Sharing Data

Siyu Chen, Yukun He,
Aoyi Shan, Shuotong Wu

Team Introduction

- Aoyi Shan, Senior, Business & Statistics
- Diane He, Senior, Statistics
- Shuotong Wu, Senior, Computer Science & Statistics
- Siyu Chen, Senior, Civil Engineering

Nine Simple Ways to Make it Easier to Use your Data

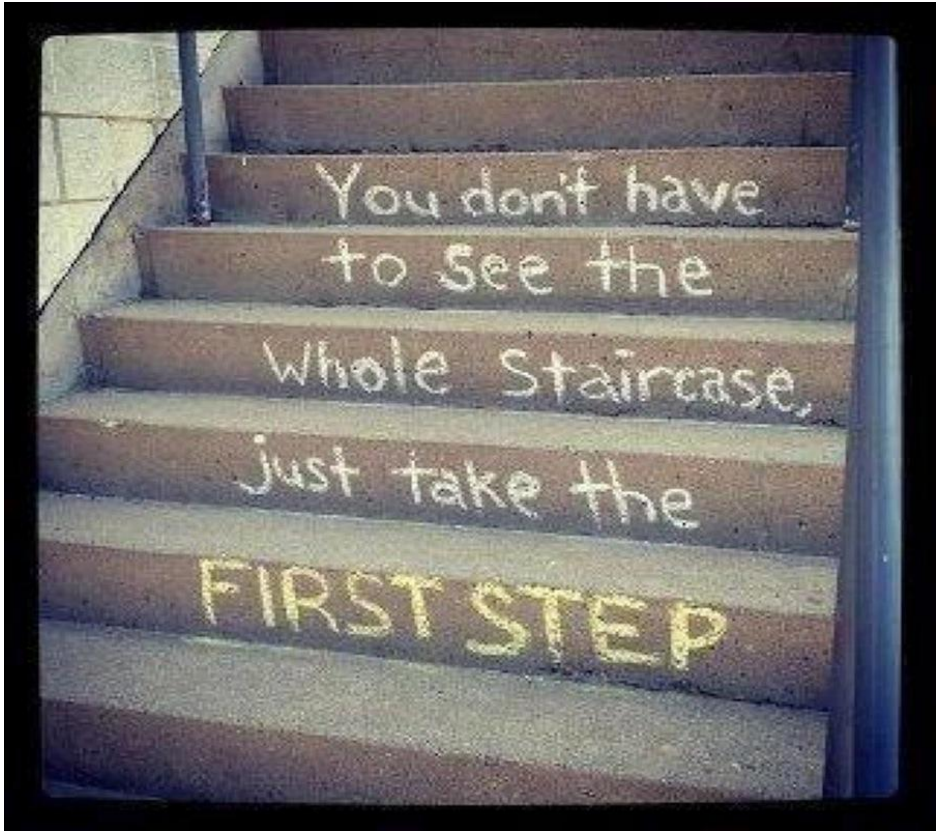
- Published in Ideas in Ecology and Evolution, a peer-reviewed, open-access, non-profit, electronic journal published at Queen’s University
- Authors: Ethan P. White, Elita Baldridge, Zachary T. Brym, Kenneth J. Locey, Daniel J. McGlinn, Sarah R. Supp
  - Department of Biology and Ecology
  - Utah State University

Common mistakes and what we expect to achieve?

- Common Mistakes
  - Data Structure
  - Metadata
  - Licensing
- Targets
  - Make your data understandable
  - Easy to analyze
  - Readily available to the wider community of scientists

1. Share your data

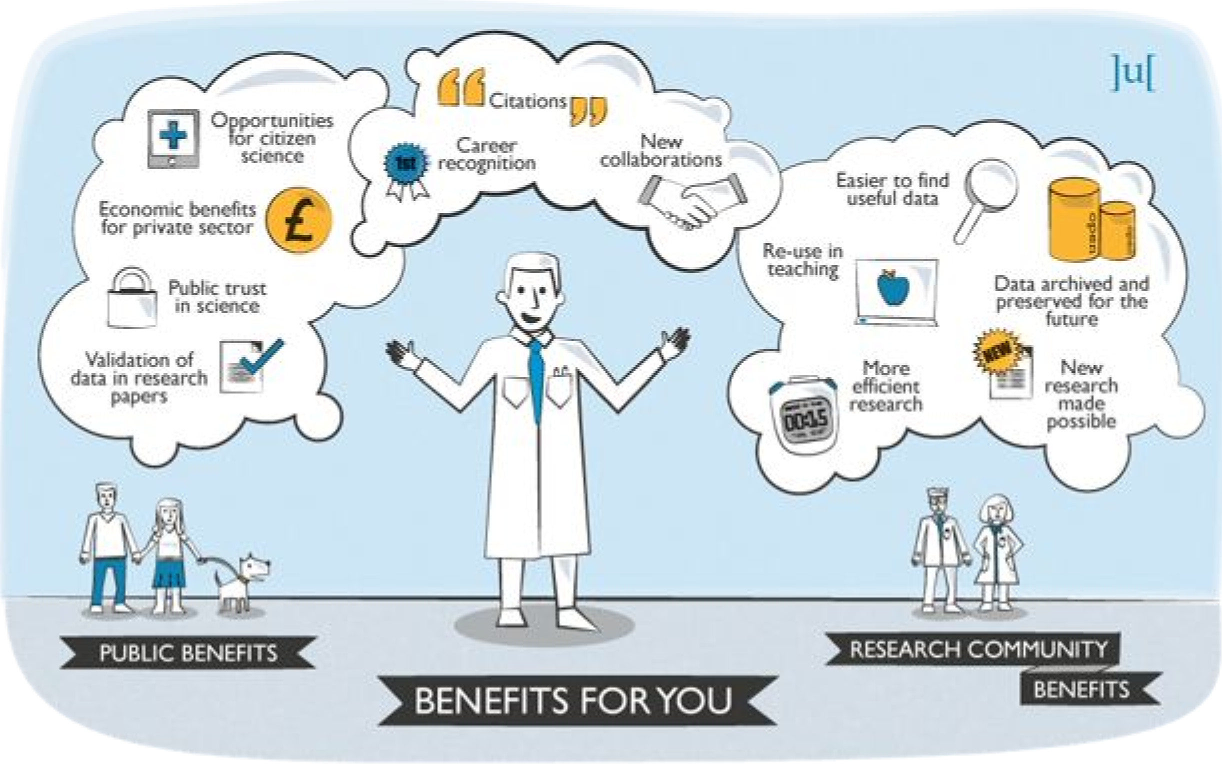
- The first and most important step in sharing your data is to share your data



Major benefits for sharing data

- For the scientific community
  - The ability to reproduce the analysis and results
  - Combined in meta-analysis to reach general conclusions
  - New approaches applied and new questions asked based on the data
- For data collectors
  - Provides credit for publication of data products
  - Make future reuse easier for the original investigator

Major benefits for sharing data



Opportunities for citizen science
Economic benefits for private sector
Public trust in science
Validation of data in research papers
Citations
1st
Career recognition
New collaborations
Easier to find useful data
Re-use in teaching
Data archived and preserved for the future
More efficient research
NEW
New research made possible
]u[
PUBLIC BENEFITS
BENEFITS FOR YOU
RESEARCH COMMUNITY BENEFITS

Why people are reluctant to share their data?

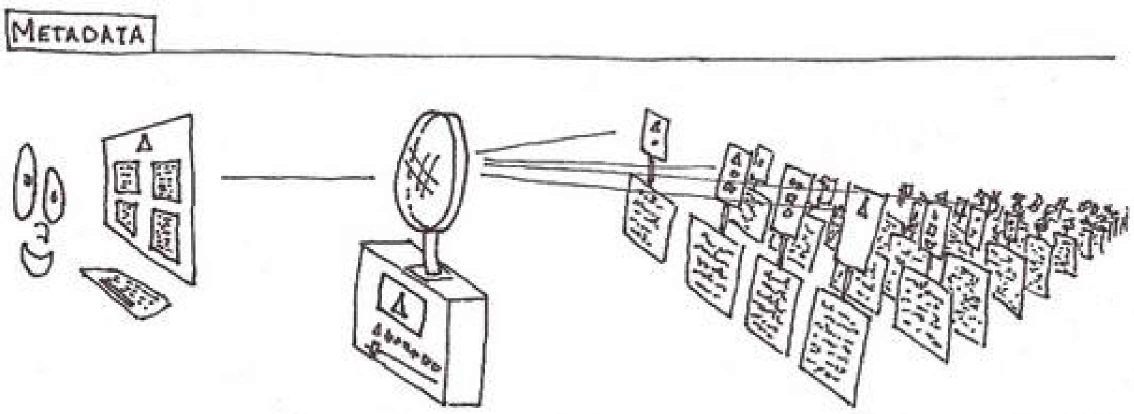
- As a result of all those advantages, data sharing is increasingly required by funding agencies, journals ad potentially by law.
- Why people are reluctant to share their data?
  - Competition for publications based on the shared data
  - A lack of recognition for sharing data
  - A perception that sharing data is technically difficult and time consuming

Solutions to address people’s concerns

- Data embargoes or limitations on direct competition
- Datasets are now considered citable entities and data providers receive recognition in the form of increased citation metrics and credit on CVs and grant applications
- Data archives have become increasingly common and easy to use

Provide metadata

- Metadata is information about the data



john-norris.net CC SA-BY 2.0

Metadata

- Benefits of providing metadata
  - Easier to figure out if a dataset is appropriate for a project
  - Easier to understand how to work with the data
- Metadata can take several forms
  - Descriptive file and column names
  - A written description of the data
  - Images
  - Structured information that can be read by computers

Metadata

- Good metadata should provide the following information
    - The what, when, where, and how of data collection
    - How to find and access the data
    - Suggestions on the suitability of the data for answering specific questions
    - Warnings about known problems or inconsistencies in the data
    - Information to check that the data are properly imported

Provide an unprocessed form of the data

- To make data as useful as possible it is best to share the raw data
    - It can be very difficult to combine data from multiple sources that have each been processed in different ways
    - Providing data in the raw form gives data users the most flexibility
        - To address different questions
        - To develop better approaches to correct the data for common limitations

Provide an unprocessed form of the data

- Provide both the raw and processed forms of the data
  - The processed data is also very important particularly when correcting data for common limitations
  - Clearly explain the differences between the two data forms in the metadata
- Share the unprocessed data along with the code that processes the data to the form used for analysis
  - This allows other scientists to assess and potentially modify the process

Use standard data formats

- It is best to store data in a standard format that can be used by many different kinds of software
- Use standard file formats
  - Use well-defined formats when they exist
    - Certain kinds of data in ecology and evolution have well established standard formats
    - Scientists and most software will be able to work with the data easily
  - Store the data in a format that can be opened by any type of software, i.e. text files

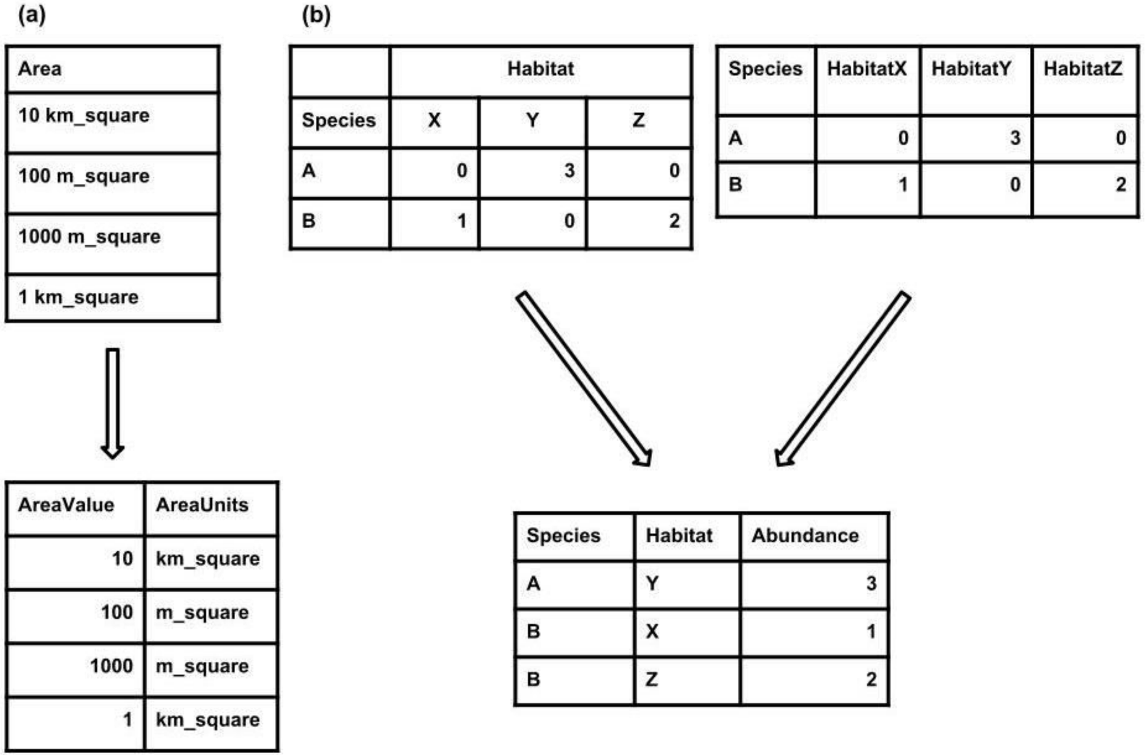
Use standard data formats

- Using formats other than plain text files (e.g, formats used by Excel) will cause problems
  - Data in other format can be difficult to load into other programs
  - Data in other format can be difficult to open if the newer versions of the software no longer support the original format
- Use descriptive names when naming files
  - It is easy to keep track of what data they contain
- Avoid spaces in file names, which can cause problems for some software
  - Spaces in file names can be avoided by using camel case (e.g, RainAvg)
  - Separating the words with underscores to avoid spaces (e.g., rain_avg)

Use standard data formats

- Use standard table formats
  - Characteristics of tabular data
    - It provides flexibility in structuring data
    - It is easy to structure data in a way that is difficult to (re)use
  - Three simple recommendations:
    - Each row should represent a single observation and each column should represent a single variable or type of measurement
    - Every cell should contain only a single value
    - There should be only one column for each type of information

Use standard data formats



Area
10 km_square
100 m_square
1000 m_square
1 km_square

AreaValue	AreaUnits
10	km_square
100	m_square
1000	m_square
1	km_square

	Habitat		
Species	X	Y	Z
A	0	3	0
B	1	0	2

Species	HabitatX	HabitatY	HabitatZ
A	0	3	0
B	1	0	2

Species	Habitat	Abundance
A	Y	3
B	X	1
B	Z	2

Use standard data formats

- Use standard table formats (Example)
    - Do not include units in the cell with the values or include multiple measurements in a single cell
        - There is no easy way for the software to treat the items within a cell as separate pieces of information
    - Avoid cross-tab structured data, where different columns contain measurements of the same variable
    - Make sure to use descriptive column names

Use standard data formats

- Use standard formats with cells
    - Be consistent
        - Capitalization of words
        - Choice of delimiters
        - Naming conventions for variables
    - Avoid special characters
    - When working with dates use the YYYY-MM-DD format

Use good null values

Missing or empty values in dataset is common.

How to represent null value?

1. Compatible with most software
2. Unlikely to cause errors in analysis

Blank, NOT a space. BEST Option

Benefits:

1. Automatically treated as null values by R, python, SQL and Excel.
2. Easily spotted in a visual examination of the data.

Problems:

1. Hard to know if a value is overlooked during data entry
2. Blanks can be confusing when spaces or tabs are used as delimiters

NA, na, N/A

NA, na: **Good Option**

Compatible with R,

But may be confused with an abbreviation(eg. North America, sodium).

N/A : **Avoid**

an alternative but often not compatible with software

Null, None

Null: **Good Option**

Compatible with SQL

A placeholder for unknown values

None: **Avoid**

Uncommon, can cause problems with data type

Compatible with Python

BAD options to Avoid

1. Numerical values, eg. 0, 999, -999.
   a. Indistinguishable from the actual numerical value.
   b. Extra step to remove from analyses and can be accidentally included in calculation
2. Non-standard text indications, eg. No data, Missing, -, +, .
   a. Uncommon
   b. Can cause problems with data type
   c. Some contain a space, eg. No data

Null Values	Problems	Compatibility	Recommendation
Blank	Spaces, Overlooked Entry during data entry	R, Python, SQL	Best Option
Na, na	Abbreviation, Data type	R	Good Option
NULL	Data type	SQL	Good Option

“Whichever null value you use, only use **one**, use it **consistently** throughout the data set, and **indicate** it clearly in the metadata.”

Make it easy to combine your data with other datasets

By Including contextual data that appears across similar data sources

Eg. taxonomy and geographic location(datum and precision).

1. add contextual data as a new column or an additional table
2. include them in metadata.

When contextual data are included in dataset ...

1. Referred as codes or abbreviations to reduce data entry and redundancy

   Eg. a single column for species ID rather than separate detailed columns about that species.

2. But hard to understand without clear definitions

   Include additional tables that contain a column for the code and additional columns that describe the item in standard way.

   Eg. site name followed by latitude and longitude

Perform basic quality control

- Make it easier to analyze your own data and decrease the chance of making mistakes.
- Basic sanity checks:
  - If a column should contain numeric values, check that there are no non-numeric values in the data
  - Check that empty cells actually represent missing data, and not mistakes in data entry, and indicate that they are empty using the appropriate null values
  - Check for consistency in unit of measurement, data type, naming schema, etc.

Use an established repository

- Data should be easy to find, accessible, and stored where it will be preserved for a long time
- Major well-established repositories will guarantee long-term persistence.
- Repositories examples:
    - Host specific data types, such as molecular sequences: DDBJ, GenBank, MG-RAST. They are often highly standardized in data type, format, and quality control.
    - Host a wide array of data types and are less standardized: Dryad, KNB, PANGAEA.
    - All-purpose repositories: figshare

How to choose a repository

- You should consider where other researchers in your discipline are sharing their data.

- No standard. It is worth considering differences among repositories in terms of use, data rights, and licensing.

How to choose a repository

- Use a repository that allows your dataset to be easily cited.
  - A easy way to guarantee that your data are citable is to confirm that the repository associates it with a persistent identifier.
  - The most popular persistent identifier is DOI (digital object identifier).
  - There are also online tools for finding good repositories for your data, like http://re3data.org
- Permanent identifiers: a set of numbers and/or characters, frequently in the form of a URL, that points to the location of a resource. PIDs are set up in such a way that even though the storage location of resource may change over time, the PID will always point to the correct location.

Use an established and open license

- An explicit license with your data is the best way to let others know exactly what they can and cannot do with the data.
- Panton Principles: a set of recommendations that address how best to make published data from scientific studies available for re-use.
- Following Panton Principles:
  - It’s better to use well established licenses in order to clearly communicate the rights and responsibilities of both the people providing the data and the people using it.
  - It’s better to use the most open license possible. Sometimes even minor restrictions on data use can have unintended consequences for the reuse of the data.

What licenses place no restrictions?

What licenses place no restrictions?

- Creative Commons Zero (CC0)

- Open Data Commons Public Domain Dedication and License (PDDL)

- Having a clear and open license will increase the chance that other scientists will be comfortable using your data

Summary

1. Share your data
2. Provide metadata
3. Provide an unprocessed form of the data
4. Use standard data formats
5. Use good null values
6. Make it easy to combine your data with other datasets
7. Perform basic quality control
8. Use an established repository
9. Use an established and open license

The End

Questions?

Thank you for listening!