the values. Fortunately there are very few cases where a particular character starts multiple countries, the most common I have encountered is "S", which starts "Scotland", "Sweden", and "Switzerland".

On a column containing names of religions I find there are a lot starting with "C". These complicate auto-completion Ignoring various abbreviations these include:

- C. Apostolic
- C.C. Baptist
- Canadian Methodist
- Canadian Presbyterian
- Catholic
- Christian
- Christian Science
- Church of England
- Church of Scotland

- Congregationalist
- Covenanted Baptist

I try to minimize these situations by substituting other values that do not create this problem. For example for “Church of England” I substitute “Anglican”, and for “Catholic” I substitute “Roman Catholic”. In the latter I am also making a theological statement. Technically all Christians who recognize the Nicene Creed are Catholics, since it contains the statement “I believe in one Holy, Catholic, and Apostolic Church.”

Another issue is with the various competing branches of the Methodist church. Some enumerators entered, for example:

1. Meth.
2. Meth. N.C.
3. Meth. E.
4. Meth. W.

However the second represents the “New Connexion Methodist Church”, and therefore should be properly abbreviated as “N.C.Meth.”, while the last represents the “Wesleyan Methodist Church”, which should be abbreviated “W.Meth.”. Making these substitutions reduces the number of keystrokes.

In the “Profession, occupation, trade ...” column of a census you frequently see the following values:

- Farmer
- Farmer's son
- Farmer ret.

In Microsoft® Excel you will be forced to type the entire word “Farmer” every time. It will only autocomplete the entry when you enter the next character. In Quattro Pro it will guess the last such value you entered as soon as you type the “F”. You can then press Tab to go to the end of the value and either continue typing (if Quattro guessed “Farmer”) or backspace to correct the guess. OpenOffice will guess “Farmer” as soon as you type the “F” and you can use the Tab key to select one of the other two values.

Other Issues

For sophisticated users be aware that, naturally, OpenOffice and Quattro do not support using VBA scripts for office automation tasks. Doing so would be a violation of Microsoft copyrights.

Jim Cobban jcobban@magma.ca
34 Palomino Dr.
Kanata, ON, CANADA
K2M 1M1
+1-613-592-9438