# Excel Archival Tool

## Automating the Spreadsheet Conversion Process

University Libraries

Data Repository for the
University of Minnesota

John McGrory

University of Minnesota
Driven to Discover℠

## Introduction

Microsoft's lack of support for between-version compatibility of Excel[1-3] can lead to some concerns for long-term archival of the data contained within Excel files. Although Excel offers built-in data export methods to capture spreadsheet data in more archival-friendly formats, manual file conversion can be tedious, especially for large data sets. Additionally, Excel allows meaningful information to be represented in multiple forms (raw data, charts and figures, cell formulae, cell formatting and styling, pivot tables, etc.), each of which requires its own conversion step (Fig 1).

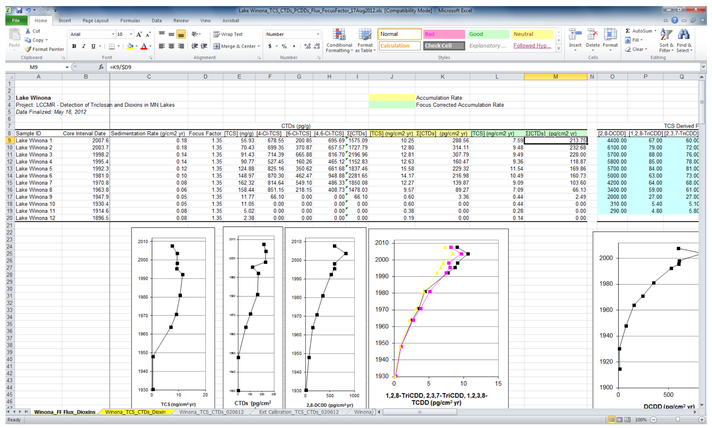


Fig 1. An example of a complex spreadsheet that contains meaningful information in many forms, including raw data, charts, table formatting, and cell formulae.

In an effort to automate the conversion process, we've developed a short program to convert data from Excel's proprietary format to suitable open source formats. Specifically, our Excel Archival Tool extracts raw spreadsheet data as CSV files, charts and figures as PNG images, cell formulae as plain text, and cell formatting and style information as an HTML snapshot (Fig 2).

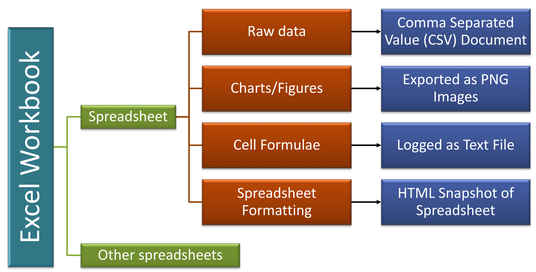



Fig 2. General program flow with breakdown of data types and corresponding conversion steps

## Program Flow and Output

With the Excel Archival Tool, one can convert a single Excel file or an entire folder containing Excel files (including subfolders within this folder). With the user interface (Fig 3), the user selects which aspects of the Excel file they would like converted, as well as a "destination folder" that will contain the final conversion products.

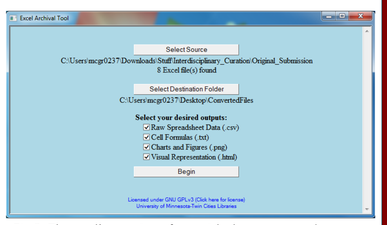



Fig 3. The tool's user interface with the source and destination folders selected. Also note the desired output list where the user can select which aspects of the Excel file(s) are to be extracted.

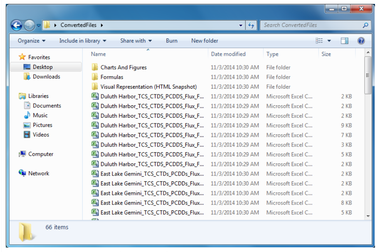


Fig 4. The output folder after the conversion process. All worksheets can be seen as CSV files, with the supplementary information contained in subfolders.

After program completion, the destination folder will open (Fig 4), showing the different Excel worksheets as CSV files, as well as subfolders which contain the additional conversion products. CSV files are named after their corresponding workbook and worksheet to facilitate organization and file identification.

- Each folder in the destination folder is organized into subfolders by workbook name.
- Each workbook folder within the "**Charts and Figures**" folder is further divided into subfolders by worksheet name, each of which contain the worksheet's images and a log of the data ranges used for those charts (Fig 5).
- The workbook subfolders inside of the "**Formulas**" folder contain text files (Fig 6) listing the cell positions and equations for all cell formulae.
- The "**Visual Representation**" folder contains an HTML file and its supporting files within each workbook subfolder (Fig 7).

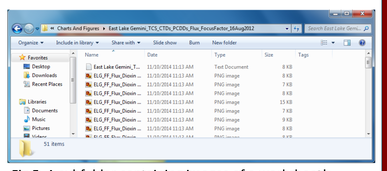


Fig 5. A subfolder containing images of a worksheet's exported charts and a text log of the data ranges used to create these charts.

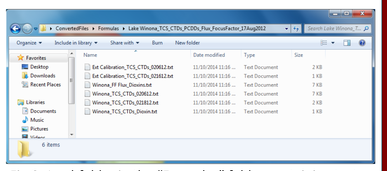


Fig 6. A subfolder in the "Formulas" folder containing text files, each of which contain the named worksheet's cell formulae.

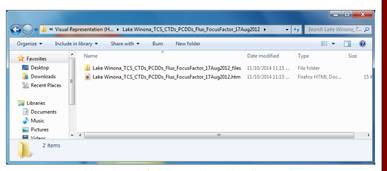
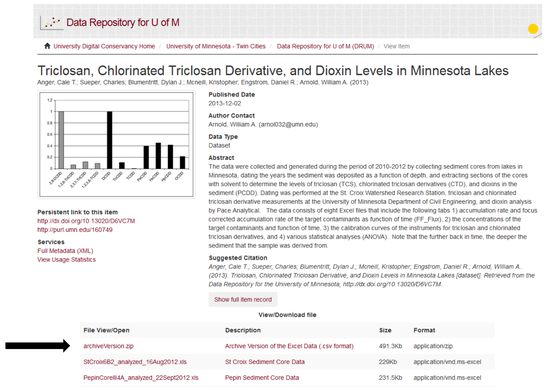


Fig 7. A workbook subfolder within the "Visual Representation" folder, containing an HTML file and its supporting files.

## In Practice

This tool was originally designed for the Data Repository for the University of Minnesota. Below is an example of how one can store the tool's output in the context of a research data repository:




## Methods and Future Directions

**Methods**

- Program core written in Visual Basic Script (VBS).
- Graphical user interface wrapper written in HTML for Applications (HTA).
- Requires a Windows environment and at least Internet Explorer 5.

**Future Directions**

- **Advanced Pivot tables:** The current version can capture the most recent output of a pivot table but is unable to gather other useful information, such as unused fields, custom labels, etc.
- **Platform expansion:** The use of VBS limits the current program's usage to Windows environments, and the HTA GUI requires Internet Explorer.

## Try the tool yourself!

Public access to the tool available at: z.umn.edu/exceltool

## References

1. https://support.office.com/en-sg/article/Use-Office-Excel-2010-with-earlier-versions-of-Excel-2fd9ffcb-6fce-485b-85af-fecfd651a5ac
2. http://coolclimate.berkeley.edu/node/424
3. https://support.office.com/en-za/article/What-s-New-Changes-made-to-Excel-functions-355d08c8-8358-4ecb-b6eb-e2e443e98aac#bm4

Example data set from the figures can be found at: http://dx.doi.org/10.13020/D6VC7M