Researcher training in spreadsheet curation

Presented by
Gene Melzack
Research Data Team, Fisher Library
Background
The Problem

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

The Problem

Spreadsheet coding error

A coding error in the RR working spreadsheet entirely excludes five countries, Australia, Austria, Belgium, Canada, and Denmark, from the analysis. The omitted countries are selected alphabetically and, hence, likely randomly with respect to economic relationships.

Defining spreadsheet curation
Defining spreadsheet curation
Defining spreadsheet curation
Existing training

DATA CARPENTRY

INTERSECT
Implementation
New training

Content

RightField

Colectica for Excel

Data Carpentry
New training

Delivery
Evaluation
Feedback

I now have a better understanding of how to manage data effectively with spreadsheets

I will be able to apply the recommendations made to my own usage of spreadsheets
Feedback

Software skills
Feedback

Qualitative data
Feedback

Delivery
Feedback

Target audience
Measuring impact

Research Data Management Plans (RDMPs)

- ‘excel’
- ‘xls’  ‘spreadsheet’
- ‘delimited’  ‘tabular’

documented spreadsheet

8.5% → 9.3%

- ‘dictionary’
- ‘definition’
- ‘readme’
Measuring impact

Sydney eScholarship Repository

.Score

.csv
.txt
.xls
.xlsx
Conclusions
Next steps