

Bridging R and Excel Projects



Jameel Alsalam

May 16, 2018

EPA R Users' Group

Jameel Alsalam

- Economist, Office of Air and Radiation, Climate Change Division
 - Reports on U.S./int'l GHG emissions projections
 - Oil and gas regulatory impacts analyses
- Me, pre-2015: Staring at Excel all day
- Me, since 2016-ish: Staring at RStudio all day
- R & tidyverse enthusiast



The R/Excel Divide

“On teams with both R and Excel users, how do we do our work?”



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

<http://github.com/jalsalam/BridgingRandExcel>

The R/Excel Divide (2)

R	Workflow Element	Excel
scripts	Approach	Point-and-click
Rmarkdown	Document Creation	Paste tables -> word
Git / Github	Collaboration	Sharepoint
scripts	Independent Components	workbook
Rstudio projects, packages	Combining Pieces Together	Links





















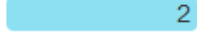
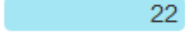


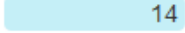






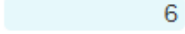




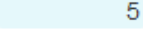






Outline

- Package landscape
- (1) Excel as input
- (2) Excel as output
- (3) Collaboration workflows
- My Questions

R/Excel Package Landscape

- Many packages address this space
- [“a million ways to connect R and Excel”](#)
- readxl, writexl, openxlsx, RExcel, XLConnect, xlsx, gdata, RODBC, BERT, officeR, excel.link, WriteXLS, RDCOMClient, tidyxl

ropenscilabs/packagemetrics comparison

package	published	dl_last_month	stars	tidyverse_happy	has_tests	vignette	last_commit	last_issue_closed	contributors	depends_count	reverse_count
readxl	2018-04-20	 307846	 442					 0.1	 21		 45
openxlsx	2017-03-23	 144296	 225					 0.7	 11	1	 32
gdata	2017-06-06	 110441								1	 55
xlsx	2014-08-02	 83023								 2	 22
RODBC	2017-05-05	 29717								1	 14
XLConnect	2018-04-05	 21912	 73					 1.3	 11	 2	 6
writexl	2018-05-10	 3524	 88					 0.1	 5		 2
excel.link	2017-05-01	 843	 29					1.9	3	 3	0

Blog post about the packagemetrics project: <https://ropensci.org/blog/2017/06/27/packagemetrics/>

My Favorites

Function	Package
Read in .xls/.xlsx data	readxl
Write output	openxlsx

Excel input with readxl

- Alternate PPT title: “an ode to Jenny Bryan”
- Great documentation: <http://readxl.tidyverse.org/>
 - .xls (legacy binary format through ~Excel 2007)
 - .xlsx (current XML-based format)
- Webinar: [What’s new with readxl](#) (November 2017)
- Vignette: [readxl workflows](#)
- Focus is importing a data rectangle, does not expose formatting or other complex aspects of Excel files.



Ideal Spreadsheet has Data in Upper Left

	A	B	C	D	E	F
1	Name	Profession	Age	Has kids	Date of birth	Date of death
2	David Bowie	musician	69	TRUE	1/8/1947	1/10/2016
3	Carrie Fisher	actor	60	TRUE	10/21/1956	12/27/2016
4	Chuck Berry	musician	90	TRUE	10/18/1926	3/18/2017
5	Bill Paxton	actor	61	TRUE	5/17/1955	2/25/2017
6	Prince	musician	57	TRUE	6/7/1958	4/21/2016
7	Alan Rickman	actor	69	FALSE	2/21/1946	1/14/2016
8	Florence Henderson	actor	82	TRUE	2/14/1934	11/24/2016
9	Harper Lee	author	89	FALSE	4/28/1926	2/19/2016
10	Zsa Zsa Gábor	actor	99	TRUE	2/6/1917	12/18/2016
11	George Michael	musician	53	FALSE	6/25/1963	12/25/2016
12						
13						
14						

Source: adapted from readxl example spreadsheet

<http://github.com/jalsalam/BridgingRandExcel>

Spreadsheets Often Untidy

	A	B	C	D	E	F	G
1			Age	<i>Child</i>		<i>Adult</i>	
2			Survived	<i>No</i>	<i>Yes</i>	<i>No</i>	<i>Yes</i>
3	Class	Sex					
4	<i>1st</i>	<i>Male</i>		0	5	118	57
5		<i>Female</i>		0	1	4	140
6	<i>2nd</i>	<i>Male</i>		0	11	154	14
7		<i>Female</i>		0	1		80
8	<i>3rd</i>	<i>Male</i>		35			75
9		<i>Female</i>		17	1		76
10	<i>Crew</i>	<i>Male</i>		0	0	670	192
11		<i>Female</i>		0	0	3	20
12							

All women in the crew worked in the victualling department.

Source: [tidyxl README](https://github.com/jalsalam/BridgingRandExcel)

<http://github.com/jalsalam/BridgingRandExcel>

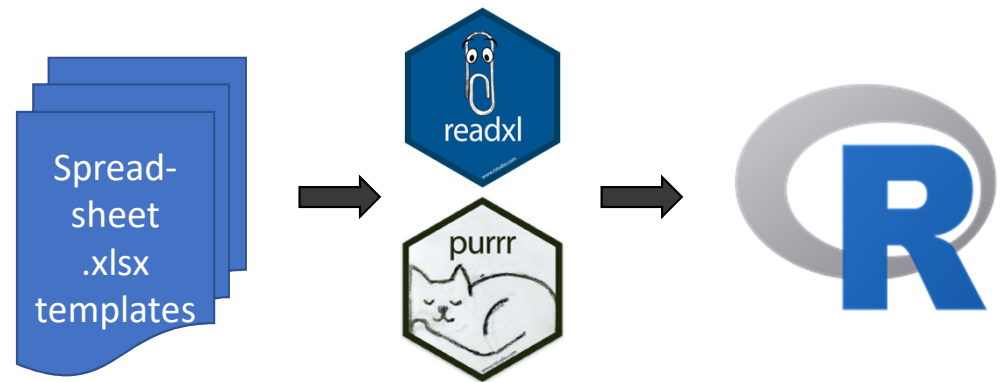
Structured Template

	A	B	C	D	E	F	G	H	I	J
1	Inventory Database Input Template					7/10/2017				
2	At the Source Lead's option you may use insert and populate the Input tab or the Transposed Input tab in your source workbook									
3	Template Instructions:									
4										
5										
6										
7										
8	Source									
9	Author Name									
10	Author Phone									
11	Publication Year									
12	Report Type									
13										
14	Chemical Production and Use									
15	Gas Quantities (Tg)									
16	Sector	Source	Subsource	Fuel	Subref	GHG Gas	1990	1991	1992	1993
17	Energy	Coal Mining	Coal Mining		Underground Liberated	CH4	3.23E+00	3.17E+00	3.12E+00	2.74E+00
18	Energy	Coal Mining	Coal Mining		Underground Recovered &Use	CH4	-2.66E-01	-2.82E-01	-3.29E-01	-4.35E-01
19	Energy	Coal Mining	Coal Mining		Surface Mining	CH4	4.30E-01	4.07E-01	4.04E-01	3.99E-01
20	Energy	Coal Mining	Coal Mining		Post-Mining (Underground)	CH4	3.68E-01	3.48E-01	3.48E-01	3.01E-01
21	Energy	Coal Mining	Coal Mining		Post-Mining (Surface)	CH4	9.32E-02	8.81E-02	8.74E-02	8.65E-02
22										
23										
24										

[old draft example from ongoing project]

<http://github.com/jalsalam/BridgingRandExcel>

Project Workflow: Many Spreadsheets



- Organize spreadsheets with standardized output format
- readxl + purrr to roll up standardized results from many analyses
- Benefits from both Excel and R:
 - Individuals can work independently on their own workbooks
 - Perform updates of cross-cutting calculations quickly and easily by re-running the R script

[jump to deaths.xlsx and many-sheets.Rmd]

Pitfalls & Best Practices

- Paper: [Data Organization in Spreadsheets](#) (Broman & Woo, 2017)
- Webinar: [Data Rectangling](#) (Bryan, 2018)
- Upgrade .xls -> .xlsx (Office Open XML specification)
- Separate the sensitive data in encrypted files
- Use a structured template
- readxl can guess wrong type for mostly-blank columns
- Stray notes/data can surprise
- Design code to fail loudly when data doesn't meet expectations

Excel as Output

- I **like** the way Excel tables look.
- Workflow 1: Raw output, formatting in Excel
- Workflow 2: Full formatting w/openxlsx or XLConnect

[jump to Table 2-1.xlsx and simple-output.Rmd]

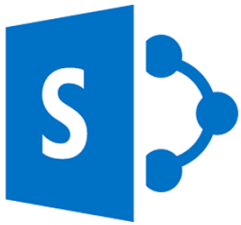
Other Approaches

(things I've only experimented with or read about)

- Fully-formatted output with: `openxlsx`
- Insert tables into word doc: `officeR/flextable`
- Dependency-free output: `writexl`
- Wrangle poorly-formatted data: `tidyxl`
- Run Excel from R: `RDCOMClient`
- Call R from Excel: Basic Excel R Toolkit ([BERT](#))

Collaboration

- Sharepoint v. Git/Github



Two Collaboration Workflows

#1 – All in Github repo

Store both data and code in repo

Pros:

- Self-contained, reproducible analysis
- Good if data files aren't too large and don't change too much

Cons:

- Requires users to learn git
- Git can't diff Excel files

#2 – Github + Sharepoint

Data in sync'd Sharepoint; code in repo

Pros:

- Different files in their 'natural' homes
- Easier for non-git users

Cons:

- Data and code are separated, less reproducible
- More complex permissions/manage

My Questions

How do your teams work across R and Excel/Office?

Input:

- How to access EPA Sharepoint directly (without syncing locally)?
- A good way to work with password-protected files?

Output:

- Make Excel output play nice with Git? (e.g., re-running shouldn't trigger a file change)
- Good approach to pretty-formatted tables?
- How to update a Word doc with updated tables?

Thank you!

Jameel Alsalam (Alsalam.Jameel@epa.gov)