# 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/

# 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

- 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

**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/UCNZoWPaMG6lkEiH8xRNnrrA

# 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**

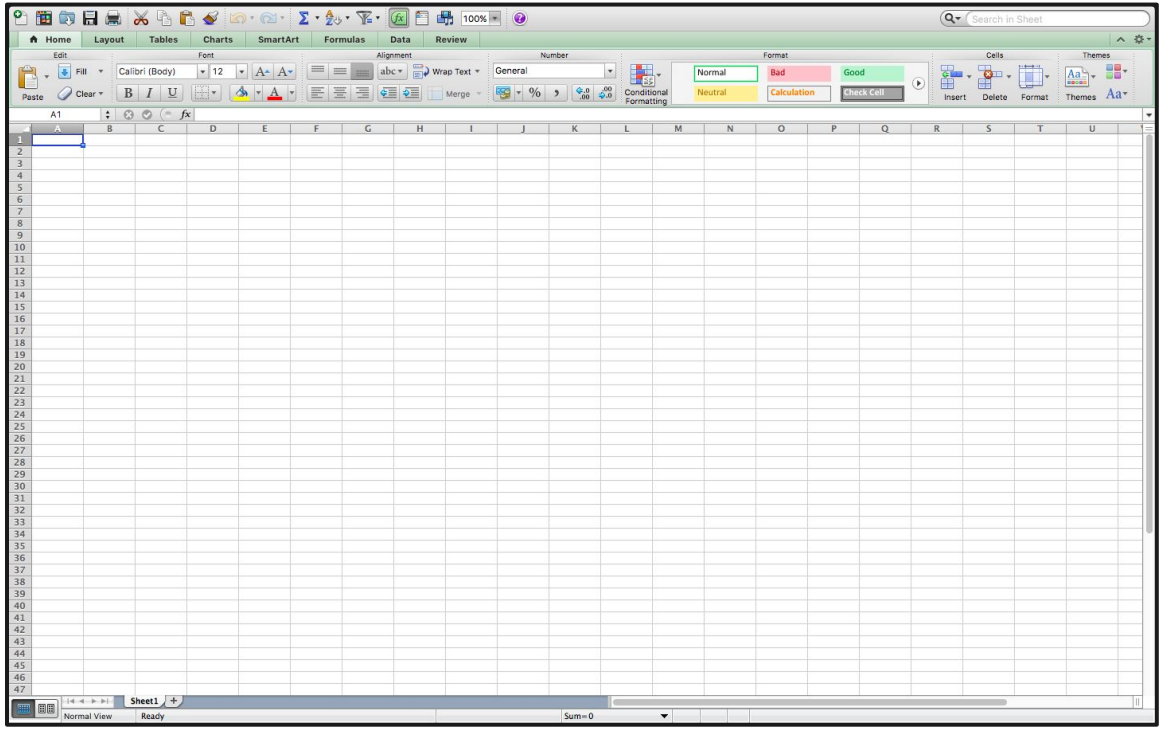
- One data type per column (for example integer, real, character, categorical).
- Missing data have consistent missing value codes assigned.
- Error-free and unambiguous computer readable format.
- QC flags.
- Precision meaningful.
- Useful time and date format.

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.

# 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



# 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.

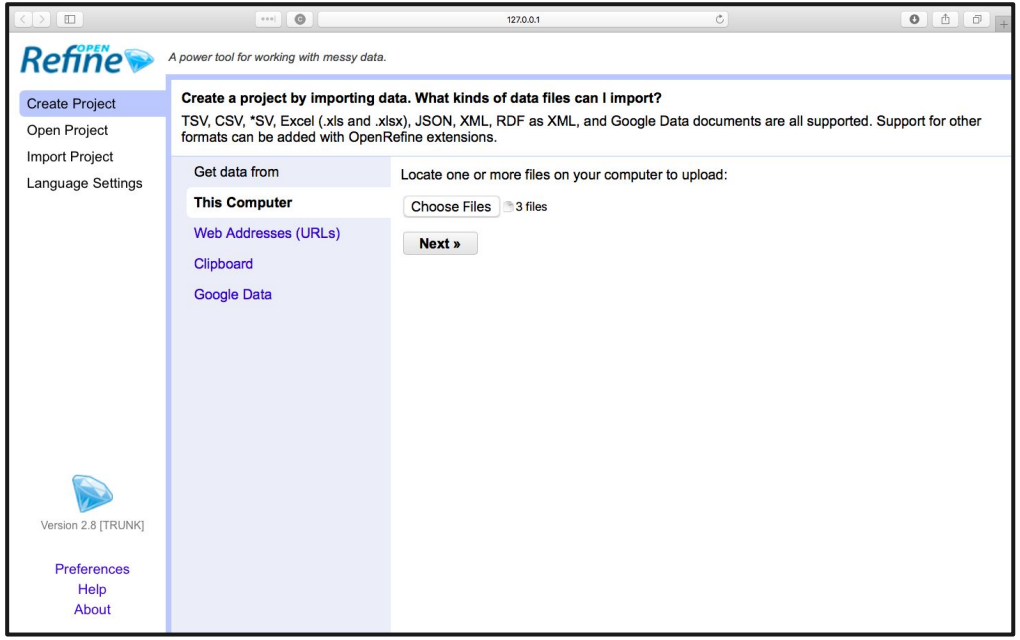
# Microsoft Excel - Limitations

- 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.

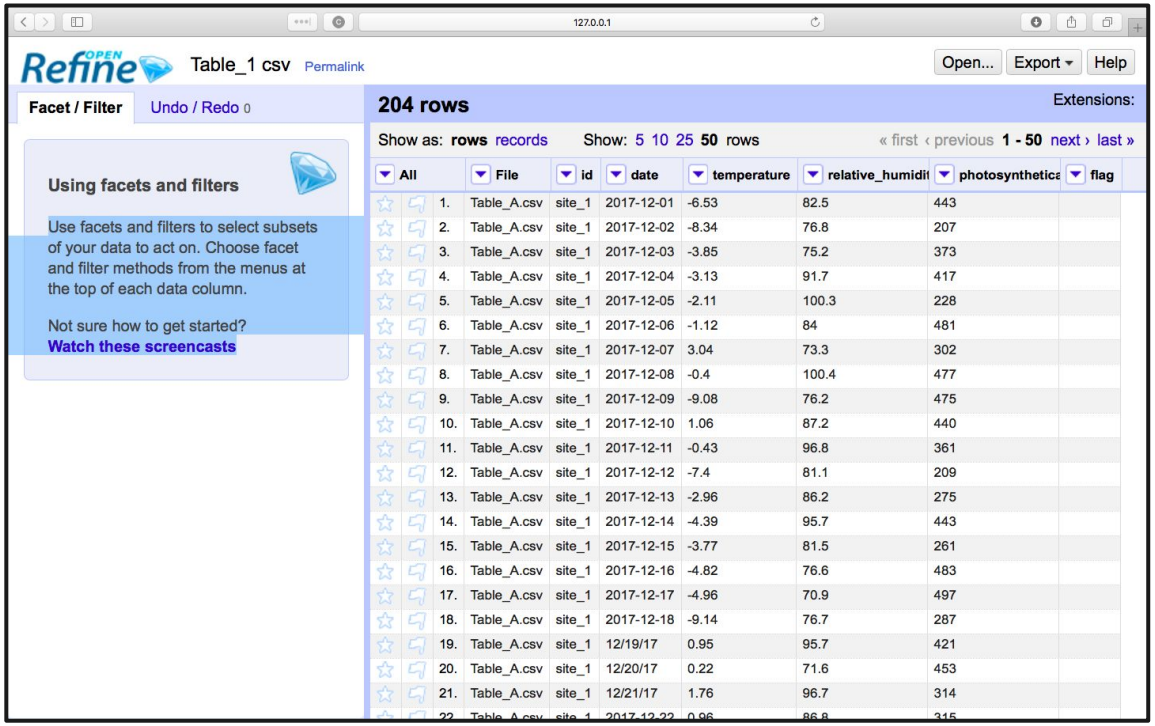
# OpenRefine

http://openrefine.org

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



# OpenRefine




| All | File | id | date | temperature | relative_humidit | photosynthetica | 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 | |

# OpenRefine - Suitability

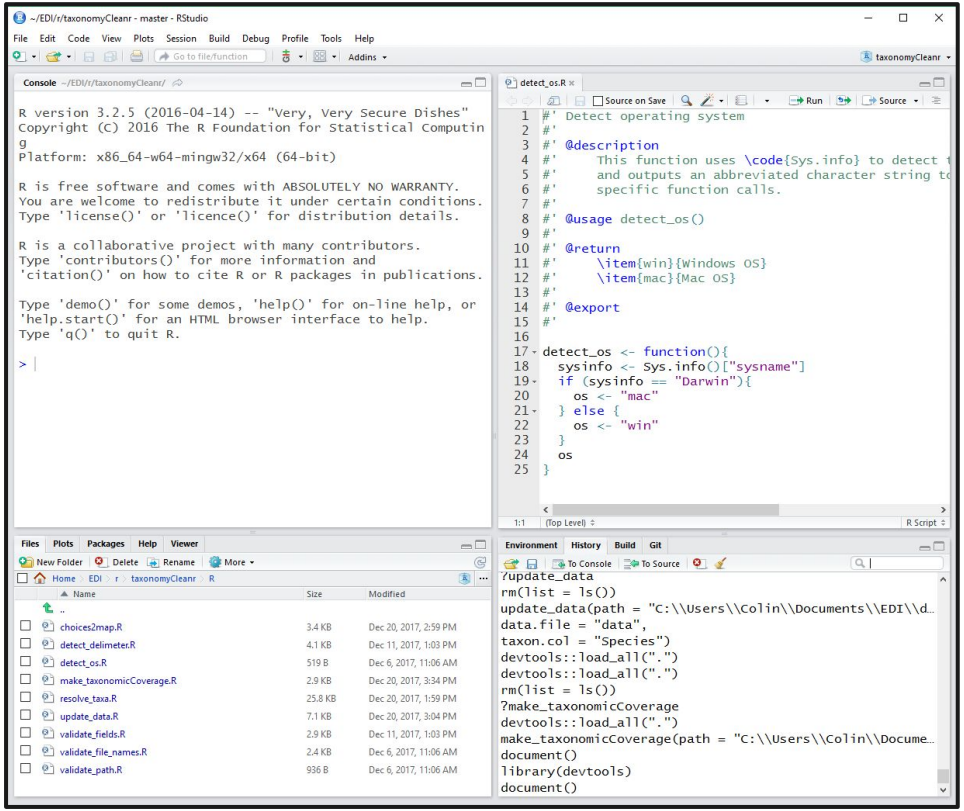
- Open source web application, designed to run on local computer (platform independent)
- Supports data manipulation and parsing for non-programmers.
- Offers common data-munging tasks in a menu based format.
- Offers faceting and clustering algorithms for:
  - easy browsing, data cleaning, General Refine 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/EDIorg/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/



# 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.

# 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/EDIorg/tutorials/tree/master/data_cleaning**

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

| **Cell problems** | R | OpenRefine |
|---|---|---|
| “White spaces” in cells | dataMaid::identifyWhitespace<br>base::gsub | Column menu ->Facet->Text or numeric facet->modify blank in “Facet/File” window |
| Multiple missing value codes | dataMaid::identifyMissing | Same procedure as for “White spaces” |
| Characters in value field | dataMaid::identifyNums<br>lapply(yourData, class) … are data classes expected? | Same procedure as for “White spaces” |
| Multiple date/time formats | lubridate::ymd, lubridate::mdy<br>readr::parse_datetime, readr::parse_date, readr::parse_time<br>… 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 | base::trimws | Column menu->Edit cells->Common transforms->trim leading and prevailing white spaces” |

# 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/EDIorg/tutorials
- **EDI tutorials on youtube**
  - youtube.com/channel/UCNZoWPaMG6IkEiH8xRNnrrA

# Thank you!

environmentaldatainitiative.org

portal.edirepository.org/nis/home.jsp

github.com/EDIorg

@ EDIgotdata

edi-got-data.slack.com