Excel is threatening the quality of research data — Data Packages are here to help

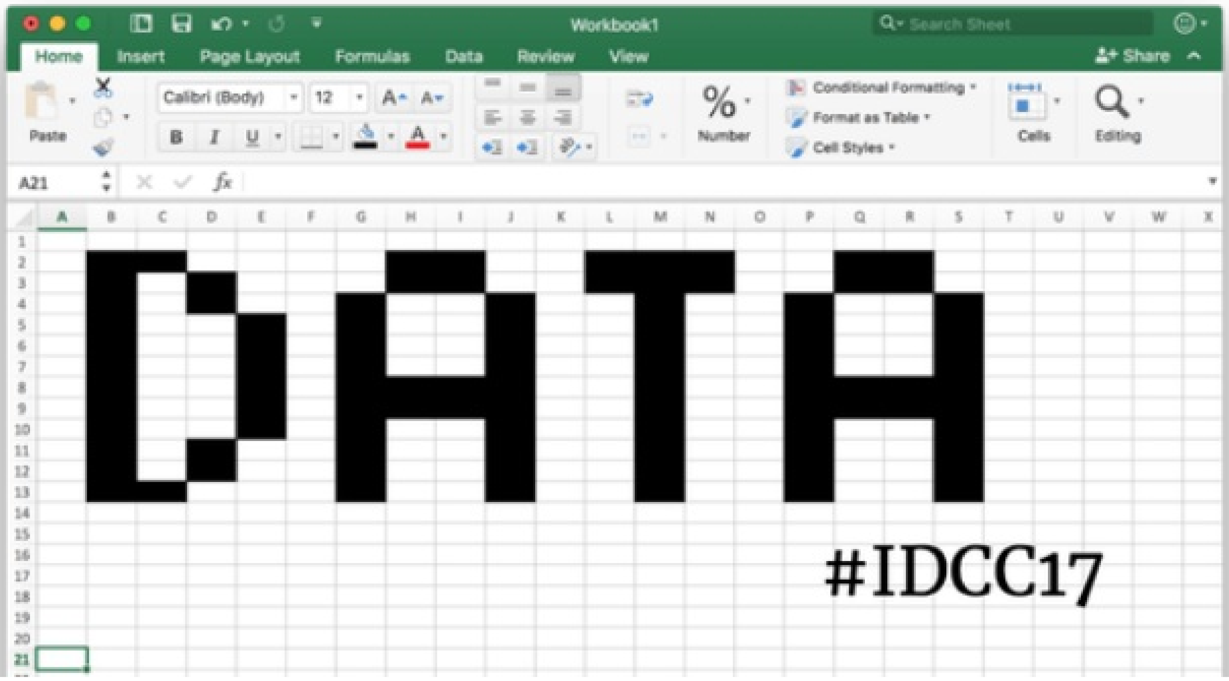
blogs.lse.ac.uk/impactofsocialsciences/2017/02/22/excel-is-threatening-the-quality-of-research-data-data-packages-are-here-to-he

2/22/2017

This week the Frictionless Data team at Open Knowledge International will be speaking about making research data quality visible at the International Digital Curation Conference (#idcc17). ***Dan Fowler*** *looks at why the popular file format Excel is problematic for research and what steps can be taken to ensure data quality is maintained throughout the research process.*

Our Frictionless Data project aims to make sharing and using data as easy and frictionless as possible by improving how data is packaged. The project is designed to support the tools and file formats researchers use in their everyday work, including basic CSV files and popular data analysis programming languages and frameworks, like R and Python Pandas. However, Microsoft Excel, both the application and the file format, remains very popular for data analysis in scientific research.

It is easy to see why Excel retains its stranglehold: over the years, an array of convenience features for visualizing, validating, and modelling data have been developed and adopted across a variety of uses. Simple features, like the ability to group related tables together, are a major advantage of the Excel format over, for example, single-table formats like CSV. However, Excel has a well-documented history of silently corrupting data in unexpected ways which leads some, like data scientist Jenny Bryan, to compile lists of "Scary Excel Stories" advising researchers to choose alternative formats or, at least, treat data stored in Excel warily.



With data validation and long-term preservation in mind, we've created Data Packages which provide researchers with an alternative format to Excel by building on simpler, well-understood, text-based file formats like CSV and JSON and adding advanced features. Added features include providing a framework for linking multiple tables together; setting column types, constraints, and relations between columns; and adding high-level metadata like licensing information. Transporting research data with open, granular metadata in this format, paired with tools like Good Tables for validation, can be a safer and more transparent option than Excel.

Why does open, granular metadata matter?

With our Tabular Data Packages, we focus on packaging data that naturally exists in tables — for example, CSV files — a clear area of importance to researchers as illustrated by guidelines issued by the Wellcome Trust's publishing platform, Wellcome Open Research. The guidelines mandate:

> *"Spreadsheets should be submitted in CSV or TAB format; EXCEPT if the spreadsheet contains variable labels, code labels, or defined missing values, as these should be submitted in SAV, SAS or POR format, with the variable defined in English."*

Guidelines like these typically mandate that researchers submit data in non-proprietary formats; SPSS, SAS, and other proprietary data formats are accepted due to the fact they provide important contextual metadata that haven't been supported by a standard, non-proprietary format. The Data Package specifications — in particular, our Table Schema specification — provide a method of assigning functional 'schemas' for tabular data. This information includes the expected type of each value in a column (string, number, date, etc.), constraints on the value (e.g. "this string can be 10 characters long at most"), and the expected format of the data (e.g. "this field should only contain strings that look like email addresses"). The Table Schema can also specify relations between tables, strings that indicate "missing" values, and formatting information.

This information can prevent incorrect processing of data at the loading step. In the absence of these table declarations, even simple datasets can be imported incorrectly in data analysis programs given the heuristic (and sometimes, in Excel's case, byzantine) nature of automatic type inference. In one example of such an issue, Zeeberg et al. and later Ziemann, Eren and El-Osta describe a phenomenon whereby gene expression data was silently corrupted by Microsoft Excel:

> *"A default date conversion feature in Excel (Microsoft Corp., Redmond, WA) was altering gene names that it considered to look like dates. For example, the tumor suppressor DEC1 [Deleted in Esophageal Cancer 1] [3] was being converted to '1-DEC.' [16]"*

These errors didn't stop at the initial publication. As these Excel files are uploaded to other databases, these errors could propagate through data repositories, an example of which took place in the now replaced LocusLink database. At a time when data sharing and reproducible research are gaining traction, the last thing researchers need is file formats leading to errors.

Zeeberg's team described various technical workarounds to avoid Excel problems, including using Excel's text import wizard to manually set column types every time the file is opened. However, the researchers acknowledge that this requires constant vigilance to prevent further errors, attention that could be spent elsewhere. Rather, a simple, open, and ubiquitous method to unambiguously declare types in column data — columns containing gene names (e.g. "DEC1") are strings not dates, and "RIKEN identifiers" (e.g. "2310009E13") are strings not floating point numbers — paired with an Excel plugin that reads this information may be able to eliminate the manual steps outlined above.

Granular metadata standards allow for new tools and integrations

By publishing this granular metadata with the data, both users and software programmes can use it to automatically import into Excel, and this benefit also accrues when similar integrations are created for other data analysis software packages, like R and Python. Further, these specifications (and specifications like them) allow for the development of whole new classes of tools to manipulate data without the overhead of Excel, while still including data validation and metadata creation.

For instance, the Open Data Institute has created Comma Chameleon, a desktop CSV editor. You can see a talk about Comma Chameleon on our Labs blog. Similarly, Andreas Billman created SmartCSV.fx to solve the issue of broken CSV files provided by clients. While initially this project depended on an *ad hoc* schema for data, the developer has since adopted our Table Schema specification.

Other approaches that bring spreadsheets together with Data Packages include Metatab which aims to provide a useful standard, modelled on the Data Package, of storing metadata within spreadsheets. To solve the general case of reading Data Packages into Excel, Nimble Learn has developed an interface for loading Data Packages through Excel's Power Query add-in.

For examples of other ways in which Excel mangles good data, it is worth reading through Quartz's Bad Data guide and checking over your data. Also, see our Frictionless Data Tools and Integrations page for a list of integrations created so far. Finally, we're always looking to hear more user stories for making it easier to work with data in whatever application you are using.

This post originally appeared on the Open Knowledge International blog and is licensed under a CC BY 4.0 license. It has been adapted from a paper on Making Research Data Quality Visible, to be presented by Jo Barratt at this week's International Digital Curation Conference (IDCC).

Note: This article gives the views of the author, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the author

Dan Fowler *contributes to various projects at Open Knowledge and currently serves as developer advocate helping to connect a community of makers and doers around open data with the technology work conducted by Open Knowledge International. He has a Master's degree in Information and Communication Technologies for Development from Royal Holloway, University of London and a Bachelor's degree in Psychology from Princeton University. Between degrees, he worked as a sysadmin for an investment bank in New York.*