



Preservation Handbook

Spreadsheets

Author AHDS History

Version 05

Date 15/07/2005

Change History



Definition

A type of application program which manipulates numerical and string data in rows and columns of cells. Spreadsheets are made up of one or more rectangular objects (known as worksheets) which contain data values, formatting codes and formulae for calculating values. The value in cells can either be entered directly or be calculated from a formula which may involve other cells. Values are generally recalculated automatically whenever a value on which it depends changes. Different cells may be displayed with different formats. Worksheets may also contain embedded images (charts, graphs or maps).

Cells may be referenced either as absolute or relative in either their horizontal or vertical indices. All copies of an absolute reference will refer to the same row, column or cell whereas a relative reference refers to a cell with a given offset from the current cell.

Within the business world, spreadsheets are used for forecasting, while within the humanities world, they are more likely to be used for trouble-free data entry, and basic data analysis.

Description

The notion of electronic spreadsheets dates to the early 1960s, however the first spreadsheet for home computers was launched in 1978. VisiCalc (1978) ran on the Mac Apple II computer. (A PC version which runs within DOS is still available from its creator's web-site.¹) This early application was followed by several subsequent applications:

Microsoft Excel

Microsoft Excel possesses the largest share of the Spreadsheet market and is the most common spreadsheet format to be deposited with the AHDS (excluding tab-delimited text). The product was initially launched for the MS DOS platform in 1985, and was replaced by versions for Microsoft Windows 3x-XP and MacOS. Excel was the first spreadsheet that allowed the user to define the appearance of spreadsheets (fonts, character attributes and cell appearance). It also introduced intelligent cell re-computation, where only cells dependent on the cell being modified are updated, while previously spreadsheets repeatedly recomputed everything or waited for a specific user command.

Numerous revisions to the core interface, file format and functionality (e.g. Visual Basic scripting support) have been made to the application. Incompatibilities are often encountered when manipulating data created in a different version of the software. Since 1993 it has been sold as part of the Microsoft Office suite. The current version is 11, also called Microsoft Office Excel 2003.

Lotus 1-2-3

A spreadsheet program launched by Lotus Software (now part of IBM) in 1983. Once hugely popular, it outsold VisiCalc and for a number of years became the leading spreadsheet for the DOS operating system. Early versions came with a separate program to produce graphs and charts, hence early Lotus spreadsheets are quite simple and are unlikely to possess a graphical component. Version 2.0 introduced Macros and add-ins that contribute to the complexity of the file format. The rise of Microsoft Windows and their promotion of Excel gradually usurped the position of 1-2-3, which was primarily aimed towards OS/2 users. Lotus 1-2-3 remains available for Windows as part of the Lotus SmartSuite package (from IBM).

Quattro Pro

¹ VisiCalc Executable for the IBM PC, <http://www.danbricklin.com/history/vcexecutable.htm> [Available 30 April 2004]



Borland's Quattro Pro (formerly Quattro) software was launched in 1988 for the MS Dos platform. Although it has superior graphics capabilities to MS Excel, it is known to have less-capable calculations than Microsoft's Excel. For example, in many versions, subtracting dates fails to produce the number of days between two dates. However, it does avoid Excel's long-standing worksheet size limitation of 65,536 rows by 256 columns, with a maximum worksheet size of one million rows by 18,276 columns. A free trial of the latest version for Windows is currently available on the Corel web site.

Gnumeric

A free spreadsheet program that is part of the GNOME desktop. Gnumeric has the ability to import and export data in several spreadsheet formats, including Excel, XML, HTML, Applix, Quattro Pro, PlanPerfect, Sylk, DIF, Oleo, SC, StarOffice, and Lotus 1-2-3. Its native format is XML, compressed with gzip.

Additional Information

- <http://www.hyperdictionary.com/dictionary/spreadsheet> [last checked 30 April 2004]
- <http://www.webopedia.com/TERM/s/spreadsheet.html> [last checked 30 April 2004]
- http://searchwin2000.techtarget.com/sDefinition/0,,sid1_gci532933,00.html [last checked 30 April 2004]
- <http://www.wotsit.org> [Available 30 April 2004]



Technical Environment

Common Formats

NB. The file extensions given here for ClarisWorks, MS Works and OpenOffice refer only to the spreadsheets created by those applications.

Format	File Extension	Notes
VisiCalc	*.vc, *.vcx	VisiCalc; the original computer spreadsheet program. Originally (1979) ran on an Apple; an IBM PC version became available in 1981. Led to the development of Lotus 1-2-3 in 1983.
MultiPlan	*.col, *.mod	One of the first spreadsheets developed by Microsoft; lacks graphics capabilities and possesses only limited database functionality. Can read and write files in SYLK, 1-2-3, or dBASE II formats. Contains 77 functions but only 13 macro commands
Lotus Improv	*.imp	Deals with data a very different way from more traditional spreadsheets; e.g. no traditional cell addresses. Doesn't support 1-2-3 or Excel macros, but uses LotusScript.
Microsoft Excel	*.xls, *.xlw	Part of Microsoft Office, the dominant spreadsheet application since version 5 (1993), includes support for VBA as a scripting language.
Quattro Pro	*.wkq, *.wb0, *.wb1, *.wb2, *.wb3	Part of Corel's WordPerfect Office suites, a spreadsheet that provides advanced graphics and presentation capabilities. Current versions are modeled on Excel and support VBA, up to 1,000,000 rows and 18,278 columns.
Lotus 1-2-3	*.wks, *.wk, *.wk1, *.wk3, *.wk3, *.fmt, *.fm1, *.fm3, *.123	An integrated spreadsheet, database and graphics package, supplemented with LotusScript, a BASIC-like language. Version 2.01 incorporates 42 macro commands and 89 functions.



ClarisWorks / AppleWorks	*.cwk, *.cws	Macintosh office application suite.
MS Works	*.wks	Part of an alternative to full-featured office suites. Macros, charts, images, or pictures may be lost when opening an Excel file in Works.
OpenOffice.org Calc	*.sxc	Part of an open source suite of business applications, compatible with Microsoft Office files as well as files created by other word processing applications, such as WordPerfect.
Comma separated values	*.csv	A file format used primarily to transfer basic data between databases and spreadsheets. Each line (up to the carriage return) is considered a record. Fields within each record are divided by a comma. Each line must have the same number of fields (commas). If a comma or leading and/or trailing blanks appear in any field value the field must be enclosed by quotes (") to indicate the information is data and not a field divider. It is for these latter reasons that CSV is not the best format for holding other than strictly numerical data.
Data Interchange Format	*.dif, *.mer	An antiquated file format to express worksheets in a textual format. This should not be used.
Symbolic Link (SYLK)	*.slk	A Microsoft data exchange format for Microsoft Excel and Multiplan spreadsheet files. The maximum number of variables you can save is 256.
Tab delimited	*.tab	Preferred preservation format



Ingest Checklist

The ingest checklist provides a list of actions which should be taken during ingest.

Level 1 (Essential)

- Purpose of spreadsheet described
- Purpose of worksheet(s) described
- Content of worksheet(s) described
- Content of each column described
- Content of each row described
- Data type of each column described
- Coding schemes fully described

Level 2 (Preferred)

- Data types specified
- Calculations checked

Level 3 (Best Practice)

- Contextual information in user documentation
- Details of how the source(s) have been converted to digital form.
- List of sources, including archival or bibliographic references.
- Copy of original material



Preservation

Significant Characteristics

At the lowest level spreadsheet files contain data and data labels. These elements must be preserved. Other elements within spreadsheet files such as formulae, macros, charts (graphs), maps and other embedded data, should only be preserved where, in the case of maps (GIS), copyright on the boundary data has been cleared, or, in the case, of formulae or macros, if the depositor specifically requests their preservation.

Technique

Treat data values as a rectangular object (one per worksheet within a workbook). Export each worksheet (rectangular object) to an ASCII or (preferred) UNICODE delimited text file. Use tab or pipe delimiters, avoid the use of commas. Do not surround textual values in quotes.

Formulas used to calculate data values may be extracted as separate rectangular objects if their complexity or explanatory value is judged worth preserving. Charts and maps (if copyright is not an issue) should be extracted to bitmap or vector images. OLE objects are preserved separately as per the relevant process for the respective digital object types (images etc) - and the links with the spreadsheet are documented. Macros should be stored as plain text files.

Check number of worksheets in a file.

This method will work for all spreadsheet applications. Another method is to use MS Excel 2002 to save as an XML file. This method preserves a) the integrity of the relationships between multiple spreadsheet worksheets; b) any formula contained within the various cells and c) all formatting within the worksheet. Note that this format does not preserve charts, maps or OLE objects. These must be treated as above. However, the output remains system specific and cannot be imported into other applications. Note also that the version of XML used by Microsoft applications is not open like the original XML standard, nor does it conform to the existing W3C XML specification.

Another method of creating XML versions of spreadsheets is to use the open source and freely available application OpenOffice.org. This method will create files saved in XML using the OpenOffice.org spreadsheet schema. The schema and the XML formats used by OpenOffice are non-proprietary and open.

Very little testing or research has been published as to the suitability of such XML spreadsheet files for long-term preservation, however, or whether the primary purpose of this format is to facilitate cross-platform interchange in the short term. The Microsoft version of XML coding is generally not regarded as an open XML that is safe for long-term preservation. However, it is suggested that it would be useful to make an additional XMLSS (the flavour of XML used with MS Excel) version of Excel spreadsheets as an adjunct to the exportation of delimited text.

Validation of Exported Data

- Check number of columns and number of rows.
- Confirming that number of records match up with documented figures.
- Check fields for unknown characters.
- Checking that variables contain expected values (e.g. sex refers to male or female, names refer to names, county variables gives counties etc.) .
- Verifying the coded variables against documented code values.
- Check precision and format of numbers.
- Check precision and format of dates.



Problems and Issues

- Check** All date values contain the correct century digit year
- Reason** Years originally entered with only two digits default to the 20th or 21st centuries if the cell was formatted as 'date' data type.
- Action** Amend dates to contain four digit years:
1. Split the date into separate values for year, month and day by using the year([date]), month([date]) and day([date]) functions
 2. Use arithmetic functions (+,-,/,*) to correct the year
 3. Use function text([year/month/day as number],"general") to convert the year, month and day numbers into text
 4. Use function concatenate([day as text],"/",[month as text],"/",[year as text]) to create a corrected date
- Example** Cell A1 contains the value 12/3/1957. The correct year is 1757.
 In cell B1 type =text(day(A1),"general")
 In cell C1 type =text(month(A1),"general")
 In cell D1 type =text(year(A1)-200,"general")
 In cell E1 type =concatenate(B1,"/",C1,"/",D1)
- Check** Cell A1 does not contain text "ID" (capital I, capital D)
- Reason** When importing a text file, Excel will mistakenly assume a file is in SYLK format if the first two characters are "ID". An error message will display if the file is opened in Microsoft Excel XP, and the file will fail to open on an earlier version.
- Action** Change value to "id" (lower case i, lower case d)
- Example** -
- Check** Headings (field names) are on first row only
- Reason** Databases, statistical packages and other software assume field names are only on the first row of imported text files, subsequent rows of field names will be misread as data
- Action** Edit field names and delete unnecessary rows so that field names are only on the first row
- Example** Cell D1 contains the heading "Total Production", cell D2 contains the heading "Coal", cell E1 is empty and cell E2 contains the heading "oil"
 Change entry in cell D1 to "Total Production: Coal"
 Change entry in cell E1 to "Total Production: Oil"
- Check** Formatting (fonts, borders, colour etc) is not used to convey information
- Reason** Formatting cannot be represented in a text file, so any information conveyed only by formatting must be expressed in a different way
- Action** Convert information represented by formatting into normal values by editing existing values or adding new fields
- Example** Values in column A are displayed in italics if the accuracy of the value is uncertain.
 Put a new field name, "uncertain" in the first row of an unused column
 For each row where the value in column A is in italics, code field "uncertain" as TRUE
 For each row where the value in column A is not in italics, code field "uncertain" as FALSE



- Check** Does worksheet contains functions
- Reason** Only the values calculated by functions, not the functions themselves are exported to tab delimited text
- Action**
- If functions are complex and there are a lot of them, create a separate tab delimited text file containing the text of the functions.
 - If functions are complex, but there are only a few, include the text of the functions in the study read file or other documentation
 - If functions are simple and easily recreated, ignore the functions
- Example** -
- Check** Does worksheet contains merged cells
- Reason** Merged cells are problematic when converted into delimited text
- Action** Unmerge all cells (Excel2000:menu:format:cells:alignment:merge cells) and move/duplicate contents as necessary to preserve column and row layout.
- Example** -
- Check** Do values include footnote/reference markers?
- Reason** The value "203¹" is interpreted by Excel as 2,301, not 203. The value "203*" is interpreted as a text string "203*", not the number 203.
- Action** Create a new field to hold the footnote markers
- Example** The value in cell G14 contains the value "118¹", the value in cell G15 contains the value "568".
Create a new field in column H called "footnotes"
In cell H14 type "1"
Do not enter a value in cell H15
- Check** Worksheet contains notes below a table
- Reason** When imported into databases and statistical packages rows with notes will be treated as data and may be truncated, converted to inappropriate data types, or confuse the import wizards
- Action** Consider creating a separate text file to store the notes
- Example** -

Additional Information

- OpenOffice.org
<http://www.openoffice.org/> [last checked 15 July 2005]
- OpenOffice.org XML document standard
<http://www.oasis-open.org/committees/download.php/12572/OpenDocument-v1.0-os.pdf> [last checked 15 July 2005]
- XML in Excel and the Spreadsheet Component
http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnexcl2k2/html/odc_xlsmiinss.asp [last checked 30 April 2004]



- XML Spreadsheet Reference

http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnexcl2k2/html/odc_xmlss.asp [last checked 30 April 2004]

- Digital Preservation Testbed. White Paper 2. Preserving spreadsheets

<http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-spreadsh-en.pdf>
[last checked 11 April 2005]