The culture of open scholarship

Karl Broman

Biostatistics & Medical Informatics
Univ. Wisconsin–Madison

kbroman.org
github.com/kbroman
@kbroman@fosstodon.org
Slides: bit.ly/broman2023

These are slides for a 10-15 min talk on open scholarship for the Big 10 Academic Alliance libraries, in Jan 2023.

Slides (pdf): https://kbroman.org/Talk_Big10Libs/oa2023.pdf

Slides with notes (pdf): https://kbroman.org/Talk_Big10Libs/oa2023_notes.pdf

Source: https://github.com/kbroman/Talk_Big10Libs

open access

open educational resources

open source

open science

In thinking about Open Scholarship, I'm generally thinking of four things: open access publications, open educational resources, open source software, and open science (by which I mean open data, methods, and materials).

I'm going to focus mostly on Open Access publications, and on academic scientists.

Concerns in the humanities can be quite different, and I'm going to focus on what I know best, which is the situation in science.

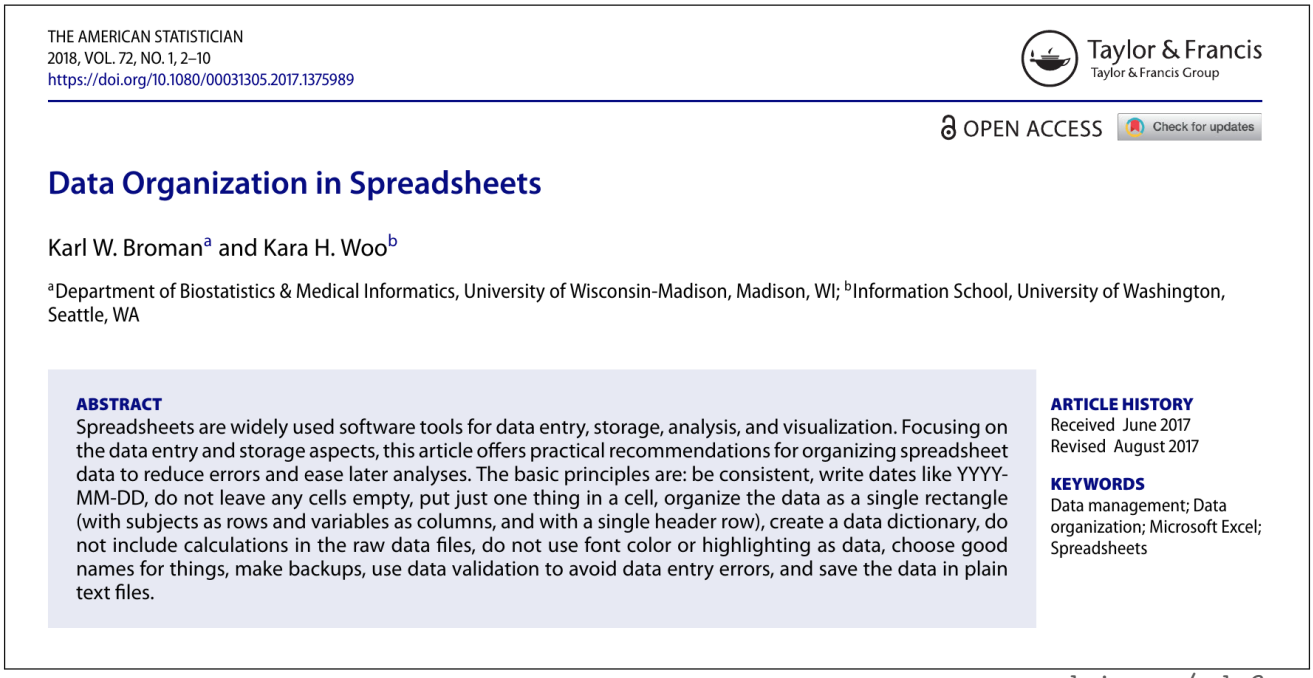
About me

- Applied statistician working in genetics
- Write & support many open-source software packages
- Co-author on 170 papers and 1 book
- Reviewer for 90 different journals
- Formerly
  - Associate Editor and Senior Editor at *Genetics*
  - Associate Editor at *Biostatistics*
  - Associate Editor at *Journal of the American Statistical Association*
  - Academic Editor at *PeerJ*
  - Editorial Board member of *BMC Biology*

I thought I should say a bit about my experience with the publication process.

My research spans genetics and statistics, but the majority of my publications are in the genetics or biology literature; I only have a couple of papers in statistics journals.

I've reviewed for a lot of different journals, and I spent a decade as an editor for the society journal, Genetics.



THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

ABSTRACT
Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

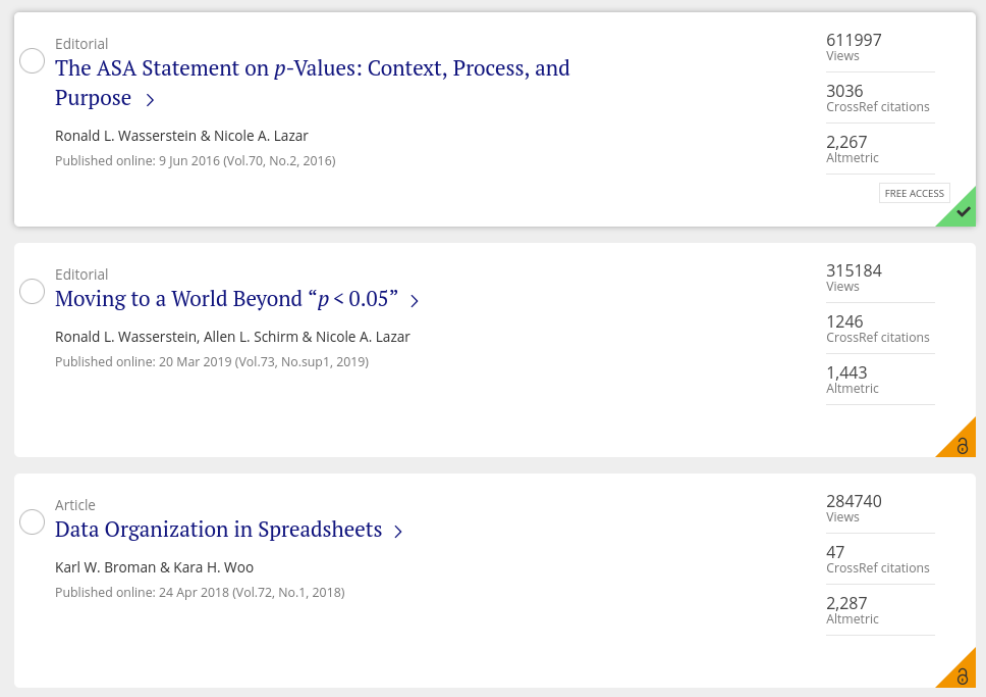
doi.org/gdz6cm

I thought I'd start with a story about the paper of mine.

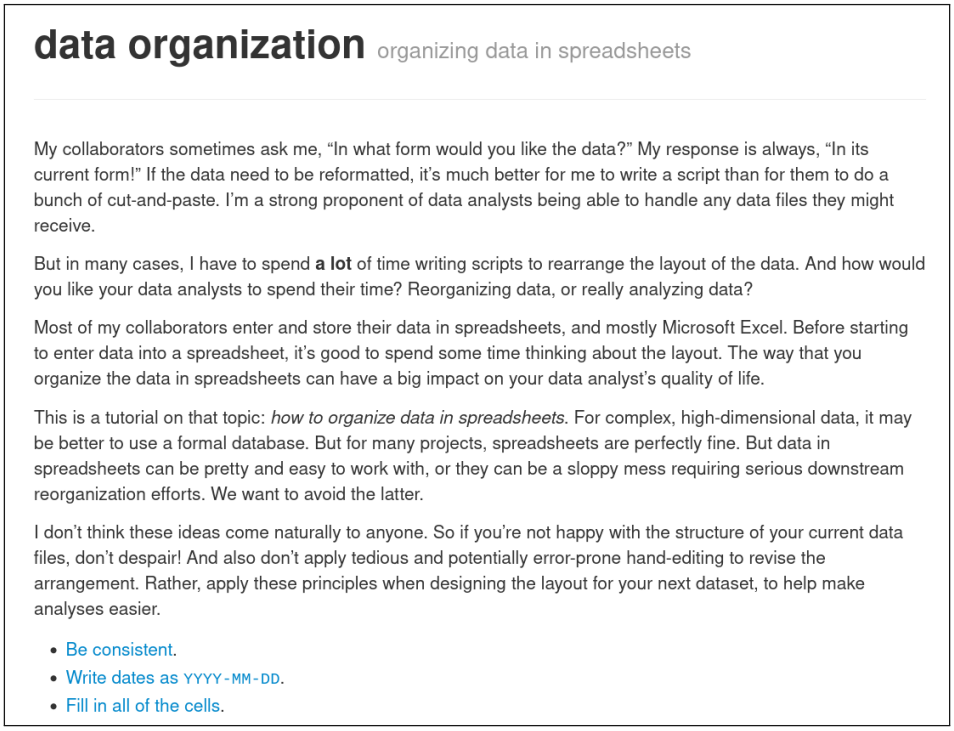
This isn't a typical research paper, but really more of a tutorial, or really a screed.

	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

The paper concerns how to organize data in spreadsheets. And really that spreadsheets shouldn't be organized like this example, but rather as a rectangle with the columns being the measured variables and one row per subject, and a single header row.



It was published in the American Statistician, a society journal, and is third-most downloaded paper in that journal, after two papers about P-values.



data organization organizing data in spreadsheets

My collaborators sometimes ask me, “In what form would you like the data?” My response is always, “In its current form!” If the data need to be reformatted, it’s much better for me to write a script than for them to do a bunch of cut-and-paste. I’m a strong proponent of data analysts being able to handle any data files they might receive.

But in many cases, I have to spend **a lot** of time writing scripts to rearrange the layout of the data. And how would you like your data analysts to spend their time? Reorganizing data, or really analyzing data?

Most of my collaborators enter and store their data in spreadsheets, and mostly Microsoft Excel. Before starting to enter data into a spreadsheet, it’s good to spend some time thinking about the layout. The way that you organize the data in spreadsheets can have a big impact on your data analyst’s quality of life.

This is a tutorial on that topic: *how to organize data in spreadsheets*. For complex, high-dimensional data, it may be better to use a formal database. But for many projects, spreadsheets are perfectly fine. But data in spreadsheets can be pretty and easy to work with, or they can be a sloppy mess requiring serious downstream reorganization efforts. We want to avoid the latter.

I don’t think these ideas come naturally to anyone. So if you’re not happy with the structure of your current data files, don’t despair! And also don’t apply tedious and potentially error-prone hand-editing to revise the arrangement. Rather, apply these principles when designing the layout for your next dataset, to help make analyses easier.

- Be consistent.
- Write dates as `YYYY-MM-DD`.
- Fill in all of the cells.

kbroman.org/dataorg

7

Before it was a paper, it was just a website. It’s one of several short tutorials that I’ve written, about things related to reproducible research and open science.

Data organization in spreadsheets

Karl W. Broman *
Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison
and
Kara H. Woo
Information School, University of Washington

September 11, 2018

Abstract

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this paper offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, don't leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header

doi.org/10.7287/peerj.preprints.3183v2

In addition, the submitted manuscript is openly available at PeerJ Preprints.

INVOICE

INVOICE NUMBER:
943345712

INVOICE DATE:
10/06/2017

TAX INVOICE

CUSTOMER NUMBER:
3551015
Please quote your customer number on all correspondence

TERMS:
Payable in 30 Days

Taylor & Francis
Taylor & Francis Group

PAID

INVOICE TO:
Biostatistics & Medical Informatics
University of Wisconsin-Madison
Biostatistics & Medical Informatics
2126 Genetics-Biotechnology
Center
425 Henry Mall
MADISON WI 53706
UNITED STATES OF AMERICA

DESPATCH TO:
Mr Karl Broman
Biostatistics & Medical Informatics
University of Wisconsin-Madison
2126 Genetics-Biotechnology
Center
425 Henry Mall
MADISON WI 53706
UNITED STATES OF AMERICA

YOUR TAX REF:

OUR REF:

OUR TAX REF:
04-3801744

ORDER NUMBER:
4490118
CUSTOMER ORDER:
10.1080/00031305.2017.1375989

ORDER REF.	QTY	ISBN/ISSN	TITLE	UNIT PRICE	DISC	NET VALUE	TAX	TAX %
T&F iOpen Access Fee	1	1537-2731	The American Statistician Online	2,950.00	0.00%	2,950.00	0.00	

Nevertheless, I paid nearly $3000 to have the published paper available open access.

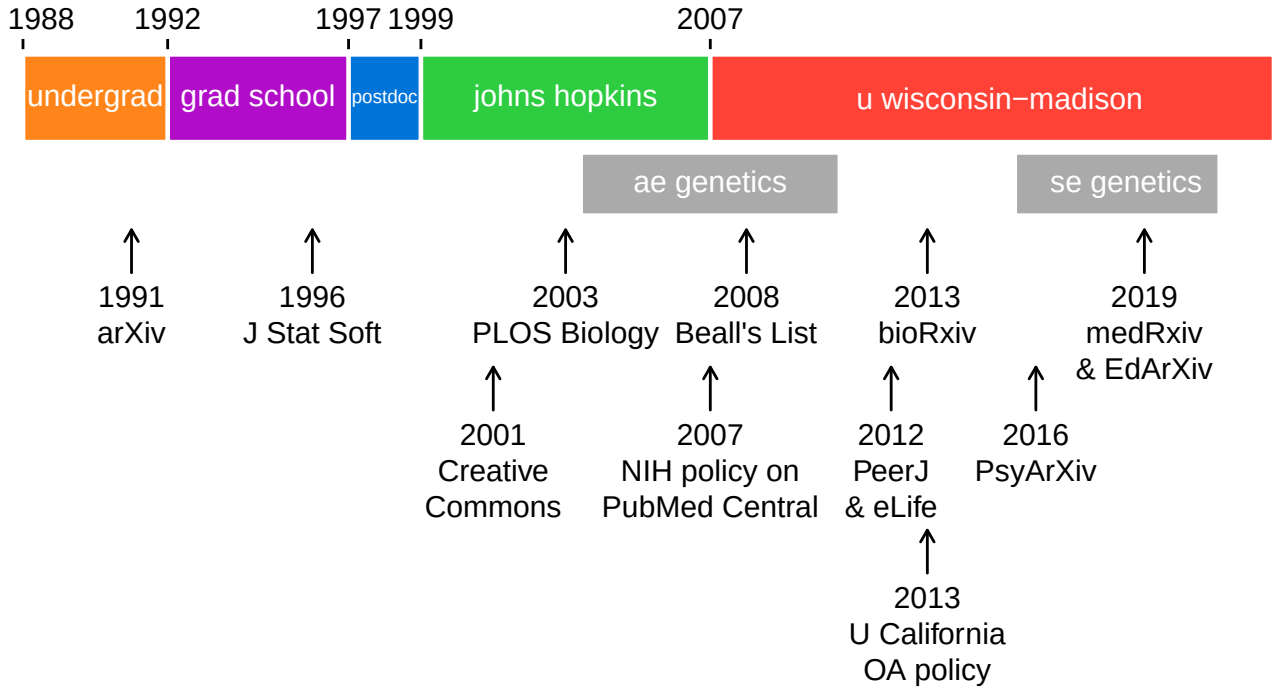
I struggled with the decision of whether to pay this fee. But I'm glad that I did.

I think audience for this paper was made much more widespread by being a formal paper (rather than a preprint, and before that basically a blog post).

bit.ly/3sIRtVY

If you want to more about this paper, see my twitter thread on 10 fun facts about the paper.

Personal timeline



My academic career is basically coincident with the history of open access. This is my personal timeline, along with key events in the history of open access. I've worked on the editorial boards of a half dozen journals, but I include here just my work for the journal Genetics, as that work was most substantial and formative, for me.

The beginnings of OA are basically the beginnings of the internet. The Journal of Statistical Software began in 1996 and has been online-only, open access, and free, with no APCs.

The start of Creative Commons and PLOS are, to me, the start of the broader OA movement. The 2007 NIH policy requiring that funded manuscripts be deposited in PubMed Central was both exciting and disappointing (disappointing for the one-year embargo).

The connection between OA and predatory publishers, and the initiation of Beall's infamous list, is my recent than I remembered.

The start of PeerJ, eLife, and bioRxiv marked a second period of hope and disappointment. medRxiv and the COVID-19 pandemic brought a third wave of change.

What is new?

- Rise of preprints in biology and medicine
- Rise of *Nature Communications*
- PubMed Central: expansion, with no embargo
- No longer stigma on OA
- Emphasis on computational reproducibility

There have been a number of new developments in the last 5 or so years. The use of preprints has really taken off, particularly with the COVID-19 pandemic. The bioRxiv preprint repository had become quite popular in computational biology, but not its use seems much more broad, and biomedical research generally has finally begun to embrace preprints.

At the same time, Nature Communications and related publications like Scientific Reports seem to be siphoning away papers from the PLOS and society journals in the biological sciences. I think this is due to the sparkle of Nature plus the ease of transfer after rejection by one of the glam Nature journals. The APCs are jaw-dropping, but researchers don't seem bothered.

It looks like the PubMed Central idea will be expanded to all government-funded work, and with no embargo. That could really shake up journals' approaches to funding.

The stigma on open access, in which OA was equated with low quality predatory publishers, has largely disappeared.

A related development has been an increased emphasis on computational reproducibility in the sciences, which has led to broader adoption of openness in science.

What isn't new?

- Attachment to Journal Impact Factor
- Attachment to Glam Journals
- Journal and conference spam
- The 20 open access enthusiasts on campus
- Researchers don't read much

Still, people focus on journal impact factors and glam journals. People complain about the focus on impact factors and then turn around and say things like "They have 3 Nature papers" when evaluating job candidates.

The journal Genetics recently sent out a newsletter that included an announcement of new associate editors. The bio for one of them had "...including more than 30 articles in Nature, Science, Cell, and Nature Genetics."

OA no longer has the stigma of predatory publishing, but it seems like 95% of the email I receive is journal or conference spam.

And while OA publishing has broadened, it still seems like most academic researchers don't much care. Hold a forum on open access publishing, and you'll likely be talking to the same 20 people as the last time.

And researchers don't read much, which is the main reason to focus on Impact Factors. It's hard to evaluate the work itself; easier to just evaluate the reputation of the journal in which it appeared. As researchers have become increasingly specialized, it's become ever harder to evaluate our colleagues' work.

Culture of open scholarship

- Community before individual
- Sharing makes better science
  - Data, methods, software, materials, manuscripts

The culture of open scholarship generally places the needs of the community before the needs of an individual: be willing to make some short-term personal sacrifices in order to achieve larger, long-term benefits for the community.

A central idea is that early and broad sharing of data, methods, and results will make for better science. We can't anticipate all possible uses of the data we generate. By making our readily available to others, in a form that is inter-operable with others' data, science as a whole will advance more rapidly.

We should focus on solving problems and gaining knowledge, above getting credit.

Traditional scholarship

What's in it for me?

But that view, of community before individual, appears rare.

The modern approach to science is centered on advancement of an individual's research group. Collaboration is useful if it advances the individual's career, and not otherwise.

While this is a gross simplification, it is also a good first approximation, and is useful for thinking about strategies to get university faculty to change their behavior.

I can't stand the word "incentivize," but that seems to be the way universities operate.

Barriers to open scholarship

- Focus on glamour/prestige
- Apathy
- Ignorance
- Concern about being scooped
- Cost
- Funding of scientific societies

How to persuade?

- Moral arguments
- Advantages for the researcher
- Institution policies
- Government policies

How to turn a successful capitalist into a socialist?

While we might think that we can point to the university's mission and ideals, talking about the Wisconsin Idea and Sifting and Winnowing and such, in practice it appears ineffective.

More successful is to persuade by appealing to personal benefits that accompany open scholarship, for example that OA publications are more widely read and cited.

I'm inclined to think that real change will come from top-down policies that require openness. Funding agencies recognize that they will get more from their investments if they require sharing of data and research products. I wish this weren't necessary, but it seems to be the case.

Privilege

white, male, US-born full professor
in cargo shorts and a hoodie
whose father was a university professor

credentials seldom questioned

I should point out the considerable advantages that I've had, and that my situation gives me considerable flexibility in adopting a completely open approach in my scholarship. When proposing solutions, we need to recognize the very different situations experienced by junior and senior faculty, and between black women and white men.

It bugs me to no end when senior faculty suggest solutions like "just remove journal titles from our CVs" and "stop sending papers to the big three journals" as these aren't real solutions for most people. Journal titles are like the schools people attended; we can wish they don't matter, but they do.

Questions

- How to relax reliance on journal prestige?
- How to support junior faculty to be open scholars?
- How to reorganize the way publishing is funded?
- How to persuade researchers to care?

I'll end with some of my own questions.

I'm sure you can sense my frustration with the state of academic science. Looking back, we have seen some important progress towards openness of data and publications, achieved through both bottom-up innovation and top-down regulation.

But the continued focus on glam journals and journal impact factors seems hard to break. The ever-increasing specialization of our research makes it ever more difficult to evaluate others' work, and so we continue to focus on short-cuts like journal impact factors.