DATA CURATION NETWORK

Microsoft Excel Data Curation Primer

Participants: Greg Janée, UC Santa Barbara (gjanee@ucsb.edu)
Sandra Sawchuk, Mount St.Vincent University (sandra.sawchuk@msvu.ca)
Ho Jung Yoo, UC San Diego (hjsyoo@ucsd.edu)

Mentor: Wendy Kozlowski, Cornell University (wak57@cornell.edu)

Format overview	
File extensions	<table><tr><td>.xlsx</td><td>Standard; most common</td></tr><tr><td>.xlsm</td><td>Contains macros</td></tr><tr><td>.xlsb</td><td>Binary equivalent of .xlsx; uncommon</td></tr><tr><td>.xls</td><td>Legacy</td></tr></table>
MIME type	<table><tr><td>.xlsx</td><td>application/vnd.openxmlformats-officedocument.spreadsheetml.sheet</td></tr><tr><td>.xlsm</td><td>application/vnd.ms-excel.sheet.macroEnabled.12</td></tr><tr><td>.xlsb</td><td>application/vnd.ms-excel.sheet.binary.macroenabled.12</td></tr></table>
Structure	ZIP archive adhering to Open Packaging Conventions containing multiple, interrelated XML documents
Versions	Many proprietary legacy versions; since 2007, an ISO standard that has undergone only minor revisions. Nevertheless, closely tied to the Microsoft Excel application, whose present version is v16.0.xx.
Primary fields or areas of use	Microsoft Excel is commonly used by many research disciplines.
Source and affiliation	Microsoft Corporation develops and manages all versions of Microsoft Excel.

Suggested Citation: Greg Janée, Sandra Sawchuk, Ho Jung Yoo. (2019). Microsoft Excel Data Curation Primer. Retrieved from the University of Minnesota Digital Conservancy. http://hdl.handle.net/11299/202816.

This work was created as part of the Data Curation Network "Specialized Data Curation" Workshop #1 co-located with the Digital Library Federation (DLF) Forum 2018 in Las Vegas, Nevada on October 17-18, 2018. See also: Primers authored by the workshop attendees at DLF. http://datacurationnetwork.org.

Metadata standards	
Key questions for curation review	• What is the purpose of the file—data storage or computational model? • Does the file have external dependencies? • If the purpose is to store data, are the worksheets in the file convertible to CSV? • Is the data sufficiently documented?
Tools for curation review	Microsoft Excel
Date created	February 7, 2019
Created by	Greg Janée (gjanee@ucsb.edu), Sandra Sawchuk (sandra.sawchuk@msvu.ca), Ho Jung Yoo (hjsyoo@ucsd.edu)
Date updated and summary of changes made	April 26, 2019 - Added data dictionary resources for different disciplines. - Added links to support 'Legacy File Format' section - Added information to support 'Document' step in the CURATE(D) checklist

Table of contents

Description of format

Overview

Microsoft Excel's widespread adoption in the corporate sector is well known, but the application has also found use in many areas of scholarship. Despite the ubiquity of tabular data in CSV (comma-separated values) format, and the availability of many tools and analysis platforms that operate on CSV files, Microsoft Excel continues to be used widely in the natural sciences and social sciences. As a consequence, Excel files are routinely deposited in data repositories and curators are likely to encounter them.

Since 2007 the Excel file format has been open and defined by ISO and ECMA standards (LoC 2017, OOXML 2019). Further, the format is supported by tools other than Microsoft Excel, notably Google Sheets, LibreOffice, and scientific programming languages such as R and Python. Nevertheless, the format continues to be closely identified with the (proprietary) Microsoft Excel application. Many features commonly used with Excel are not supported by those other tools.

The Excel file format is technically a ZIP archive, organized according to Open Packaging Conventions, that contains a number of interrelated XML files adhering to the Spreadsheet ML language. But because the individual files have no independently reusable value, from a curation and reuse perspective, in practice, an Excel file is always treated as an indivisible unit.

The close association of the Excel file format with a proprietary application limits its long-term reusability. And reusability of Excel data is further hindered by both the complexity of the file format (it is difficult to create new tools that fully support the format) and incompatibilities that can appear when moving Excel files across platforms (macOS vs. Windows) and even across versions of Microsoft Excel on the same platform. As a consequence, the curation recommendation for Excel files is to, where feasible, convert the files into more preservation friendly and reusable formats such as CSV for data tables and PNG for graphs and plots.

Characteristics

At its core, an Excel spreadsheet stores tabular data arranged in a two-dimensional array of cells. Typically, columns in a table represent attributes or variables, and each row represents a record, sample, or subject. The first row or rows may contain column headers, or variable names. Cells themselves may contain numbers, text, dates, and other data types, or they may contain formulas which perform calculations that reference the values of other cells. Beyond this basic model, an Excel file may further contain:

- **Multiple worksheets**, which may be independent or interrelated.
- **Visual effects**: borders, shading, font styling, merged cells, etc. Such formatting may be conditional on cell values.
- **Hidden content:** hidden columns, rows, worksheets, and cell comments, which are not displayed until the user actively reveals them within the spreadsheet.
- **Filters:** transformed views that non-destructively reorder and/or hide the underlying data.
- **Charts:** visualizations (graphs, plots, etc.) that are computed from the underlying data.
- **Pivot tables**: dynamically generated tables that, like filters, display transformed views of the underlying data.

- **References to data values**. A worksheet may reference cell values in another worksheet in the same Excel file, as well as values in separate files referenced by filename. The latter practice creates external dependencies.
- **Macros and scripts**. These are written in the Visual Basic for Applications (VBA) programming language and may perform computation as well as mimic keystrokes and mouse clicks. Due to security considerations, Microsoft Excel is typically configured to disable macros by default. Macros are usually stored in the Excel file, but can also be referenced in a separate, or even a hidden background workbook stored in a system startup folder, usually with the filename `Personal.xlsb`, thus creating another external dependency.

Typical purposes and functions

Researchers use Excel for various purposes spanning a large range of the research data life cycle. Although some of these uses are commonly seen in the research workflow, many may not be considered of value for archiving by researchers, and therefore may seldom be deposited into repositories. Typical purposes of Excel spreadsheets include the following. Note that a single file may be used for more than one purpose.

- **Data entry.** Due to ease of set up and use, Excel may be used to create data entry forms, validate entries, or input data originally collected in a non-digital format (e.g., recorded on paper). See example (Morgan-Short 2016).
- **Data storage and transformation.** Tabular data may be transcribed from non-digital sources, output from instrumentation, or exported from other software and stored in Excel format. These data are also frequently transformed and cleaned in Excel for import into other applications for statistical analysis and visualization. Tabular data resulting from these activities are the most common types that are archived for long-term preservation and reuse. See example (Bassan et al. 2018).
- **Plotting and reporting.** Although not developed as a visualization package, Excel has basic charting and table formatting functions that can be used to generate displays of plots and other graphics. These charts and tables may be used for reporting purposes in publications and presentations. See example (Piet et al. 2017).
- **Computation or modeling.** The emphasis of a spreadsheet may be on formulas that compute and display outputs in response to various user inputs. These spreadsheets may not store data at all, but rather function as computational tools. See example (Bemrah et al. 2016).

What to look for

Problems opening the file

- **Legacy version.** Older versions of Excel files (file suffix .xls, corresponding to Excel 97-2004) may be opened by Microsoft Excel in reduced-functionality "Compatibility Mode." The older file formats are proprietary and therefore not amenable to long-term preservation. Remedy: save the file in the current .xlsx format. See Use Excel with earlier versions of Excel for more information on how to work with earlier versions of Excel files. Excel files are generally upward compatible, but the linked Microsoft support document lists some rare conversion problems that may occur. It's also possible to save a workbook for compatibility with an earlier version of Excel.

- **External references, missing or not.** An external reference (also called a link) is a reference to a cell or range of cells in another Excel file. A missing reference will result in an unusable Excel file, but even if all referenced files are present, external references cause significant usability problems. Due to the way external references are recorded, every new user of the Excel file will need to re-establish the connections between the primary file and any referenced files. Remedy: consider reorganizing multiple Excel files into multiple worksheets within a single file. Unfortunately, there is no automated means of finding all external references. An external reference in a formula will include the external file's path and filename in square brackets ([ ]), followed by the worksheet name, an exclamation point (!), and the cells in that sheet that the formula depends on, e.g., `=SUM('C:\Data\[Compiled.xlsx]Surveys'!C10:C25)`. For more information, see Find links (external references) in a workbook.
- **Macros, missing or not.** Macros hinder usability (due to significant security considerations, macros in Microsoft Excel are typically disabled by default) and portability (macros are supported only by Microsoft Excel). And as with external references to cell values, external macros may be referenced, and the same considerations apply. Note that macros may be stored in a user's `Personal.xlsb` macro file. Remedy: consider removing macros if they are not necessary. If macros are necessary, consider consolidating them in the Excel file.
- **Password protection.** Microsoft Excel's password protection feature is incompatible with preservation and reuse. If the data is sensitive, access should be restricted through repository access and authentication mechanisms instead. Remedy: remove any passwords.

Content problems

For an Excel file whose purpose is data storage, there are many characteristics that can compromise the preservation and reuse of that data. The most preservation-friendly and reusable format for tabular data is comma-separated values (CSV), and recommended practice is to export data as CSV and to archive the CSV file(s) alongside the Excel original. Therefore, any characteristics that preclude conversion to CSV should be considered problematic. The specific problems to look for listed below are derived primarily from Strasser et al. (2014) and (Carpentry 2019):

- **Multiple tables in a worksheet.** Multiple tables in a single worksheet, particularly if the tables have different dimensions or are side-by-side, will greatly confuse any CSV reader and any conversion to CSV because the structure, which is only visually apparent, will be lost. Remedy: move multiple tables to separate worksheets.
- **Ragged tables.** Tools that read and write CSV files generally assume a rectangular structure, that is, that each row has the same number of column values. There may be cases when different row or column lengths in a table is intentional, however, in exporting such a table to CSV, Microsoft Excel will rectangularize the data by adding blank cells, and this in turn may cause the meaning behind the different row or column lengths to be lost. Remedy: reorganize tables or designate a coding scheme for blank cells so that tables are rectangular.
- **Merged cells.** The structure of merged cells will not be maintained when data are exported to CSV. Remedy: unmerge cells and annotate appropriately so that information is not lost.
- **Blank cells.** Blank cells within a contiguous data table are potentially problematic when the table is read by other programs. Remedy: confirm that blank cells are intentional and that the semantics of a blank cell (no data? not applicable? to be determined?) are documented. Preferably, designate a coding scheme for missing data or other explanations for blank cells.

- **Embedded charts.** Embedded content will not be visible when data are exported to CSV. Also, these elements are visible only if the file is opened with Microsoft Excel. Remedy: move embedded content to their own worksheets or save as image files.
- **Embedded comments.** Comments will not be visible when data are exported to CSV. Also, these elements are visible only if the file is opened with Microsoft Excel. Remedy: create a new column titled “Comments” and place comments there.
- **Columns that have mixed data types.** Mixed data types (e.g., numbers and text) in the same column may reflect underlying errors, and even if not, many programs cannot handle mixed data types. Remedy: ensure that columns are uniform with respect to data type. Split data into multiple columns if necessary.
- **Multiple values in one column.** Data is most reusable when it is stored in discrete, independently processable units. For example, location data compacted into a single column "city, state" will be more reusable if city and state are stored as separate columns; similarly, dates are most reusable if stored as separate year, month, and day columns. Remedy: add additional columns as needed.
- **Special characters.** Characters outside the basic alphanumeric character set (mathematical symbols, characters with diacritics, etc.) may cause problems for other programs or may be modified upon export. At minimum, use of the broader Unicode character set will necessitate careful understanding and documentation of the Unicode encoding in use. Remedy: use alphanumeric characters only.
- **Dates.** Microsoft Excel can display dates in many different syntaxes, some of which can be ambiguous or difficult to parse by other programs. Additionally, dates are stored internally by Microsoft Excel in an Excel-specific, numeric format, and this internal representation can occasionally be exposed. A related problem is that Excel, in its legacy versions (but not in recent versions), supported two, incompatible date systems. See Date systems in Excel for information on differences and conversion between the systems. Remedy: consider storing dates as explicit and separate year, month, and day columns.
- **Visual effects.** Color, borders, conditional formatting, and other visual effects can greatly enhance the readability of a spreadsheet, but if such effects convey semantics that are not otherwise present in the spreadsheet, those semantics will be lost in the conversion to CSV. Remedy: consider adding additional columns to the table to convey any unapparent information.
- **Missing column headers.** Every table should have at least one header row (and preferably exactly one header row) to name and describe the columns. Note that column headers are not a substitute for more complete documentation; see the Appendix for guidelines on creating a data dictionary. Remedy: assess the completeness of column documentation and augment as necessary.
- **No primary key.** Generally speaking, every table should have a column or combination of columns that constitute a “primary key”— a quantity that is unique to each row, and that can thus serve to identify and distinguish the row from every other row. For example, measurements taken over time should be identified by date; measurements of samples should identify the sample. Remedy: if no primary key can be identified, consider adding a “row ID” column.

Software for viewing or analyzing data

- Microsoft Excel is the *de facto* tool.
- The Spreadsheet Inquire add-in for Excel can be used to analyze Excel files for problems and inconsistencies, and to make hidden content (e.g., cell comments) visible. It is available only with certain versions of Microsoft Excel for Windows, however.
- The Excel Archival Tool programmatically converts Excel files to open source formats (specifically, CSV and PNG).

Preservation actions

- Convert legacy .xls files to .xlsx. In contrast to .xls, .xlsx is an open format, and it is the currently supported format. Microsoft Excel will still open .xls files, but in a reduced-functionality "compatibility mode."
- Convert .xlsb files to .xlsx. The file formats are functionally equivalent (.xlsb is simply a more compact form), but .xlsb is non-standard and supported only by Microsoft Excel.
- If an Excel file contains macros or VBA scripts, determine if they are integral to the purpose of the spreadsheet. If they're not integral, create a version of the file without them to enhance reusability and portability. If they are integral, contact the depositor to understand their purpose, and then, to the extent possible, test the macros to ensure that they achieve that purpose.
- When data storage is the primary purpose, save worksheets as separate CSV files and save charts as image files. To do the latter, right-click on a chart and select "Save as picture." Consider using the Excel Archive Tool to automate this process.
- When data entry or computation is the primary purpose, conversion to CSV may cause significant loss of key formulas and formats, and is therefore not recommended.

Excel CURATED checklist

- **Check.** Check that the file can be opened, and that it has no missing external dependencies. Ensure that documentation, particularly a data dictionary, has been provided.
- **Understand.** For an Excel file whose purpose is data storage, check for the content problems listed above: missing column headers, use of color, ragged tables, etc. Review the data dictionary for clarity and completeness (see Appendix).
- **Request.** Remediations of content problems are likely outside the scope of what curators can comfortably perform, and beyond what the file creator would feel comfortable having performed by somebody else. Consider requesting that the creator make the changes "to enhance reusability."
- **Augment.** Evaluate the data from the perspective of a non-specialist. Add supporting and contextual metadata as necessary to enhance discovery.
- **Transform.** For an Excel file whose purpose is data storage, convert the data to CSV and archive along with the Excel original.
- **Evaluate.** After converting a data Excel file to CSV file(s), confirm that the conversion process was successful and that the data is faithful to the original Excel file.
- **Document.** Describe any revisions made to the original Excel file, for each new version. Record names responsible for and dates of versioning activities.

Appendix: Creating a data dictionary

A data dictionary describes the meaning or purpose of the data; its relationships to other data; and its origin, usage, and structure (OSF 2019). Data dictionaries can be added to Excel spreadsheets in the form of supplementary worksheet(s) in the same Excel file, or as an additional "README" file accompanying the original Excel file.

Introductory or contextual information: Explain the context of the original study and the role of this data in that study. Include information about versioning, access restrictions, and any information that would be useful for someone to know if they were unable to speak to the data creator or depositor. This information may be gleaned from the study's ethics application or data management plan. Ensure that abbreviations are spelled out. If there are URLs or DOIs that refer to the data source or provide extra information, include them.

Information about the data itself: Each file, worksheet in the file, and table in each worksheet (ideally there is one table per worksheet) should be described. At minimum, provide the following information for each table:

- Overall description of the table, and the meaning of the rows in the table.
- For each variable or column in the table:
    - The column name as it appears in the spreadsheet.
    - The full column name. Column names are often abbreviated for space considerations; consider fully spelling out each column name to aid comprehension and avoid ambiguity.
    - Description: Write a brief definition that describes the column both independently and in the context of the other columns and the table. Consider consulting a metadata librarian about using standardized metadata for these descriptions.
    - Data type: text, integer, date, etc.
    - Unit of measure, if applicable.
    - Acceptable or allowed values: If the column is numeric, list the range of acceptable values. If textual column values are restricted to a vocabulary, list all acceptable values and indicate their meanings.
    - If blank cells occur in the table, explain their meaning (No data? Not applicable? etc.).

Additional resources: Data dictionaries can be made for data in any discipline. For examples and guides to creating data dictionaries, see the Canadian Heritage Information Network (2013) for the Humanities; the USGS for the geological sciences; DataONE for environmental sciences; and National Network of Libraries of Medicine for medical sciences.

References

1. (LoC 2017) "XLSX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5." *Sustainability of Digital Formats: Planning for Library of Congress Collections*, Library of Congress, 27 July 2017, https://www.loc.gov/preservation/digital/formats/fdd/fdd000398.shtml.
   Excellent overview of the history and structure of the format.
2. (OOXML 2019) "Office Open XML." *Wikipedia*, retrieved 23 January 2019, https://en.wikipedia.org/wiki/Office_Open_XML.
   Describes the Excel format in the context of the suite of Office formats.
3. Kara Morgan-Short (2016). Effects of attention to form on second language comprehension: A multi-site replication study - Data entry Excel file UPDATED. [Data set]. Open Science Framework. https://osf.io/d5s2t/. *Download example:* https://osf.io/psmfd/
4. Bassan, Arianna, Ceriani, Lidia, Richardson, Jane, Livaniou, Anastasia, Ciacci, Andrea, Baldin, Rossella, … Dorne, Jean Lou. (2018). OpenFoodTox: EFSA's chemical hazards database (Version 2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1252752. *Download example*: https://zenodo.org/record/1252752/files/OpenFoodToxTX22291_2018.xlsx?download=1
5. Piet, GJ, van Overzee, HMJ, Miller, DCM, & Royo Gelabert, E. (2017). Data Figures Indicators of the 'wild seafood' provisioning ecosystem service based on the surplus production of commercial fish

stocks [Data set]. Zenodo. http://doi.org/10.5281/zenodo.570917. *Download example*: https://zenodo.org/record/570917/files/Data%20figures.xlsx?download=1. *Data are associated with article*: https://doi.org/10.1016/j.ecolind.2016.08.003.

6. Bemrah Aouachria, Nawel, Bakker, Martine, König, Jürgen, Leblanc, Jean-Charles, Lindtner, Oliver, Tlustos, Christina, … Tasiopoulou, Stavroula. (2016, September 23). Food Additives Intake Model (FAIM) - Version 1.1 - July 2013. Zenodo. http://doi.org/10.5281/zenodo.154725. *Download example*: https://zenodo.org/record/154725/files/faimtemplate.xls?download=1.
7. Carly Strasser, John Kunze, Stephen Abrams, and Patricia Cruse (2014). "DataUp: A tool to help researchers describe and share tabular data." *F1000 Research* **3**(6), 12 September 2014, doi:10.12688/f1000research.3-6.v2, https://f1000research.com/articles/3-6/v2.
   Describes problematic spreadsheet practices and their remediations.
8. (Carpentry 2019) "Data Organization in Spreadsheets for Ecologists." *Data Carpentry*, The Carpentries, retrieved 23 January 2019, https://datacarpentry.org/spreadsheet-ecology-lesson/.
   Best practices in creating spreadsheets.
9. (OSF 2019) "How to Make a Data Dictionary." Open Science Framework, retrieved 23 January 2019, http://help.osf.io/m/bestpractices/l/618767-how-to-make-a-data-dictionary.
   Not specific to Excel, but a general guide.
10. (Canadian Heritage Information Network 2013) "CHIN Data Dictionaries". Government of Canada, retrieved 4 April 2019, https://app.pch.gc.ca/application/ddrcip-chindd/description-about.app?lang=en
11. (U.S. Geological Survey) "Data Dictionaries." *Data Management,* U.S. Geological Survey, retrieved 4 April 2019, https://www.usgs.gov/products/data-and-tools/data-management/data-dictionaries.
12. (Data ONE) "Create a data dictionary." DataONE, retrieved 4 April 2019, https://www.dataone.org/best-practices/create-data-dictionary.
13. (National Network of Libraries of Medicine) "Data Dictionary." National Network of Libraries of Medicine, retrieved 4 April 2019, https://nnlm.gov/data/thesaurus/data-dictionary.