



The Environmental Data Initiative (EDI)



How to clean and format data using R, OpenRefine,
Excel



EDI is funded by the NSF DEB



5 Phases of Publishing Ecological Data

1. Assemble data and metadata
2. **Format and QC data tables**
3. Create EML metadata
4. Submit your data package (data and metadata) to repository
5. Cite your data package

<https://environmentaldatainitiative.org/resources/assemble-data-and-metadata/>

A horizontal bar with a teal segment on the left and an orange segment on the right.

Overview

1. Common data set issues
2. Features of clean ecological data for archiving
3. Tools for cleaning data: Suitability and limitations
4. Resources
5. Data cleaning exercise
6. Discussion

ugly.xls

	E	F	G	H	I	J
1	My Ugly Data Spreadsheet					
2	Date: 5/23/2005					
3	Meter Type	YSI_Model_30		cond_bot	586	
4	Tide State	slack-high		conductivity_top	<30	
5	Salinity	Bottom	0.6	Top	0.1	
6						
7	Date:	10/2/2005				
8	TIDESTATE	-0.34		MeterType	YSI Model 30	
9	Salinity_bottom	0.3		Conductivity Top	349	
10	Salinity_Top	None		cond_bot	39%	
11						

Sheet1 / Sheet2 / Sheet3 /

Common data set issues

From J. Porter: “Creating clean data for archiving”, webinar (1) in EDI series on “5 phases of data publishing”.

<https://www.youtube.com/channel/UCNZoWPaMG6IkEiH8xRNnrrA>

- Inconsistencies abound!
 - Dates sometimes in cells with label, sometimes not
 - Labels for values change between mini-tables
 - Data codes change (YSI_Model_30 is not YSI Model 30)
 - Symbols (% , <) mixed in with numbers

A short horizontal bar with a teal-to-orange gradient.

Features of clean ecological data for archiving

Structure

- Easy-to-use-structure (rectangular), new data adds rows not columns.
- Each variable has its own column, each data point has its own cell.
- Easy to maintain and update, i.e. consistent over time.
- Data archiving format might differ from analysis format (subset, human readable).

File types

- text/csv



Features of clean ecological data for archiving

Data values

- One data type per column (for example integer, real, character, categorical).
- Missing data have consistent missing value codes assigned.
- Error-free and unambiguous computer readable format.
- QC flags.
- Precision meaningful.
- Useful time and date format.

Cook et al. (2001): Best practices for preparing ecological data sets to share and archive.
Bulletin of the Ecological Society of America, p. 138-141.

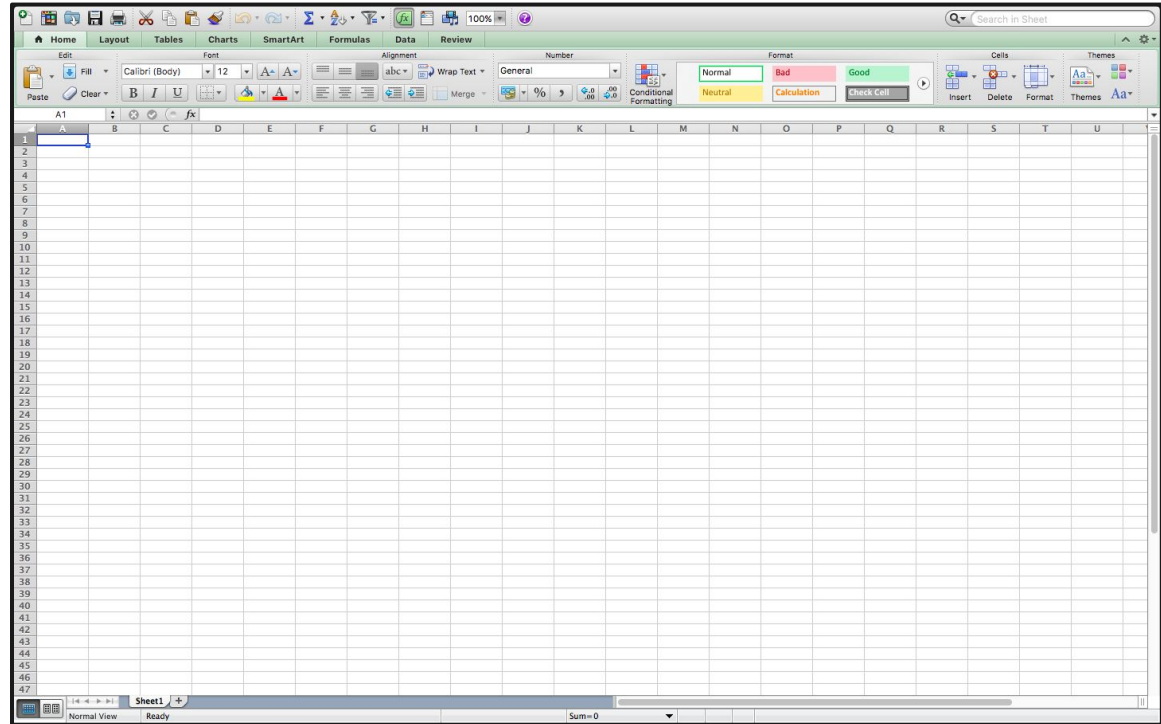
A horizontal bar with a teal segment on the left and an orange segment on the right.

Tools for cleaning data

- **Microsoft Excel (or other spreadsheet software)**
 - <https://products.office.com/en-us/excel>
- **OpenRefine**
 - <http://openrefine.org>
- **R**
 - <https://www.rstudio.com>
 - <https://cran.r-project.org>



Microsoft Excel



A horizontal bar with a teal segment on the left and an orange segment on the right.

Microsoft Excel - Suitability

- Supports data manipulation for non-programmers.
- Initial editing if data set is messy, for example having irregular structure.
- Small data sets.
- For looking at data in a tabular structure.
- One time transformation (if data set is not too large and complex).
- Simple statistical analysis.
- Quick plots.
- Basic data transformations are easily learned.

A horizontal bar with a teal segment on the left and an orange segment on the right.

Microsoft Excel - Limitations

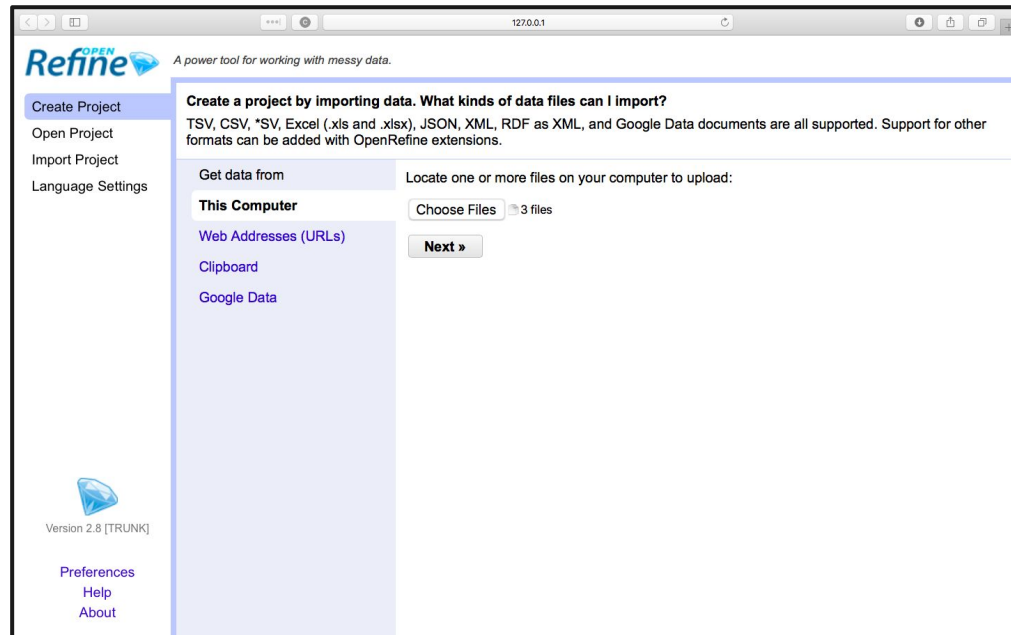
- Limited to certain computer platforms, not open source.
- Reproducibility not guaranteed, manual data manipulation is prone to errors.
- Not suitable for repetitive workflow.
- Publication ready and diverse plots not easily achieved.
- Advanced statistical analysis not possible.
- Handling of large data sets difficult.
- Not flexible in data operations.



OpenRefine

<http://openrefine.org>

<http://www.datacarpentry.org/OpenRefine-ecology-lesson/>





OpenRefine

The screenshot shows the OpenRefine web interface. The browser address bar shows '127.0.0.1'. The page title is 'Table_1 csv'. The interface includes a 'Facet / Filter' sidebar on the left with a 'Using facets and filters' section. The main area displays a table with 204 rows. The table has columns: 'id', 'date', 'temperature', 'relative_humidity', 'photosynthetic', and 'flag'. The data is sorted by 'date' in descending order. The table shows data for various dates in December 2017, with values for temperature, relative humidity, and photosynthesis. The interface also includes a 'Show as: rows records' dropdown and a 'Show: 5 10 25 50 rows' dropdown. The 'Extensions:' button is visible in the top right corner.

Using facets and filters


Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

204 rows

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last »

		id	date	temperature	relative_humidity	photosynthetic	flag
1.	Table_A.csv	site_1	2017-12-01	-6.53	82.5	443	
2.	Table_A.csv	site_1	2017-12-02	-8.34	76.8	207	
3.	Table_A.csv	site_1	2017-12-03	-3.85	75.2	373	
4.	Table_A.csv	site_1	2017-12-04	-3.13	91.7	417	
5.	Table_A.csv	site_1	2017-12-05	-2.11	100.3	228	
6.	Table_A.csv	site_1	2017-12-06	-1.12	84	481	
7.	Table_A.csv	site_1	2017-12-07	3.04	73.3	302	
8.	Table_A.csv	site_1	2017-12-08	-0.4	100.4	477	
9.	Table_A.csv	site_1	2017-12-09	-9.08	76.2	475	
10.	Table_A.csv	site_1	2017-12-10	1.06	87.2	440	
11.	Table_A.csv	site_1	2017-12-11	-0.43	96.8	361	
12.	Table_A.csv	site_1	2017-12-12	-7.4	81.1	209	
13.	Table_A.csv	site_1	2017-12-13	-2.96	86.2	275	
14.	Table_A.csv	site_1	2017-12-14	-4.39	95.7	443	
15.	Table_A.csv	site_1	2017-12-15	-3.77	81.5	261	
16.	Table_A.csv	site_1	2017-12-16	-4.82	76.6	483	
17.	Table_A.csv	site_1	2017-12-17	-4.96	70.9	497	
18.	Table_A.csv	site_1	2017-12-18	-9.14	76.7	287	
19.	Table_A.csv	site_1	12/19/17	0.95	95.7	421	
20.	Table_A.csv	site_1	12/20/17	0.22	71.6	453	
21.	Table_A.csv	site_1	12/21/17	1.76	96.7	314	
22.	Table_A.csv	site_1	2017-12-22	0.96	86.8	315	

A short horizontal bar with a teal segment on the left and an orange segment on the right.

OpenRefine - Suitability

- Open source web application, designed to run on local computer (platform independent)
- Supports data manipulation and parsing for non-programmers.
- Offers common data-munging tasks in a menu based format.
- Offers faceting and clustering algorithms for:
 - easy browsing, data cleaning, General Refine Expression Language (GREL)
- Complete provenance/undo history of transformations/modifications is available and interactive or by saving a script.
 - For repetitive tasks previous actions can be reapplied.
- Relatively intuitive & easy to learn.



OpenRefine - Limitations

- Plotting features are very limited.
- Publication ready and diverse plots not possible.
- Advanced statistical analysis not possible.
- Not flexible in data manipulation operations.



R

- Some R useful packages:
 - tidyverse (dplyr, tidyr, ggplot2)
 - dataMaid
- Cheat sheets:
 - <https://www.rstudio.com/resources/cheatsheets/>
 - https://cdn.rawgit.com/EDlorg/tutorials/3e4c1299/data_cleaning/cheatsheet_dataMaid.pdf
- Data Carpentry lesson “R for data analysis and visualization of Ecological Data”:
<http://www.datacarpentry.org/R-ecology-lesson/>

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The top toolbar contains icons for file operations, running code, and other functions. The top status bar shows the current project path: ~/EDI/r/taxonomyCleanr - master - RStudio.

The left pane displays the R console output, which includes the R version (3.2.5), copyright information, platform details (x86_64-w64-mingw32/x64), and a welcome message. It also provides instructions on how to use R, including how to get help and how to quit.

The right pane shows a script editor with a file named 'detect_os.R'. The script contains a function 'detect_os()' that uses 'Sys.info()' to detect the operating system and returns a character string representing the OS. The script is as follows:

```
1 #' Detect operating system
2 #'
3 #' @description
4 #'   This function uses \code{Sys.info} to detect the
5 #'   operating system and outputs an abbreviated character string to
6 #'   specific function calls.
7 #'
8 #' @usage detect_os()
9 #'
10 #' @return
11 #'   \item{win}{Windows OS}
12 #'   \item{mac}{Mac OS}
13 #'
14 #' @export
15 #'
16
17 detect_os <- function(){
18   sysinfo <- Sys.info()["sysname"]
19   if (sysinfo == "Darwin"){
20     os <- "mac"
21   } else {
22     os <- "win"
23   }
24   os
25 }
```

The bottom pane shows the Environment, History, and Git tabs. The Environment tab is active, showing the current workspace. It includes a table of objects in the environment:

Name	Size	Modified
..		
choices2map.R	3.4 KB	Dec 20, 2017, 2:59 PM
detect_delimites.R	4.1 KB	Dec 11, 2017, 1:03 PM
detect_os.R	519 B	Dec 6, 2017, 11:06 AM
make_taxonomicCoverage.R	2.9 KB	Dec 20, 2017, 3:34 PM
resolve_taxa.R	25.8 KB	Dec 20, 2017, 1:59 PM
update_data.R	7.1 KB	Dec 20, 2017, 3:04 PM
validate_fields.R	2.9 KB	Dec 11, 2017, 1:03 PM
validate_file_names.R	2.4 KB	Dec 6, 2017, 11:06 AM
validate_path.R	936 B	Dec 6, 2017, 11:06 AM

The bottom right pane shows the History tab, which displays a list of executed R commands. The commands include:

```
rm(list = ls())
update_data(path = "C:\\Users\\Colin\\Documents\\EDI\\d.
data.file = "data",
taxon.col = "Species")
devtools::load_all(".")
devtools::load_all(".")
rm(list = ls())
?make_taxonomicCoverage
devtools::load_all(".")
make_taxonomicCoverage(path = "C:\\Users\\Colin\\Docume
document()
library(devtools)
document()
```

A short horizontal bar with a teal segment on the left and an orange segment on the right.

R - Suitability

- **User friendly and comprehensive data analysis and statistics possible for “tidy” data:**
 - Each column is a variable, each row is an observation
- **For repetitive tasks by reapplying programs to different datasets.**
- **Complete provenance of data manipulation available.**
- **Very flexible tool for data manipulation.**
- **Excellent tool for visualization.**
- **Used in research and data science community.**
 - Free, open source scripting language & extended support network and tools.
 - Compiles and runs on a wide variety of computer platforms: Windows, MacOS, Unix, Linux.

A horizontal bar with a teal segment on the left and an orange segment on the right.

R - Limitations

- Steep learning curve.
- Not easy to use with very irregular data sets.
- Time intensive to create the data cleaning program



Data cleaning exercise

Exercise:

- Messy dataset: 4 data tables (A,B,C,D) with problems
- Step-by-step instructions for cleaning with: OpenRefine and R

Download:

- https://github.com/EDlorg/tutorials/tree/master/data_cleaning



Data formatting	R	OpenRefine
Add column	<code>dplyr::mutate</code>	column menu ->edit column->add column
Unify column names	<code>base::which %in%</code>	use text editor or spread sheet before importing tables into OpenRefine
Concatenate tables row wise	<code>base::rbind dplyr::bind_rows</code>	Importing tables with identical column names in same project
Transform from wide to long	<code>tidyr::gather</code>	Column menu->Transpose
Transform from long to wide	<code>tidyr::spread</code>	Column menu->Transpose
Check uniqueness of rows	<code>dataMaid::isKey dplyr::filter + duplicated base::unique</code>	column menu->Facet ->Customized Facet->Duplicates Facet



Cell problems	R	OpenRefine
"White spaces" in cells	<code>dataMaid::identifyWhitespace</code> <code>base::gsub</code>	Column menu ->Facet->Text or numeric facet->modify blank in "Facet/File" window
Multiple missing value codes	<code>dataMaid::identifyMissing</code>	Same procedure as for "White spaces"
Characters in value field	<code>dataMaid::identifyNums</code> <code>lapply(yourData, class) ... are data classes</code> <code>expected?</code>	Same procedure as for "White spaces"
Multiple date/time formats	<code>lubridate::ymd</code> , <code>lubridate::mdy</code> <code>readr::parse_datetime</code> , <code>readr::parse_date</code> , <code>readr::parse_time</code> ... if parsing fails then search for offending entities	Same procedure as for "White spaces" or Column menu->Edit cells->Common transforms->To data
Identify and clear leading and prevailing "white spaces"	<code>base::trimws</code>	Column menu->Edit cells->Common transforms->trim leading and prevailing white spaces"

A short horizontal bar with a teal-to-orange gradient is positioned above the section header.

EDI Resources

- **EDI website on “5 phases of data publishing”**
 - environmentaldatainitiative.org/resources/assemble-data-and-metadata
- **Contact EDI’s data curation team**
 - info@environmentaldatainitiative.org
- **EDI tutorials on gitHub**
 - github.com/EDIdorg/tutorials
- **EDI tutorials on youtube**
 - youtube.com/channel/UCNZoWPaMG6lkEiH8xRNnrrA



Thank you!



environmentaldatainitiative.org



portal.edirepository.org/nis/home.jsp



github.com/EDIdorg



@ EDIgotdata



edi-got-data.slack.com