IJDC | *General Article*

Researcher Training in Spreadsheet Curation

Gene Melzack
University of Sydney

Abstract

Spreadsheets are commonly used across most academic disciplines, however their use has been associated with a number of issues that affect the accuracy and integrity of research data. In 2016, new training on spreadsheet curation was introduced at the University of Sydney to address a gap between practical software skills training and generalised research data management training. The approach to spreadsheet curation behind the training was defined and the training's distinction from other spreadsheet curation training offering described.

The uptake of and feedback on the training were evaluated. Training attendance was analysed by discipline and by role. Quantitative and qualitative feedback were analysed and discussed. Feedback revealed that many attendees had been expecting and desired practical spreadsheet software skills training. Issues relating to whether or not practical skills training should and can be integrated with curation training were discussed. While attendees were found to be predominantly from science disciplines, qualitative feedback suggests that humanities attendees have specific needs in relation to managing data with spreadsheets that are currently not being met. Feedback also suggested that some attendees would prefer the curation training to be delivered as a longer, more in depth, hands on workshop.

The impact of the training was measured using data collected from the University's Research Data Management Planning (RDMP) tool and the Sydney eScholarship Repository. RDMP descriptions of spreadsheet data and records of tabular datasets published in the repository were analysed and assessed for quality and for accompanying data documentation. No significant improvements in data documentation or quality were found, however it is likely too soon after the launch of the training program to have seen much in the way of impact.

Identified next steps include clarifying the marketing material promoting the training to better communicate the curation focus, investigating the needs of humanities researchers working with qualitative data in spreadsheets, and incorporating new material into the training in order to address those needs. Integrating curation training with practical skills training and modifying the training to be more hands on are changes that may be considered in future, but will not be implemented at this stage.

Received 20 January 2017

Correspondence should be addressed to Gene Melzack, Fisher Library, The University of Sydney, Eastern Avenue, Camperdown NSW 2006. Email: gene.melzack@sydney.edu.au

An earlier version of this paper was presented at the 12th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Background

The Problem

The use of spreadsheets for collecting, analysing, and storing research data is common, but not without potential problems. Ziemann, Eren, and El-Osta (2016) alerted the genomics research community to issues with data integrity in their field due to Microsoft Excel's automatic format conversion; a high profile paper in economics was refuted in part on the basis of a spreadsheet error in data analysis (Herndon, Ash and Pollin, 2013); and Barchard and Pace (2011) document the impact of the type of data entry methods commonly used with spreadsheets on data accuracy and statistical results. Due to their familiarity and availability, spreadsheets are used widely by many researchers across most fields of research. There is therefore a need to train researchers on how to effectively curate their data while using spreadsheets.

Defining Spreadsheet Curation

The goal of curating spreadsheets is to capture and preserve research data so that it can be retrieved, understood, and used in the future. Researchers use spreadsheets for so many different purposes and with so many different types of data that one size spreadsheet curation does not fit all. The information captured and capture methods used vary depending on the origins and nature of the data, as do the file formats needed to preserve the data. What it means to preserve data in a usable format depends on how and with what tools the data is likely to be used, and for what purpose. For example, instrument data that is previewed in Excel before being analysed in MATLAB will need to be curated differently from qualitative data entered into Google Sheets before being exported as an OpenDocument Spreadsheet and analysed in NVivo. Both will in turn need to be curated differently from data that has been cleaned and processed in Excel using custom macros and then analysed using pivot tables and Excel charting functions. Different curation recommendations apply to these different examples of spreadsheet data. Spreadsheet curation requires a qualitative judgement on what features of the data are significant and/or necessary in order to understand and use the data, and different curation strategies will apply on the basis of this judgement.

Existing Training

In May 2016, the University of Sydney Library's Research Data team were providing introductory research data management sessions to teach university researchers the basic principles of data curation, while Intersect, the NSW eResearch support agency was offering Microsoft Excel training to build researchers' technical skills. Intersect was also offering training sessions on how to use statistical and computational analysis software packages. This left a gap in the training offerings: the introductory research data management training addressed best practice in data curation but didn't allude to any spreadsheet-specific curation issues, while the Excel training addressed software usage skills but not best practice in data curation.

The Data Carpentry Spreadsheets for Ecology (Bahlai and Pawlik, 2014) lesson was developed in 2014 to address a similar gap. However, this course was tailored for the needs of scientists rather than a cross-discipline audience. And while Data Carpentry workshops have occasionally been run in Sydney, they're not offered on a regular basis, and the Data Carpentry Spreadsheets for Ecology lesson had never been run locally.

Implementation

New Training

To address this gap in training provision, the Research Data team developed a new training package on curating research data using spreadsheets. This training package was designed to raise awareness of key data curation issues in relation to spreadsheets among researchers and provide advice on simple tools and practices to mitigate some common curation risks. A secondary intention of the training was to promote the other existing eResearch training offerings to researchers who want to improve their technical skills or even move on to using other, more curation-friendly data collection, storage, or analysis tools. Unlike the Data Carpentry training, which assumed that spreadsheets are primarily a tool for collecting and cleaning data in preparation for performing data analysis using statistical software, this new training addressed all the potential uses of spreadsheets by researchers in various disciplines, including data collection, cleaning, storage, analysis, and visualisation, and provided best practice guidance for multiple different use cases.

The new training and the Data Carpentry spreadsheet training cover a lot of common ground. Both sessions address best practice for structuring and formatting data in spreadsheets, including guidelines on organising data with variables in columns, observations in rows, and one data value per cell, as well as on avoiding colour-coding and special characters. Both sessions also include recommendations for data validation using Excel, handling date formats, and exporting data in open file formats, specifically comma separated values (.csv).

The new training goes beyond existing spreadsheet training materials by referring to additional tools and providing additional guidelines for implementing the recommendations made, as well as providing advice on creating data documentation to accompany spreadsheets. For example, while existing training refers only to the Microsoft Excel data validation feature, the new training also refers to RightField (Wolstencroft et al., 2011), a tool that integrates with Microsoft Excel to implement existing ontologies for standardising data entry. The new training acknowledges that it is sometimes appropriate for researchers to organise their data into multiple worksheets and makes some recommendations for doing so, including making use of the Protect Sheet feature in Microsoft Excel to prevent accidental modifications to worksheets containing original raw data.

The new training addresses the needs of researchers who perform analysis within spreadsheets by providing recommendations on suitable formats for capturing both the analysis and its visualisations. Researchers using formulas or macros are advised to save their work in Microsoft Excel (.xlsx), OpenDocument Spreadsheet (.ods), or eXtensible Markup Language (.xml) formats while researchers who use spreadsheets to generate charts are advised to save their data as a comma separated values file and their charts

separately as images in Tagged Image File Format (.tiff) or Portable Network Graphics (.png) format. The training also warns researchers against using the header 'ID' in the first column of a spreadsheet when saving data in comma separated values format, as this can sometimes lead to difficulties re-opening the file with spreadsheet software.

A substantial portion of the new training is dedicated to data documentation: what and how to document spreadsheet data. The recommendations include: to document the creators of the dataset; file naming conventions and related files; data sources, including temporal and geospatial metadata and data collection methods; data definitions, including variable names, data codes, and units and precision; and technical requirements, including the hardware and software used to generate or analyse the data. Two methods for creating data documentation are offered: manual creation in a separate README document, or automated creation through the Colectica tool, which integrates with Microsoft Excel, allowing data documentation to be generated from metadata entered into a sidebar within the spreadsheet itself.

Evaluation

Uptake

The pilot training session took place in September 2016 and subsequent training sessions are being offered on a monthly basis. The training has so far been in high demand and sessions have booked out quickly. The session is only one hour long, so even time-pressured researchers feel able to attend. Full sessions were run in October and November 2016, with the training program now on hiatus over the summer, to resume in March 2017 once the new academic year is underway. Attendees spanned 13 of the University's 16 faculties, with the Nursing School, the College of the Arts, and the Conservatorium of Music the only faculties not represented among attendees. Four of the top five faculties in terms of training attendance were science disciplines; 46 attendees came from Medicine, 24 from Science, 14 from Pharmacy, and 13 from Health Sciences. The faculty of Arts and Social Sciences, the only humanities faculty in the top five, sent 15 attendees. Ten of the training attendees were University administrative staff who were not affiliated with any particular faculty. The faculties of Agriculture and Environment, Architecture, Design and Planning, Business, Dentistry, Engineering and Information Technologies, Law, and Veterinary Science sent between one and three attendees each.

The majority of attendees were higher degree by research students (52.6%). The next two biggest groups of attendees were professional staff (23.0%) and academic staff (15.1%). More than half of the professional staff who attended (13 out of 23) were research support staff affiliated with a particular faculty.

These attendance numbers suggest that there is higher demand for spreadsheet curation training from science disciplines. To verify this, attendance numbers were examined in the context of the size of the faculties, as indicated by postgraduate enrolment numbers, given that the majority of attendees are higher degree by research students. When attendance was considered as a proportion of postgraduate enrolments, attendance from science disciplines was found to be significantly higher than attendance from humanities disciplines. Since much of the common material shared with the Data Carpentry lesson is geared towards science disciplines, the stronger showing from

science attendees justifies a slant towards that audience. However, around 20% of attendees are from arts and humanities disciplines, so their needs should not be neglected out of hand.

Feedback

Feedback from attendees has been very positive, with the majority responding that they learned something new and that they'll be able to apply what they learned. 94.8% of respondents affirmed the statement "I now have a better understanding of how to manage data effectively with spreadsheets", with 33 marking Agree and 20 marking Strongly Agree. Two responses were neutral (Neither Agree nor Disagree), with only one denying the statement (Disagree or Strongly Disagree). Similarly, 92.7% of respondents confirmed that "I will be able to apply the recommendations made to my use of spreadsheets", with 37 answering Agree and 14 Strongly Agree. Four responses to this statement were neutral, but none disagreed.

In answer to the question "Is there anything that wasn't covered in the session that you think should have been?" five respondents asked for more in depth practical training on how to use various features in Excel, such as formulas and charts, and three asked for more on SPSS and other data analysis tools. Some of these topics are covered in the Excel and statistical software R training sessions provided by Intersect, however this feedback does suggest a need for SPSS training at the university. This feedback also indicates that the content of the session did not always meet the expectations of some attendees, though the same attendees acknowledge the value of the content even if it is not what they were anticipating. In marketing the session to researchers it has been a challenge to differentiate best practice curation training from practical software skills training. A number of attendees made contact after the training for referral on to the other training and statistical consultancy services on offer.

In an ideal world, practical software skills and best practice data curation would be taught simultaneously and without making a clear distinction between the two in the minds of researchers. However, there are a number of practical reasons why this has not occurred. One such practical consideration is time. Adding curation concerns into a practical Excel session makes it longer and requires more of a block time commitment from researchers. Another practical concern is in finding appropriately skilled teachers. Software instructors are often unfamiliar with data curation best practice, while data curators do not necessarily have the in depth software skills required to teach a technical class. This could be resolved through two subject matter experts coming together to collaborate on teaching sessions, however this leads to the next practical hurdle: the fact that the relevant subject matter experts are employed by different agencies with different priorities and remits. The practical software skills training is provided by Intersect, which is a state-wide agency with a remit to cultivate e-research skills in NSW researchers. The data curation training is provided by the Research Data team at the University of Sydney Library, whose focus is on improving the research data management and data curation skills of University of Sydney researchers. While it may yet be possible to work together with Intersect to collaborate on delivering combined practical and curation-focused spreadsheet training, this would require Intersect to modify a statewide training program for the sake of one institution, and the modified training could only be made available to University of Sydney researchers, as the Research Data team do not have the time, resources, or remit to extend their curation training to other institutions.

Because of these practical hurdles to offering combined practical software skills and data curation training, there are no plans to modify the spreadsheet training in response to the feedback requesting more practical Excel, SPSS, and data analysis training, but to continue to refer interested researchers on to other available training and services. Instead, the title and description of the training will be modified to make it more apparent that this session is a best practice session, designed to complement the existing practical skills training offered elsewhere. It may be worth investigating whether this training can be integrated with or delivered in more close association with the practical skills training in future.

Four comments in response to the question about what else should have been covered requested more information on using spreadsheets for qualitative research, including managing qualitative data with spreadsheets for use with NVivo. The training prompted one researcher to request a follow up consultation with the Library's research data service in relation to their qualitative data needs. This feedback suggests that arts and humanities researchers dealing with qualitative data have their own unique needs that are not necessarily being met by the current training, so there is an intention to investigate those needs further and modify the training to better address them.

Some attendees have also requested a more hands on approach. In response to the question "Do you have any feedback about the way the session was run?" four commenters requested more interactive delivery. It has been challenging to strike the right balance between providing information in a shorter, lecture-style format to larger audiences versus using a hands-on workshop style, like that employed by the Data Carpentry Spreadsheets for Ecology lesson, which necessitates longer sessions with smaller participatory audiences. The former approach is currently used due to the fact that it makes the training more accessible to more researchers and can be run more often, however the latter approach would provide a richer learning experience for researchers, if the resources were available to provide the training in this format. However, given that 22 of the responses to this question were broad positive statements about the training's delivery, and in the interest of serving as wide an audience as possible, the intention for now is to continue offering the training in its current format.

There were also three responses to the question about how the session was run that suggested running separate sessions for administrative staff and researchers. The name of the training session was "Managing research data with spreadsheets" and these same comments indicated a confusion over what was meant by the term 'research data'. Administrative staff were taking 'research data' to refer to information and metrics about research outputs that are used to assess impact, rather than data that underpins or is itself a research output, and were expecting the training to focus on using spreadsheets to analyse research information and metrics within an administrative context. Since the training is primarily aimed at researchers with the goal of improving research data curation, the content and focus of the training will not change in response to these comments. However, the title and marketing of the training will change in an attempt to better communicate the purpose of the training to the intended audience. This training session will become part of a suite of "Research data management best practice" training sessions, so will be renamed "Research data management best practice series - spreadsheets", and the description of the training on our publicity materials will make it clear that researchers are the intended audience for the training.

 doi:10.2218/ijdc.v12i2.507

Measuring Impact

Researchers at the University of Sydney are required to complete research data management plans (RDMPs). Limited information about the quality of spreadsheet data can be gleaned from RDMPs. However, to measure the effectiveness of the spreadsheet curation training, RDMPs that document the use of spreadsheets, Excel, or tabular datasets were identified and examined for whether these RDMPs also included mention of data documentation such as READMEs or data dictionaries. A python script was created to analyse the Excel report generated by the RDMP tool. The script searches the description field in which researchers are asked to describe their data and identifies plans that include any variations on the words 'excel', 'spreadsheet', 'xls', 'delimited', or 'tabular' in order to identify plans describing spreadsheet or tabular datasets. The script also identified plans that included the words 'readme', 'dictionary', or 'definition' in their descriptions. Plans that were identified as referring to both spreadsheets and documentation were classified as documented spreadsheets.

Prior to the launch of the training, only 8.5% of RDMPs that referenced spreadsheets or tabular datasets also referred to accompanying data documentation. It is hoped that, over time, this percentage will increase as a result of the training. Since the spreadsheet curation training was launched the number of documented spreadsheets has thus far risen to 9.3%. However, this increase is not statistically significant and represents merely four new plans, only one of which was authored by a researcher who had attended the training.

The Sydney eScholarship Repository accepts submissions of datasets for publication, including spreadsheets and tabular datasets. The volume and quality of the datasets in the repository before the launch of the spreadsheet training was audited as a benchmark for measuring the effectiveness of the spreadsheet curation training. Spreadsheet and tabular datasets were initially identified by searching the repository for records with attachments with the file extensions .csv, .txt, .xls, and .xlsx. The attachments were examined and classified as a spreadsheet or tabular dataset. The audit then assessed whether the identified datasets were accompanied by data documentation and were organised and formatted according to the recommendations made in the training. This involved assessing whether the dataset included: an additional title row, blank rows or columns, multiple tables in a single worksheet, special characters in column headings, colour coding, Excel comments, or merged columns. Datasets were assigned a value of 0 if they failed to meet a recommendation and 1 if they met the recommendation. Total scores of between zero and eight were calculated based on the number of recommendations met. The higher the score, the more recommendations met, and therefore the higher the assessed quality of the dataset. Assessed datasets were divided into pre and post training groups, based on whether they were submitted to the repository before or after the launch of the training.

The pre-training datasets were of slightly higher quality than the post-training datasets on all statistical measures, however this drop in quality was not a statistically significant change, given that the post-training dataset consisted of only four datasets, none of which were created by training attendees. At this point it is too early to see any improvements in spreadsheet curation in repository submissions due to the new training program. Attendees of spreadsheet curation training sessions run thus far will, if they do implement the recommendations, only be in the early stages of applying the training to their research data. We would therefore expect to see the impact of the session in the processes at the early stages of the research data lifecycle, such as research data management planning, but not in later processes such as data publication. This is

consistent with the modest improvements in spreadsheet curation noted in RDMPs and the lack of a discernible difference in the quality of spreadsheet repository submissions. It is hoped that as these metrics continue to be tracked over time, they will show improvements in spreadsheet curation in both RDMPs and in the repository in response to the training.

Conclusions

Next Steps

Spreadsheet curation training is the first instalment in a series of best practice research data management training sessions, designed to complement existing and new training offerings. In response to the feedback so far, the name and details of the intended audience will be updated in the marketing materials for the current spreadsheet curation training. Additional work will be done to identify the spreadsheet curation needs of humanities researchers working with qualitative data and the training will be updated to address those needs where possible. Further changes, such as integrating the curation training with the practical skills training, or introducing a less frequent hands-on workshop, will be considered as the program continues. Alongside the spreadsheet curation training, a new training package on curating data for use with scripting software, based in part on the Project TIER protocol (Ball and Medeiros, 2012), is in development. This will complement the existing Intersect training on programming with Python, R, and MATLAB, in particular their "Reproducible scientific analysis with R" session. Additional best practice training is also being developed on data sharing and publication, and on data visualisation, as a further complement to Intersect's technical offerings to address the gap in teaching best practice, tool-independent visualisation.

References

Bahlai, C. & Pawlik, A. (2014). Data carpentry spreadsheets for ecology. Retrieved from http://www.datacarpentry.org/spreadsheet-ecology-lesson/

Ball, R.J. & Medeiros, N. (2012). Teaching students to document their empirical research. *The Journal of Economic Education, 43*(2), 182–189. doi:10.1080/00220485.2012.659647

Barchard, K.A. & Pace, L.A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behaviour, 27*(5), 1834–1839. doi:10.1016/j.chb.2011.04.004

Herndon, T., Ash, M., & Pollin, R. (2013). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics, 38*(2), 257–279. doi:10.1093/cje/bet075

Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J. L., . . . Goble, C. (2011). Rightfield: embedding ontology annotation in spreadsheets. *Bioinformatics, 27*(14), 2021–2022. doi:10.1093/bioinformatics/btr312

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome Biology, 17*(177). doi:10.1186/s13059-016-1044-7