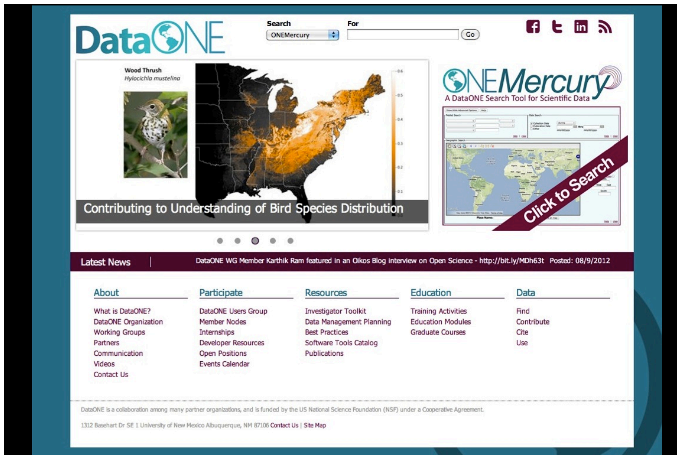
# Data & Data Management

Jarrett Byrnes
9/6/2012



Contributing to Understanding of Bird Species Distribution

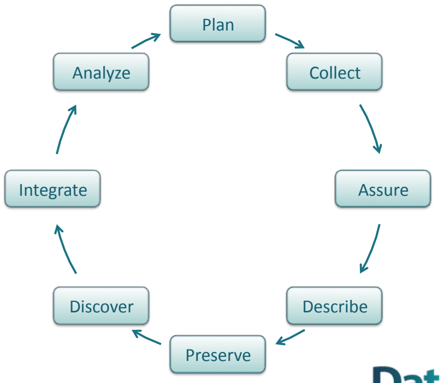
DataONE WG Member Karthik Ram featured in an Oikos Blog interview on Open Science - http://bit.ly/MDh63t Posted: 08/9/2012

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement.

1312 Basehart Dr SE 1 University of New Mexico Albuquerque, NM 87106

http://dataone.org

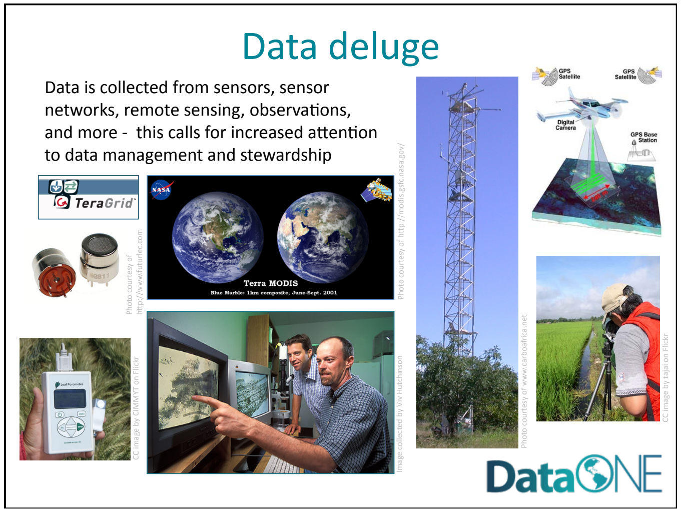
## What is the Data Life Cycle?



## For Each Stage of the Data Lifecycle…

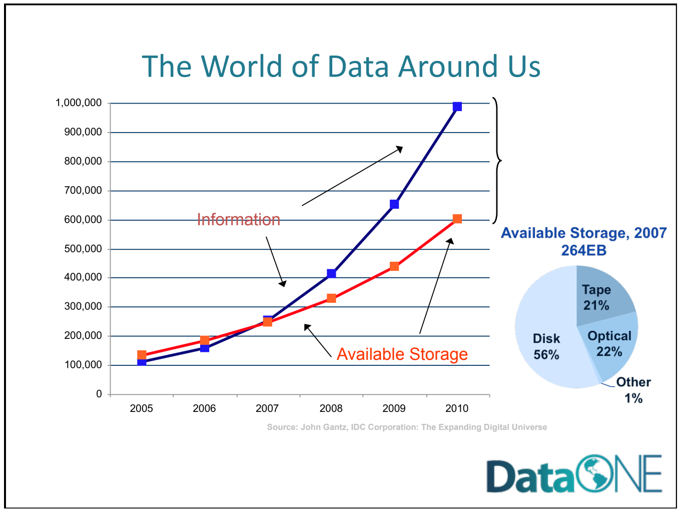
- …there are best practices…..and….tools to help!
- Your well-managed and accessible data can contribute to science in ways you may not even imagine today!
- `http://www.dataone.org/best-practices`
- `http://www.dataone.org/education-modules`

Images collected by DataOne.org



## Data deluge

Data is collected from sensors, sensor networks, remote sensing, observations, and more - this calls for increased attention to data management and stewardship



## The World of Data Around Us

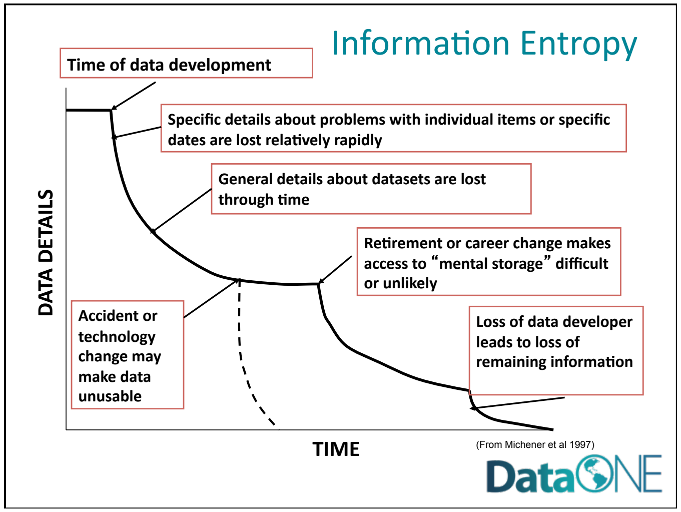
Source: John Gantz, IDC Corporation: The Expanding Digital Universe

## The World of Data Around Us: Data Loss

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations and requirements

CC image by Sharyn Morrow on Flickr

CC image by momboleum on Flickr



## Information Entropy

Time of data development

Specific details about problems with individual items or specific dates are lost relatively rapidly

General details about datasets are lost through time

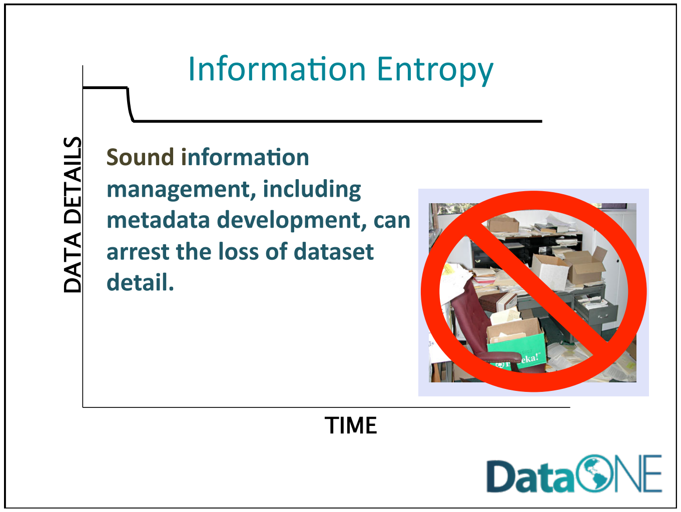
Retirement or career change makes access to “mental storage” difficult or unlikely

Accident or technology change may make data unusable

Loss of data developer leads to loss of remaining information

(From Michener et al 1997)

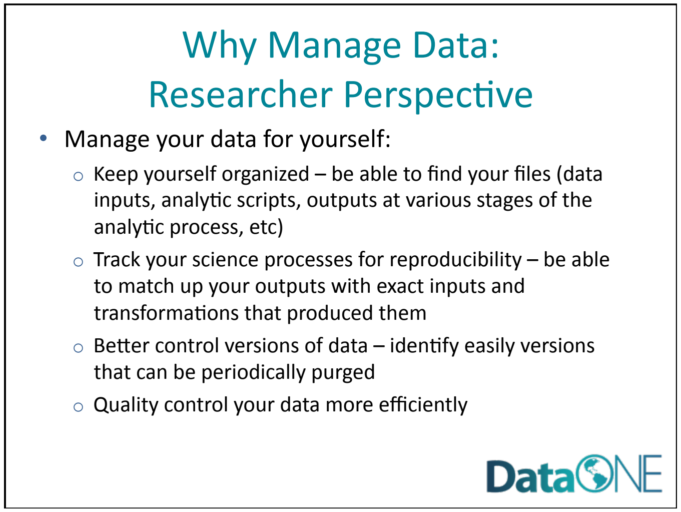
DataONE



## Information Entropy

Sound information management, including metadata development, can arrest the loss of dataset detail.

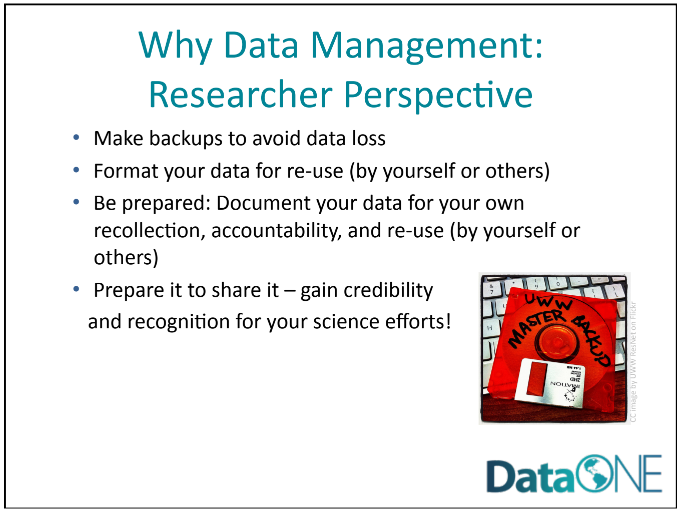
DataONE



## Why Manage Data: Researcher Perspective

- Manage your data for yourself:
  - Keep yourself organized – be able to find your files (data inputs, analytic scripts, outputs at various stages of the analytic process, etc)
  - Track your science processes for reproducibility – be able to match up your outputs with exact inputs and transformations that produced them
  - Better control versions of data – identify easily versions that can be periodically purged
  - Quality control your data more efficiently

DataONE



## Why Data Management: Researcher Perspective

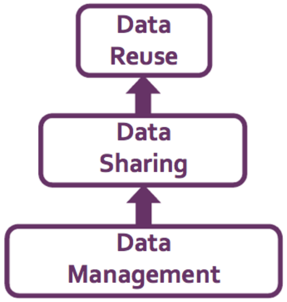
- Make backups to avoid data loss
- Format your data for re-use (by yourself or others)
- Be prepared: Document your data for your own recollection, accountability, and re-use (by yourself or others)
- Prepare it to share it – gain credibility and recognition for your science efforts!

UWW MASTER BACKUP

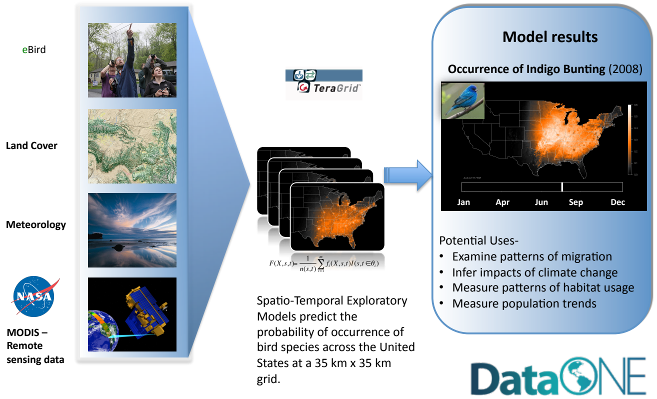
CC image by UWW ResNet on Flickr

DataONE

## Data Management Facilitates Sharing and Re-use…



## Well-Managed Data Can Result in Re-use, Integration and New Science



$$F(X,s,t)=\frac{1}{n(s,t)}\sum_{i=1}^{m} f_i(X,s,t)I(s,t\in\theta_i)$$

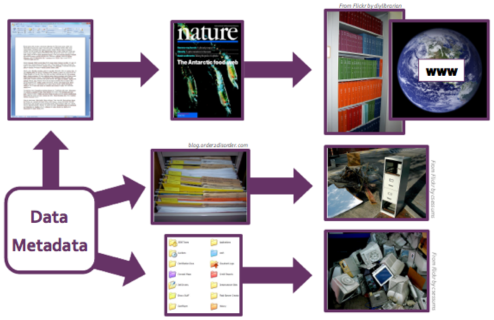
Spatio-Temporal Exploratory Models predict the probability of occurrence of bird species across the United States at a 35 km x 35 km grid.

Potential Uses-

- Examine patterns of migration
- Infer impacts of climate change
- Measure patterns of habitat usage
- Measure population trends

Slide courtesy of DataOne

## Where a majority of data end up now…



From Flickr by diylibrarian

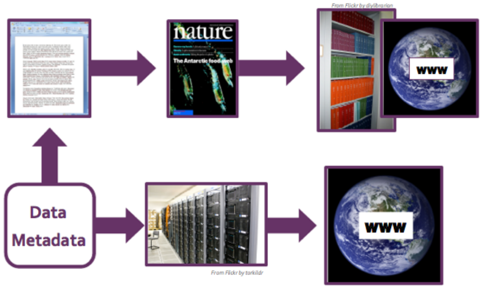
blog.order2disorder.com

From Flickr by csessums

From Flickr by csessums

*Recreated from Klump et al. 2006*

## Imagine if data were more accessible….



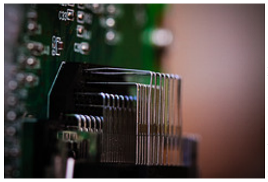
From Flickr by diylibrarian

From Flickr by torkildr

*Recreated from Klump et al. 2006*

## Why Share Data?

Data sharing requires effort, resources, and faith in others. Why do it?



CC image by Jessica Lucia on Flickr

For the benefit of:

- ○ the public
- ○ the research sponsor
- ○ the research community
- ○ the researcher

DataONE

## Value of Data Sharing: To the Scientist

Scientists that share data gain the benefit of:

- ○ Authority
- ○ Citation
- ○ Collaboration

CC image by SLU Madrid Campus on Flickr

DataONE

## Concerns About Data Sharing

Even if the value of data sharing is recognized, concerns remain as to the impacts of increased data exposure.

CC image by CyberHades on Flickr

DataONE

## Concerns About Data Sharing

| Concern | Solution |
|---|---|
| inappropriate use due to misunderstanding of research purpose or parameters | ? |
| security and confidentiality of sensitive data | ? |
| lack of acknowledgement / credit | ? |
| loss of advantage when competing for research dollars | ? |

DataONE

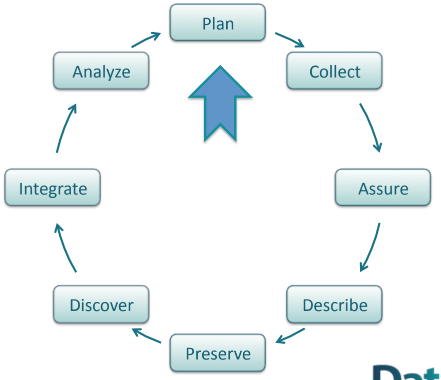
## Concerns About Data Sharing

| Concern | Solution |
|---|---|
| inappropriate use due to misunderstanding of research purpose or parameters | ✓ metadata |
| security and confidentiality of sensitive data | ✓ metadata |
| lack of acknowledgement / credit | ✓ metadata |
| loss of advantage when competing for research dollars | ✓ metadata |

## Data Management Planning & Meta-Data

CC image by Joe Hall on Flickr

## The Data Life Cycle



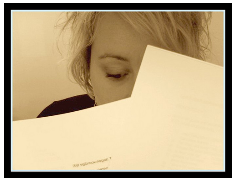
## What is a Data Management Plan?

- Formal document
- Outlines what you will do with your data **during** and **after** you complete your research
- Ensures your data is safe for the **present** and the **future**

*From University of Virginia Library*

## Why Prepare a DMP?

- Save time
  - ○ Less reorganization later
- Increase research efficiency
  - ○ Ensures you and others will be able to understand and use data in future



CC image by Cathdew on Flickr

DataONE

## Components of a General DMP

1. Information about data & data format
2. Metadata content and format
3. Policies for access, sharing and re-use
4. Long-term storage and data management
5. Budget

DataONE

## 1. Information About Data & Data Format

### 1.1 Description of data to be produced

- Experimental
- Observational
- Raw or derived
- Physical collections
- Models and their outputs
- Simulation outputs
- Curriculum materials
- Software
- Images
- Etc…

CC image by Jeffery Beall on Flickr

DataONE

## 1. Information About Data & Data Format

### 1.2 How data will be acquired

- When?
- Where?

### 1.3 How data will be processed

- Software used
- Algorithms
- Workflows

CC image by Ryan Sandridge on Flickr

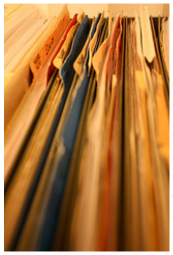
DataONE

## 1. Information About Data & Data Format

1.4 File formats

- Justification
- Naming conventions

1.5 Quality assurance & control during sample collection, analysis, and processing



CC image by Artform Canada on Flickr

DataONE

## 1. Information About Data & Data Format

1.6 Existing data

- If existing data are used, what are their origins?
- Will your data be combined with existing data?
- What is the relationship between your data and existing data?

1.7 How data will be managed in short-term

- Version control
- Backing up
- Security & protection
- Who will be responsible

DataONE

## 2. Metadata Content & Format

**Metadata defined:**

- Documentation and reporting of data
- Contextual details: Critical information about the dataset
- Information important for using the data
- Descriptions of temporal and spatial details, instruments, parameters, units, files, etc.

DataONE

## 2. Metadata Content & Format

2.1 What metadata are needed

- Any details that make data meaningful

2.2 How metadata will be created and/or captured

- Lab notebooks? GPS units?
- Auto-saved on instrument?

2.3 What format will be used for the metadata

- Standards for community
- Justification for format chosen

DataONE

## 3. Policies for Access, Sharing, Reuse

3.4 Intellectual property & copyright issues

- Who owns the copyright?
- Institutional policies
- Funding agency policies
- Embargos for political/commercial reasons

3.5 Intended future uses/users for data

3.6 Citation

- How should data be cited when used?
- Persistent citation?

CC image by buddawiggi on Flickr

## 4. Long-term Storage & Data Management

4.1 What data will be preserved

4.2 Where will it be archived

- Most appropriate archive for data
- Community standards

3.6 Data transformations/formats needed

- Consider archive policies

4.4 Who will be responsible

- Contact person for archive

## 5. Budget

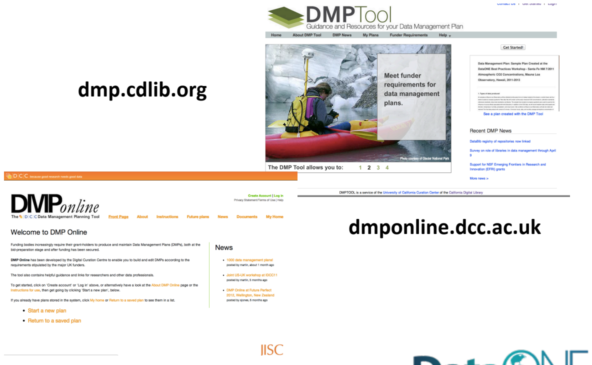
5.1 Anticipated costs

- Time for data preparation & documentation
- Hardware/software for data preparation & documentation
- Personnel
- Archive costs

5.2 How costs will be paid

CC image by Adria Richards on Flickr

## Tools for Creating Data Management Plans



dmp.cdlib.org

DMPTool
Guidance and Resources for your Data Management Plan

Home | About DMP Tool | DMP News | My Plans | Funder Requirements | Help

Get Started!

Meet funder requirements for data management plans.

Photo courtesy of Glacier National Park

The DMP Tool allows you to: 1 2 3 4

Data Management Plan: Sample Plan Created at the DataONE Best Practices Workshop - Santa Fe NM 7/2011
Atmospheric CO2 Concentrations, Mauna Loa Observatory, Hawaii, 2011-2013

1. Types of data produced

See a plan created with the DMP Tool

Recent DMP News

DataBib registry of repositories now linked

Survey on role of libraries in data management through April 9

Support for NSF Emerging Frontiers in Research and Innovation (EFRI) grants

More news >

DMPTOOL is a service of the University of California Curation Center of the California Digital Library

DCC because good research needs good data

DMPonline
The DCC Data Management Planning Tool

Create Account | Log in
Privacy Statement/Terms of Use | Help

Front Page | About | Instructions | Future plans | News | Documents | My Home

Welcome to DMP Online

Funding bodies increasingly require their grant-holders to produce and maintain Data Management Plans (DMPs), both at the bid-preparation stage and after funding has been secured.

**DMP Online** has been developed by the Digital Curation Centre to enable you to build and edit DMPs according to the requirements stipulated by the major UK funders.

The tool also contains helpful guidance and links for researchers and other data professionals.

To get started, click on 'Create account' or 'Log in' above, or alternatively have a look at the About DMP Online page or the Instructions for use, then get going by clicking 'Start a new plan', below.

If you already have plans stored in the system, click My home or Return to a saved plan to see them in a list.

- Start a new plan
- Return to a saved plan

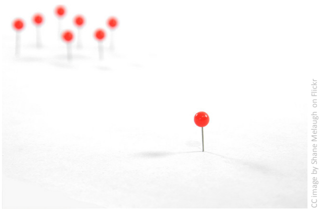
News

- 1000 data management plans!
  posted by martin, about 1 month ago
- Joint US-UK workshop at IDCC11
  posted by martin, 5 months ago
- DMP Online at Future Perfect 2012, Wellington, New Zealand
  posted by sjones, 6 months ago

JISC

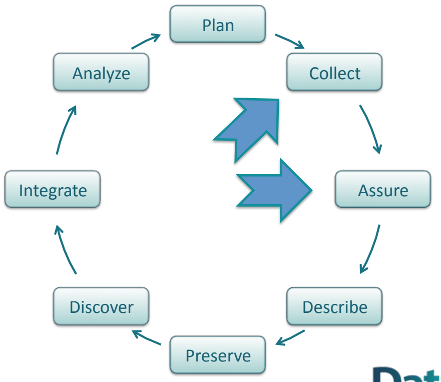
dmponline.dcc.ac.uk

# Data Entry and Quality Control



CC image by Shane Melaugh on Flickr

# The Data Life Cycle

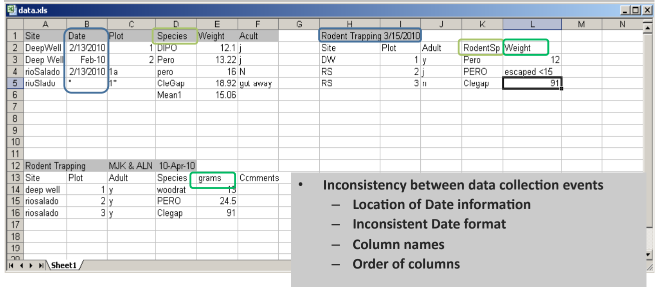


# Goals of Data Entry

- create data sets that are:
  - ○ Valid
  - ○ Organized to support ease of use

CC image by Travis S on Flickr

# Example: Poor Data Entry



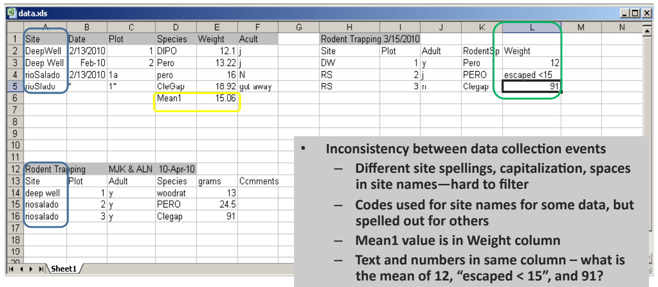
data.xls

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Site | Date | Plot | Species | Weight | Acult | | Rodent Trapping 3/15/2010 | | | | |
| 2 | DeepWell | 2/13/2010 | 1 | DIPO | 12.1 | j | | Site | Plot | Adult | RodentSp | Weight |
| 3 | Deep Well | Feb-10 | 2 | Pero | 13.22 | j | | DW | 1 | y | Pero | 12 |
| 4 | rioSalado | 2/13/2010 | 1a | pero | 16 | N | | RS | 2 | j | PERO | escaped <15 |
| 5 | riuSladu | " | 1* | CleGap | 18.92 | gut away | | RS | 3 | n | Clegap | 91 |
| 6 | | | | Mean1 | 15.06 | | | | | | | |
| 7 | | | | | | | | | | | | |
| 8 | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | |
| 10 | | | | | | | | | | | | |
| 11 | | | | | | | | | | | | |
| 12 | Rodent Trapping | | MJK & ALN | 10-Apr-10 | | | | | | | | |
| 13 | Site | Plot | Adult | Species | grams | Ccmments | | | | | | |
| 14 | deep well | 1 | y | woodrat | 13 | | | | | | | |
| 15 | riosalado | 2 | y | PERO | 24.5 | | | | | | | |
| 16 | riosalado | 3 | y | Clegap | 91 | | | | | | | |

Sheet1

- Inconsistency between data collection events
  - Location of Date information
  - Inconsistent Date format
  - Column names
  - Order of columns

## Example: Poor Data Entry



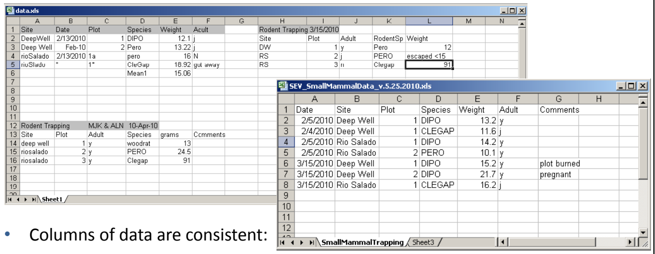
data.xls

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Site | Date | Plot | Species | Weight | Acult | | Rodent Trapping 3/15/2010 | | | | | | |
| 2 | DeepWell | 2/13/2010 | 1 | DIPO | 12.1 | j | | Site | Plot | Adult | RodentSp | Weight | | |
| 3 | Deep Well | Feb-10 | 2 | Pero | 13.22 | j | | DW | 1 | y | Pero | 12 | | |
| 4 | rioSalado | 2/13/2010 | 1a | pero | 16 | N | | RS | 2 | j | PERO | escaped <15 | | |
| 5 | riuSladu | " | 1* | CleGap | 18.92 | gul away | | RS | 3 | n | Clegap | 91 | | |
| 6 | | | | Mean1 | 15.06 | | | | | | | | | |
| 7 | | | | | | | | | | | | | | |
| 8 | | | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | | | |
| 10 | | | | | | | | | | | | | | |
| 11 | | | | | | | | | | | | | | |
| 12 | Rodent Trapping | | MJK & ALN | 10-Apr-10 | | | | | | | | | | |
| 13 | Site | Plot | Adult | Species | grams | Ccmments | | | | | | | | |
| 14 | deep well | 1 | y | woodrat | 13 | | | | | | | | | |
| 15 | riosalado | 2 | y | PERO | 24.5 | | | | | | | | | |
| 16 | riosalado | 3 | y | Clegap | 91 | | | | | | | | | |
| 17 | | | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | | | |
| 19 | | | | | | | | | | | | | | |

Sheet1

- Inconsistency between data collection events
  - Different site spellings, capitalization, spaces in site names—hard to filter
  - Codes used for site names for some data, but spelled out for others
  - Mean1 value is in Weight column
  - Text and numbers in same column – what is the mean of 12, "escaped < 15", and 91?

## Best Practices



data.xls

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Site | Date | Plot | Species | Weight | Acult | | Rodent Trapping 3/15/2010 | | | | | | |
| 2 | DeepWell | 2/13/2010 | 1 | DIPO | 12.1 | j | | Site | Plot | Adult | RodentSp | Weight | | |
| 3 | Deep Well | Feb-10 | 2 | Pero | 13.22 | j | | DW | 1 | y | Pero | 12 | | |
| 4 | rioSalado | 2/13/2010 | 1a | pero | 16 | N | | RS | 2 | j | PERO | escaped <15 | | |
| 5 | riuSladu | " | 1* | CleGap | 18.92 | gul away | | RS | 3 | n | Clegap | 91 | | |
| 6 | | | | Mean1 | 15.06 | | | | | | | | | |
| 12 | Rodent Trapping | | MJK & ALN | 10-Apr-10 | | | | | | | | | | |
| 13 | Site | Plot | Adult | Species | grams | Ccmments | | | | | | | | |
| 14 | deep well | 1 | y | woodrat | 13 | | | | | | | | | |
| 15 | riosalado | 2 | y | PERO | 24.5 | | | | | | | | | |
| 16 | riosalado | 3 | y | Clegap | 91 | | | | | | | | | |

Sheet1

SEV_SmallMammalData_v.5.25.2010.xls

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Date | Site | Plot | Species | Weight | Adult | Comments | |
| 2 | 2/5/2010 | Deep Well | 1 | DIPO | 13.2 | y | | |
| 3 | 2/4/2010 | Deep Well | 1 | CLEGAP | 11.6 | j | | |
| 4 | 2/5/2010 | Rio Salado | 1 | DIPO | 14.2 | y | | |
| 5 | 2/5/2010 | Rio Salado | 2 | PERO | 10.1 | y | | |
| 6 | 3/15/2010 | Deep Well | 1 | DIPO | 15.2 | y | plot burned | |
| 7 | 3/15/2010 | Deep Well | 2 | DIPO | 21.7 | y | pregnant | |
| 8 | 3/15/2010 | Rio Salado | 1 | CLEGAP | 16.2 | j | | |
| 9 | | | | | | | | |
| 10 | | | | | | | | |
| 11 | | | | | | | | |
| 12 | | | | | | | | |

SmallMammalTrapping / Sheet3

- Columns of data are consistent: only numbers, dates, or text
- Consistent Names, Codes, Formats (date) used in each column
- Data are all in one table, which is much easier for a statistical program to work with than multiple small tables which each require human intervention

## Best Practices

- Create descriptive column names without spaces or special characters
  - Soil T30 → Soil_Temp_30cm
  - Species-Code → Species_Code (avoid using -,+,*,^ in column names. Some software may interpret these symbols as an operator)
- Use a descriptive file name. For instance, a file named SEV_SmallMammalData_v.5.25.2010.csv indicates the project the data is associated with (SEV), the theme of the data (SmallMammalData) and also when this version of the data was created (v.5.25.2010). This name is much more helpful than a file named mydata.xls.

## Best Practices

- Missing data
  - Preferably leave field empty (NULL = no value)
  - In numeric fields, use a distinct value such as 9999 to indicate a missing value – but only if this is in your meta-data!
  - In text fields, use NA ("Not Applicable" or "Not Available")
  - Use Data flags in a separate column to qualify missing value

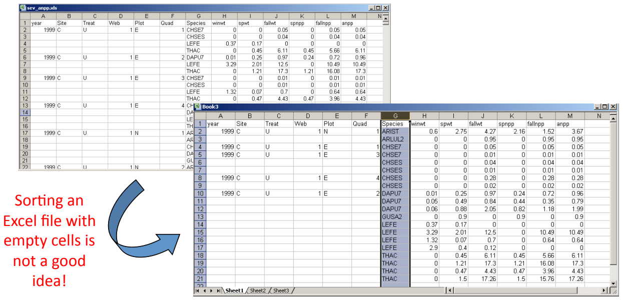
| Date | Time | NO3_N_Conc | NO3_N_Conc_Flag |
|---|---|---|---|
| 20081011 | 1300 | 0.013 | |
| 20081011 | 1330 | 0.016 | |
| 20081011 | 1400 | | M1 |
| 20081011 | 1430 | 0.018 | |
| 20081011 | 1500 | 0.001 | E1 |

M1 = missing; no sample collected

E1 = estimated from grab sample

## Best Practices

- Enter complete lines of data



sev_anpp.xls

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | year | Site | Treat | Web | Plot | Quad | Species | winwt | spwt | fallwt | spnpp | fallnpp | anpp | |
| 2 | 1999 | C | U | 1 | E | 1 | CHSE7 | 0 | 0 | 0.05 | 0 | 0.05 | 0.05 | |
| 3 | | | | | | | CHSES | 0 | 0 | 0.04 | 0 | 0.04 | 0.04 | |
| 4 | | | | | | | LEFE | 0.37 | 0.17 | 0 | 0 | 0 | 0 | |
| 5 | | | | | | | THAC | 0 | 0.45 | 6.11 | 0.45 | 5.66 | 6.11 | |
| 6 | 1999 | C | U | 1 | E | 2 | DAPU7 | 0.01 | 0.25 | 0.97 | 0.24 | 0.72 | 0.96 | |
| 7 | | | | | | | LEFE | 3.29 | 2.01 | 12.5 | 0 | 10.49 | 10.49 | |
| 8 | | | | | | | THAC | 0 | 1.21 | 17.3 | 1.21 | 16.08 | 17.3 | |
| 9 | 1999 | C | U | 1 | E | 3 | CHSE7 | 0 | 0 | 0.01 | 0 | 0.01 | 0.01 | |
| 10 | | | | | | | CHSES | 0 | 0 | 0.01 | 0 | 0.01 | 0.01 | |
| 11 | | | | | | | LEFE | 1.32 | 0.07 | 0.7 | 0 | 0.64 | 0.64 | |
| 12 | | | | | | | THAC | 0 | 0.47 | 4.43 | 0.47 | 3.96 | 4.43 | |
| 13 | 1999 | C | U | 1 | E | 4 | CH | | | | | | | |
| 14 | | | | | | | DA | | | | | | | |
| 15 | | | | | | | LE | | | | | | | |
| 16 | | | | | | | TH | | | | | | | |
| 17 | 1999 | C | U | 1 | N | 1 | AR | | | | | | | |
| 18 | | | | | | | AR | | | | | | | |
| 19 | | | | | | | CH | | | | | | | |
| 20 | | | | | | | DA | | | | | | | |
| 21 | | | | | | | GU | | | | | | | |
| 22 | 1999 | C | U | 1 | N | 2 | AR | | | | | | | |

Book3

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | year | Site | Treat | Web | Plot | Quad | Species | winwt | spwt | fallwt | spnpp | fallnpp | anpp | |
| 2 | 1999 | C | U | 1 | N | 1 | ARIST | 0.6 | 2.75 | 4.27 | 2.16 | 1.52 | 3.67 | |
| 3 | | | | | | | ARLUL2 | 0 | 0 | 0.95 | 0 | 0.95 | 0.95 | |
| 4 | 1999 | C | U | 1 | E | 1 | CHSE7 | 0 | 0 | 0.05 | 0 | 0.05 | 0.05 | |
| 5 | 1999 | C | U | 1 | E | 3 | CHSE7 | 0 | 0 | 0.01 | 0 | 0.01 | 0.01 | |
| 6 | | | | | | | CHSES | 0 | 0 | 0.04 | 0 | 0.04 | 0.04 | |
| 7 | | | | | | | CHSES | 0 | 0 | 0.01 | 0 | 0.01 | 0.01 | |
| 8 | 1999 | C | U | 1 | E | 4 | CHSES | 0 | 0 | 0.28 | 0 | 0.28 | 0.28 | |
| 9 | | | | | | | CHSES | 0 | 0 | 0.02 | 0 | 0.02 | 0.02 | |
| 10 | 1999 | C | U | 1 | E | 2 | DAPU7 | 0.01 | 0.25 | 0.97 | 0.24 | 0.72 | 0.96 | |
| 11 | | | | | | | DAPU7 | 0.05 | 0.49 | 0.84 | 0.44 | 0.35 | 0.79 | |
| 12 | | | | | | | DAPU7 | 0.06 | 0.88 | 2.05 | 0.82 | 1.18 | 1.99 | |
| 13 | | | | | | | GUSA2 | 0 | 0.9 | 0 | 0.9 | 0 | 0.9 | |
| 14 | | | | | | | LEFE | 0.37 | 0.17 | 0 | 0 | 0 | 0 | |
| 15 | | | | | | | LEFE | 3.29 | 2.01 | 12.5 | 0 | 10.49 | 10.49 | |
| 16 | | | | | | | LEFE | 1.32 | 0.07 | 0.7 | 0 | 0.64 | 0.64 | |
| 17 | | | | | | | LEFE | 2.9 | 0.4 | 0.12 | 0 | 0 | 0 | |
| 18 | | | | | | | THAC | 0 | 0.45 | 6.11 | 0.45 | 5.66 | 6.11 | |
| 19 | | | | | | | THAC | 0 | 1.21 | 17.3 | 1.21 | 16.08 | 17.3 | |
| 20 | | | | | | | THAC | 0 | 0.47 | 4.43 | 0.47 | 3.96 | 4.43 | |
| 21 | | | | | | | THAC | 0 | 1.5 | 17.26 | 1.5 | 15.76 | 17.26 | |

Sorting an Excel file with empty cells is not a good idea!

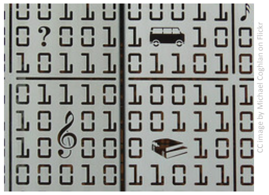
DataONE

## Best Practices

- For the long term, store data in a consistent format that can be read well in to the future and that can be used by any application now or in the future
- Appropriate file types include:
  - Non-proprietary: Open, documented standard
  - Common usage by research community: Standard representation (ASCII, Unicode)
  - Unencrypted
  - Uncompressed
- ASCII formatted files will be readable into the future
  - Use ASCII (comma-separated) for tabular data

DataONE

## Quality Control



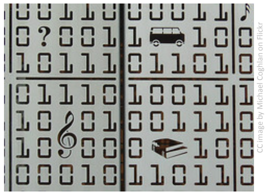
CC image by Michael Coghlan on Flickr

DataONE

## Definitions

**Data Contamination**

- Process or phenomenon, other than the one of interest, that affects the variable value
- Erroneous values

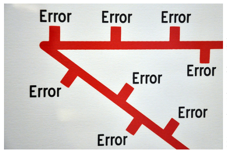


CC image by Michael Coghlan on Flickr

DataONE

## Definitions: Types of Errors

- Errors of **Commission**
  - Incorrect or inaccurate data entered
  - Examples: malfunctioning instrument, mistyped data
- Errors of **Omission**
  - Data or metadata not recorded
  - Examples: inadequate documentation, human error, anomalies in the field



CC image by Nick J Webb on Flickr

DataONE

## Defining QA/QC

- Strategies for preventing errors from entering a dataset
- Activities to ensure quality of data before collection
- Activities that involve monitoring and maintaining the quality of data during the study

DataONE

## QA/QC Before Collection

- Define & enforce standards
  - Formats
  - Codes
  - Measurement units
  - Metadata
- Assign responsibility for data quality
  - Be sure assigned person is educated in QA/QC

DataONE

## QA/QC During Data Entry

- Double entry
  - Data keyed in by two independent people
  - Check for agreement with computer verification
- Record a reading of the data and transcribe from the recording
- Use text-to-speech program to read data back

CC image by weskriesel on Flickr

DataONE

## QA/QC During Data Entry

- Design data storage well
  - Minimize number of times items that must be entered repeatedly
  - Use consistent terminology
  - Atomize data: one cell per piece of information
- Document changes to data
  - Avoids duplicate error checking
  - Allows undo if necessary

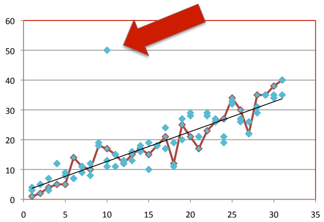
## QA/QC After Data Entry

- Make sure data line up in proper columns
- No missing, impossible, or anomalous values
- Perform statistical summaries

CC image by chesapeakeclimate on Flickr

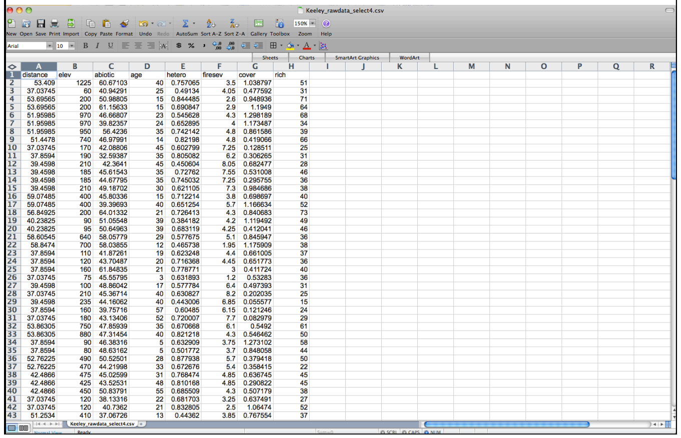
## QA/QC After Data Entry

- Look for outliers
  - Outliers are extreme values for a variable given the statistical model being used
  - The goal is not to eliminate outliers but to identify potential data contamination



## R and Data Screening

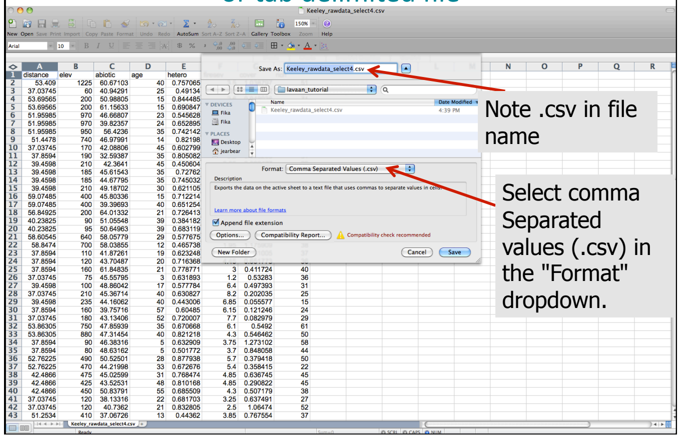
## R and Data: Use Data in a Long Table Format



Keeley_rawdata_select4.csv

|  | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | distance | elev | abiotic | age | hetero | firesev | cover | rich |
| 2 | 53.409 | 1225 | 60.67103 | 40 | 0.757065 | 3.5 | 1.038797 | 51 |
| 3 | 37.03745 | 60 | 40.94291 | 25 | 0.49134 | 4.05 | 0.477592 | 31 |
| 4 | 53.69565 | 200 | 50.98805 | 15 | 0.844485 | 2.6 | 0.948936 | 71 |
| 5 | 53.69565 | 200 | 61.15633 | 15 | 0.690847 | 2.9 | 1.1949 | 64 |
| 6 | 51.95985 | 970 | 46.66807 | 23 | 0.545628 | 4.3 | 1.298189 | 68 |
| 7 | 51.95985 | 970 | 39.82357 | 24 | 0.652895 | 4 | 1.173487 | 34 |
| 8 | 51.95985 | 950 | 56.4236 | 35 | 0.742142 | 4.8 | 0.861586 | 39 |
| 9 | 51.4478 | 740 | 46.97991 | 14 | 0.82198 | 4.8 | 0.419066 | 66 |
| 10 | 37.03745 | 170 | 42.08806 | 45 | 0.602799 | 7.25 | 0.128511 | 25 |
| 11 | 37.8594 | 190 | 32.59387 | 35 | 0.805082 | 6.2 | 0.306265 | 31 |
| 12 | 39.4598 | 210 | 42.3641 | 45 | 0.450604 | 8.05 | 0.682477 | 28 |
| 13 | 39.4598 | 185 | 45.61543 | 35 | 0.72762 | 7.55 | 0.531008 | 46 |
| 14 | 39.4598 | 185 | 44.67795 | 35 | 0.745032 | 7.25 | 0.295755 | 36 |
| 15 | 39.4598 | 210 | 49.18702 | 30 | 0.621105 | 7.3 | 0.984686 | 38 |
| 16 | 59.07485 | 400 | 45.80336 | 15 | 0.712214 | 3.8 | 0.698697 | 40 |
| 17 | 59.07485 | 400 | 39.39693 | 40 | 0.651254 | 5.7 | 1.166634 | 52 |
| 18 | 56.84925 | 200 | 64.01332 | 21 | 0.726413 | 4.3 | 0.840683 | 73 |
| 19 | 40.23825 | 90 | 51.05548 | 39 | 0.384182 | 4.2 | 1.119492 | 49 |
| 20 | 40.23825 | 95 | 50.64963 | 39 | 0.683119 | 4.25 | 0.412041 | 46 |
| 21 | 58.60545 | 640 | 58.05779 | 29 | 0.577675 | 5.1 | 0.845947 | 36 |
| 22 | 58.8474 | 700 | 58.03855 | 12 | 0.465738 | 1.95 | 1.175909 | 38 |
| 23 | 37.8594 | 110 | 41.87261 | 19 | 0.623248 | 4.4 | 0.661005 | 37 |
| 24 | 37.8594 | 120 | 43.70487 | 20 | 0.716368 | 4.45 | 0.651773 | 36 |
| 25 | 37.8594 | 160 | 61.84835 | 21 | 0.778771 | 3 | 0.411724 | 40 |
| 26 | 37.03745 | 75 | 45.55795 | 3 | 0.631893 | 1.2 | 0.53283 | 36 |
| 27 | 39.4598 | 100 | 48.86042 | 17 | 0.577784 | 6.4 | 0.497393 | 31 |
| 28 | 37.03745 | 210 | 45.36714 | 40 | 0.630827 | 8.2 | 0.202035 | 25 |
| 29 | 39.4598 | 235 | 44.16062 | 40 | 0.443006 | 6.85 | 0.055577 | 15 |
| 30 | 37.8594 | 160 | 39.75716 | 57 | 0.60485 | 6.15 | 0.121246 | 24 |
| 31 | 37.03745 | 180 | 43.13406 | 52 | 0.720007 | 7.7 | 0.082979 | 29 |
| 32 | 53.86305 | 750 | 47.85939 | 35 | 0.670668 | 6.1 | 0.5492 | 61 |
| 33 | 53.86305 | 880 | 47.31454 | 40 | 0.821218 | 4.3 | 0.546462 | 50 |
| 34 | 37.8594 | 90 | 46.38316 | 5 | 0.632909 | 3.75 | 1.273102 | 58 |
| 35 | 37.8594 | 80 | 48.63162 | 5 | 0.501772 | 3.7 | 0.848058 | 44 |
| 36 | 52.76225 | 490 | 50.52501 | 28 | 0.877938 | 5.7 | 0.379418 | 50 |
| 37 | 52.76225 | 470 | 44.21998 | 33 | 0.672676 | 5.4 | 0.358415 | 22 |
| 38 | 42.4866 | 475 | 45.02599 | 31 | 0.768474 | 4.85 | 0.636745 | 45 |
| 39 | 42.4866 | 425 | 43.52531 | 48 | 0.810168 | 4.85 | 0.290822 | 45 |
| 40 | 42.4866 | 450 | 50.83791 | 55 | 0.685509 | 4.3 | 0.507179 | 38 |
| 41 | 37.03745 | 120 | 38.13316 | 22 | 0.681703 | 3.25 | 0.637491 | 27 |
| 42 | 37.03745 | 120 | 40.7362 | 21 | 0.832805 | 2.5 | 1.06474 | 52 |
| 43 | 51.2534 | 410 | 37.06726 | 13 | 0.44362 | 3.85 | 0.767554 | 37 |

## R doesn't Play Nice with Excel – save data in a comma or tab delimited file



Save As: Keeley_rawdata_select4.csv

lavaan_tutorial

| Name | Date Modified |
|---|---|
| Keeley_rawdata_select4.csv | 4:39 PM |

Format: Comma Separated Values (.csv)

Description

Exports the data on the active sheet to a text file that uses commas to separate values in cells.

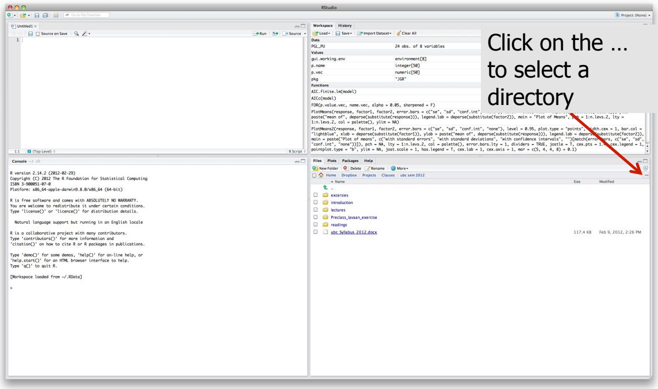
Learn more about file formats

☑ Append file extension

Note .csv in file name

Select comma Separated values (.csv) in the "Format" dropdown.

## Step 1) Set your Working Directory



| Data |  |
|---|---|
| PGL_PU | 24 obs. of 8 variables |
| **Values** |  |
| gui.working.env | environment[8] |
| p.name | integer[50] |
| p.vec | numeric[50] |
| pkg | "JGR" |

Functions

```
AIC.finite.lm(model)
AICc(model)
FDR(p.value.vec, name.vec, alpha = 0.05, sharpened = F)
PlotMeans(response, factor1, factor2, error.bars = c("se", "sd", "conf.int",
paste("mean of", deparse(substitute(response))), legend.lab = deparse(substitute(factor2)), main = "Plot of Means", pch = 1:n.levs.2, lty =
1:n.levs.2, col = palette(), ylim = NA)
PlotMeans2(response, factor1, factor2, error.bars = c("se", "sd", "conf.int", "none"), level = 0.95, plot.type = "points", width.cex = 1, bar.col =
"lightblue", xlab = deparse(substitute(factor1)), ylab = paste("mean of", deparse(substitute(response))), legend.lab = deparse(substitute(factor2)),
main = paste("Plot of means", c("with standard errors", "with standard deviations", "with confidence intervals", "")[match(error.bars, c("se", "sd",
"conf.int", "none"))]), pch = NA, lty = 1:n.levs.2, col = palette(), error.bars.lty = 1, dividers = TRUE, jostle = T, cex.pts = 1.4, cex.legend = 1,
pointplot.type = "b", ylim = NA, jost.scale = 1, has.legend = T, cex.lab = 1, cex.axis = 1, mar = c(5, 4, 4, 8) + 0.1)
```

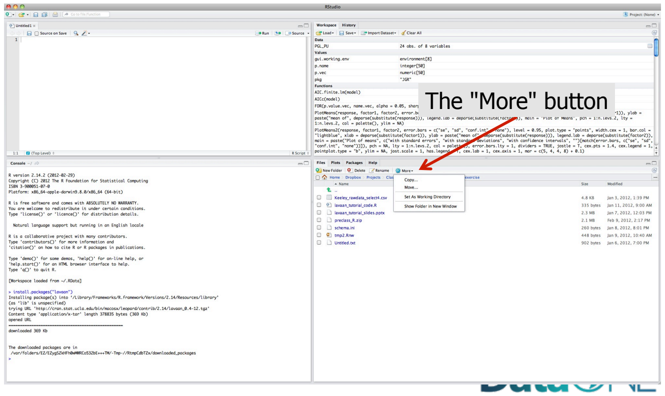
```
R version 2.14.2 (2012-02-29)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

>
```

Home › Dropbox › Projects › Classes › ubc sem 2012

| Name | Size | Modified |
|---|---|---|
| .. |  |  |
| excersies |  |  |
| introduction |  |  |
| lectures |  |  |
| Preclass_lavaan_exercise |  |  |
| readings |  |  |
| ubc_Syllabus_2012.docx | 117.4 KB | Feb 9, 2012, 2:26 PM |

Click on the … to select a directory

## Step 1) Set your Working Directory



```
R version 2.14.2 (2012-02-29)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> install.packages("lavaan")
Installing package(s) into '/Library/Frameworks/R.framework/Versions/2.14/Resources/library'
(as 'lib' is unspecified)
trying URL 'http://cran.stat.ucla.edu/bin/macosx/leopard/contrib/2.14/lavaan_0.4-12.tgz'
Content type 'application/x-tar' length 378835 bytes (369 Kb)
opened URL
==================================================
downloaded 369 Kb


The downloaded packages are in
 /var/folders/EZ/EZygSZkHFh0wMMRCo532bE+++TM/-Tmp-//RtmpCdbTZx/downloaded_packages
>
```

Home › Dropbox › Projects › Clas… exercise

| Name | Size | Modified |
|---|---|---|
| .. |  |  |
| Keeley_rawdata_select4.csv | 4.8 KB | Jan 3, 2012, 1:39 PM |
| lavaan_tutorial_code.R | 335 bytes | Jan 11, 2012, 9:00 AM |
| lavaan_tutorial_slides.pptx | 2.3 MB | Jan 7, 2012, 12:03 PM |
| preclass_R.zip | 2.1 MB | Feb 9, 2012, 2:17 PM |
| schema.ini | 260 bytes | Jan 8, 2012, 8:01 PM |
| tmp2.Rnw | 448 bytes | Jan 9, 2012, 10:40 AM |
| Untitled.txt | 902 bytes | Jan 6, 2012, 7:00 PM |

The "More" button

# Loading Data

```
westNile <- read.csv("./data/SwaddleWestNile2002NCEAS-BAD.csv")
```

Note:

- File path (./ is this directory, ../ is back one directory)
- Quotes
- Our data is now an object in R

# Look at Your Data

```
head(westNile)

##   State                     Infected.County WNV.incidence
## 1    AL      Autauga, AL                            2.290
## 2    AL     Calhoun , AL                            0.891
## 3    AL Chambers, AL                                2.734
## 4    AL Dallas , AL                                 2.157
## 5    AL Marengo , AL                                8.874
## 6    AL Marion, AL                                  3.204
##   Species.Richness Corvid.Abundance
## 1               66                8
## 2               67               64
## 3               41               69
## 4               60               66
## 5               69               64
## 6               NA    NOT AVAILABLE
```

# Look at Columns 3 through 4

```
head(westNile[, 3:4])

##   WNV.incidence Species.Richness
## 1         2.290               66
## 2         0.891               67
## 3         2.734               41
## 4         2.157               60
## 5         8.874               69
## 6         3.204               NA
```

- Data Frame is treates as a Matrix.
- $[rows, columns]$

# Look at Your Individual Columns

```
names(westNile)

## [1] "State"            "Infected.County"  "WNV.incidence"
## [4] "Species.Richness" "Corvid.Abundance"
```

(Note that spaces are now .s)

```
westNile$Species.Richness

##   [1] 66 67 41 60 69 NA 56 65 54 52 81 51 47 59 49 51 72 53
##  [19] 54 49 61 81 62 70 71 57 87 64 50 62 71 70 59 63 58 51
##  [37] 46 66 53 59 58 56 58 43 65 51 51 63 54 60 53 39 62 67
##  [55] 68 82 70 76 58 60 72 59 72 62 82 63 68 39 67 66 63 47
##  [73] 59 61 65 79 54 56 30 48 56 68 58 42 51 64 73 55 61 65
##  [91] 61 74 65 61 51 93 42 63 68 58 68 61 56 60 81 66 53 49
## [109] 68 72 76 57 76 55 76 56 73 59 73 57 90 50 73 64 78 75
## [127] 61 80 59 69
```

# Missing Data is NA

```
westNile$Species.Richness

##   [1] 66 67 41 60 69 NA 56 65 54 52 81 51 47 59 49 51 72 53
##  [19] 54 49 61 81 62 70 71 57 87 64 50 62 71 70 59 63 58 51
##  [37] 46 66 53 59 58 56 58 43 65 51 51 63 54 60 53 39 62 67
##  [55] 68 82 70 76 58 60 72 59 72 62 82 63 68 39 67 66 63 47
##  [73] 59 61 65 79 54 56 30 48 56 68 58 42 51 64 73 55 61 65
##  [91] 61 74 65 61 51 93 42 63 68 58 68 61 56 60 81 66 53 49
## [109] 68 72 76 57 76 55 76 56 73 59 73 57 90 50 73 64 78 75
## [127] 61 80 59 69
```

Note the NA. This is missing data.

```
westNile$Species.Richness[6]

## [1] NA
```

# Let's look at another

```
westNile$Corvid.Abundance

##    [1] 8             64            69
##    [4] 66            64            NOT AVAILABLE
##    [7] 59            129           54
##   [10] 100           62            82
##   [13] 102           35            31
##   [16] 13            51            60
##   [19] 10            87            53
##   [22] 9999          34            86
##   [25] 75            102           216
##   [28] 71            43            57
##   [31] 98            84            44
##   [34] 109           165           44
##   [37] 68            48            34
##   [40] 63            9999          52
##   [43] 24            39            41
##   [46] 32            47            23
```

# Cleaner Data

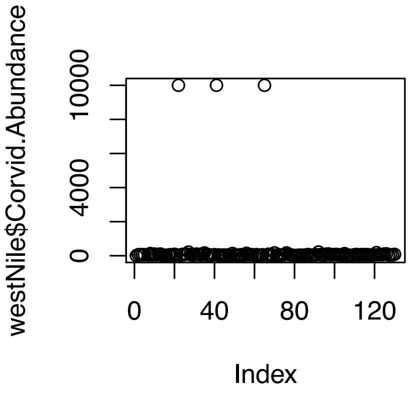
```
westNile <- read.csv("./data/SwaddleWestNile2002NCEAS-BAD.csv",
    na.strings = "NOT AVAILABLE")
```

# And, Fixed!

```
westNile$Corvid.Abundance

##    [1]    8.00   64.00   69.00   66.00   64.00      NA
##    [7]   59.00  129.00   54.00  100.00   62.00   82.00
##   [13]  102.00   35.00   31.00   13.00   51.00   60.00
##   [19]   10.00   87.00   53.00 9999.00   34.00   86.00
##   [25]   75.00  102.00  216.00   71.00   43.00   57.00
##   [31]   98.00   84.00   44.00  109.00  165.00   44.00
##   [37]   68.00   48.00   34.00   63.00 9999.00   52.00
##   [43]   24.00   39.00   41.00   32.00   47.00   23.00
##   [49]  135.00   49.00   32.00   27.00   63.00   15.00
##   [55]   45.00  144.00   61.00   71.00   57.00   29.00
##   [61]   66.00   36.00   46.00   57.00 9999.00   54.00
##   [67]   91.00   19.00   56.00  168.00   14.00   71.00
##   [73]   43.00   48.00   70.00  170.00   75.00   63.00
##   [79]    6.00   18.00   21.00   29.00   34.00   18.00
##   [85]   39.00   57.00   71.00   26.00   31.00   47.00
##   [91]   63.00  220.00   70.00   42.00   36.00  101.00
```

## What about fixing many bad values?

```
plot(westNile$Corvid.Abundance)
```



## What about fixing many bad values?

```
which(westNile$Corvid.Abundance == 9999)

## [1] 22 41 65
```

== is makes a COMPARISON and returns a logical value
Can also use <, >, and more.

```
westNile$Corvid.Abundance == 9999

##   [1] FALSE FALSE FALSE FALSE FALSE    NA FALSE FALSE FALSE
##  [10] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [19] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##  [28] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##  [46] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [55] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [64] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [82] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

## Replace the 9999s

```
westNile$Corvid.Abundance[which(westNile$Corvid.Abundance ==
    9999)] <- NA
```

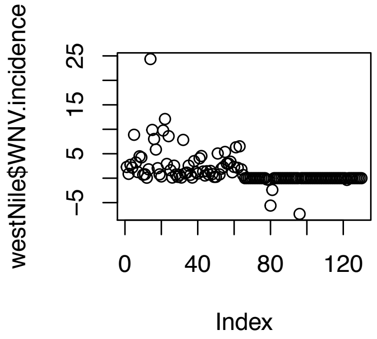
The which approach is often good, as once you spot a single problem observation, there may be others like it.

## Exercise

1. Is everything OK with West Nile Virus Incidence?
2. Let's say a database overwrote some 0 values - fix these values!

# The Fix

```
plot(westNile$WNV.incidence)
```



```
westNile$WNV.incidence[which(westNile$WNV.incidence < 0)] <- 0
```

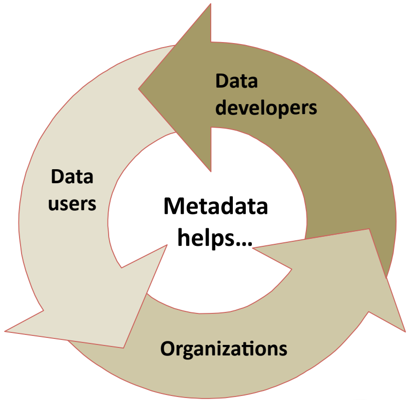
# Metadata

## What is Metadata?

**Metadata is: Data ‘reporting’**

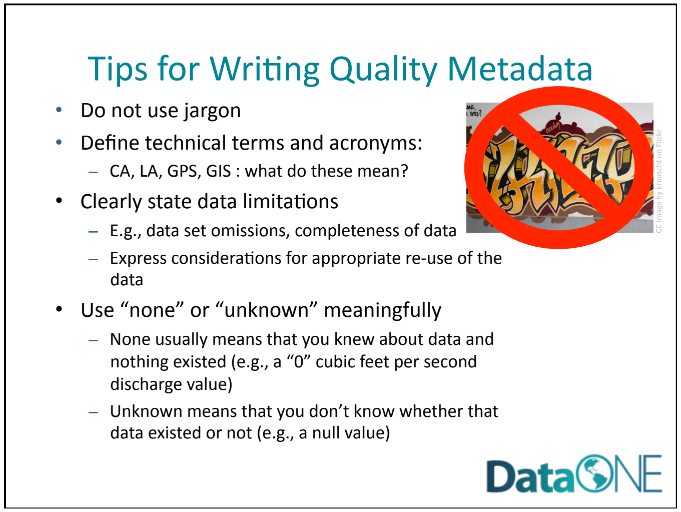
- **WHO** created the data?
- **WHAT** is the content of the data?
- **WHEN** were the data created?
- **WHERE** is it geographically?
- **HOW** were the data developed?
- **WHY** were the data developed?

## The Value of Metadata



## Multiple Metadata Standards Exist: Examples

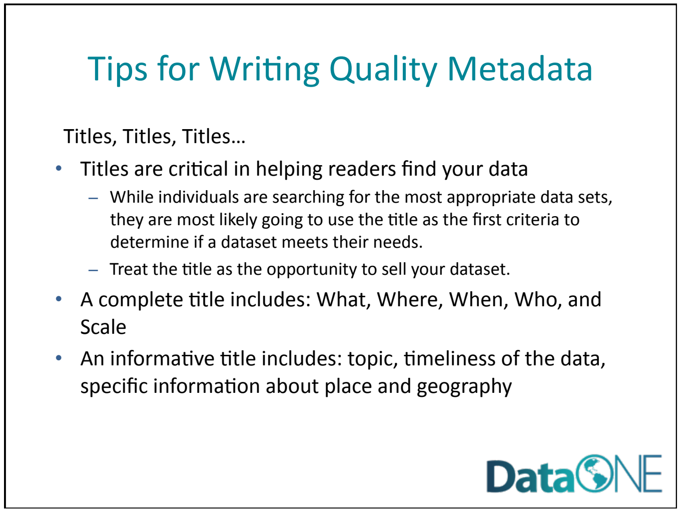
- **Ecological Metadata Language (EML)**
  - Focus on ecological data
  - http://knb.ecoinformatics.org/eml_metadata_guide.html
- **Darwin Core**
  - Emphasis on museum specimens
  - http://rs.tdwg.org/dwc/index.htm
- **Geography Markup Language (GML)**
  - Emphasis on geographic features (roads, highways, bridges)
  - http://www.opengeospatial.org/standards/gml



## Tips for Writing Quality Metadata

- Do not use jargon
- Define technical terms and acronyms:
  - CA, LA, GPS, GIS : what do these mean?
- Clearly state data limitations
  - E.g., data set omissions, completeness of data
  - Express considerations for appropriate re-use of the data
- Use “none” or “unknown” meaningfully
  - None usually means that you knew about data and nothing existed (e.g., a “0” cubic feet per second discharge value)
  - Unknown means that you don’t know whether that data existed or not (e.g., a null value)

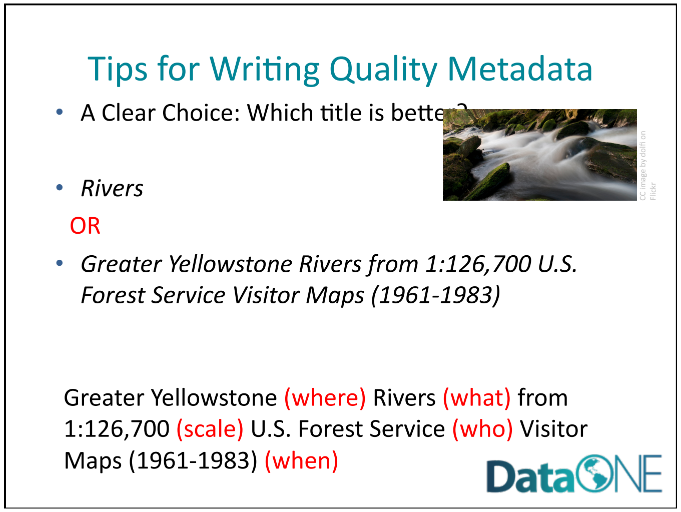
CC image by kruuscht on Flickr



## Tips for Writing Quality Metadata

Titles, Titles, Titles…

- Titles are critical in helping readers find your data
  - While individuals are searching for the most appropriate data sets, they are most likely going to use the title as the first criteria to determine if a dataset meets their needs.
  - Treat the title as the opportunity to sell your dataset.
- A complete title includes: What, Where, When, Who, and Scale
- An informative title includes: topic, timeliness of the data, specific information about place and geography



## Tips for Writing Quality Metadata

- A Clear Choice: Which title is better?

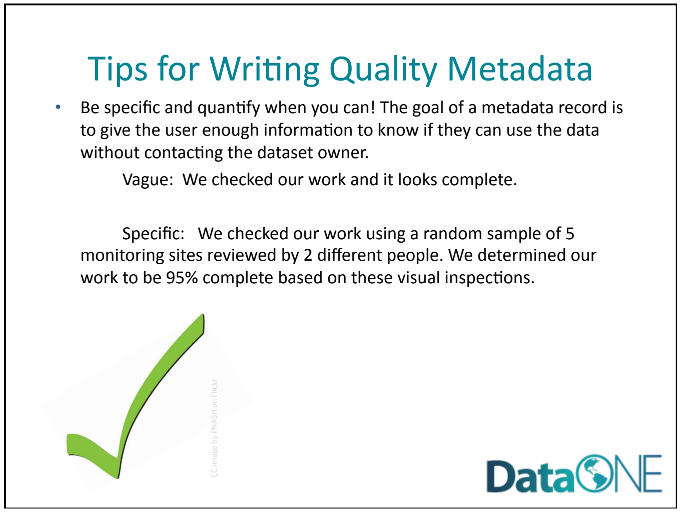
- *Rivers*

OR

- *Greater Yellowstone Rivers from 1:126,700 U.S. Forest Service Visitor Maps (1961-1983)*

CC image by dolfi on Flickr

Greater Yellowstone (where) Rivers (what) from 1:126,700 (scale) U.S. Forest Service (who) Visitor Maps (1961-1983) (when)



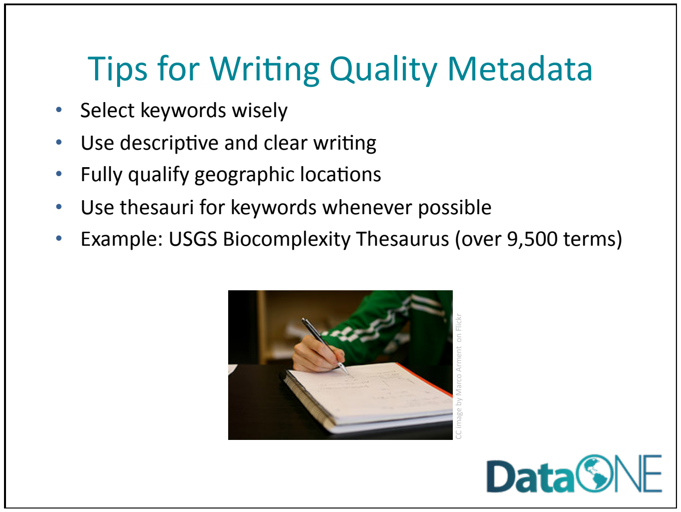
## Tips for Writing Quality Metadata

- Be specific and quantify when you can! The goal of a metadata record is to give the user enough information to know if they can use the data without contacting the dataset owner.

  Vague: We checked our work and it looks complete.

  Specific: We checked our work using a random sample of 5 monitoring sites reviewed by 2 different people. We determined our work to be 95% complete based on these visual inspections.

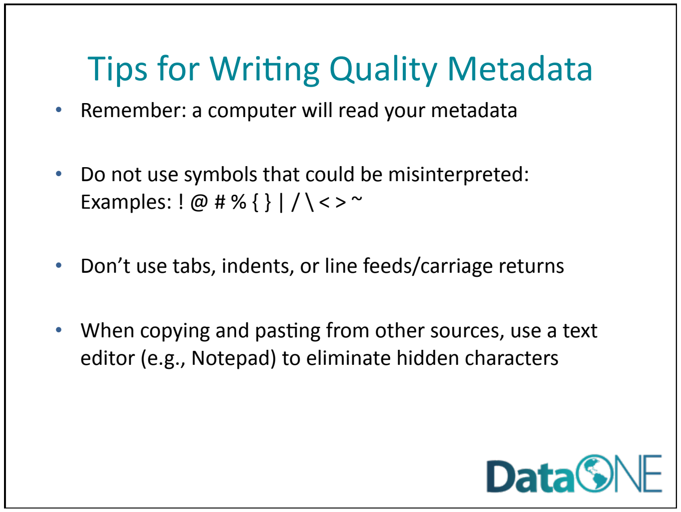
CC image by PNASH on Flickr



## Tips for Writing Quality Metadata

- Select keywords wisely
- Use descriptive and clear writing
- Fully qualify geographic locations
- Use thesauri for keywords whenever possible
- Example: USGS Biocomplexity Thesaurus (over 9,500 terms)

CC image by Marco Arment on Flickr



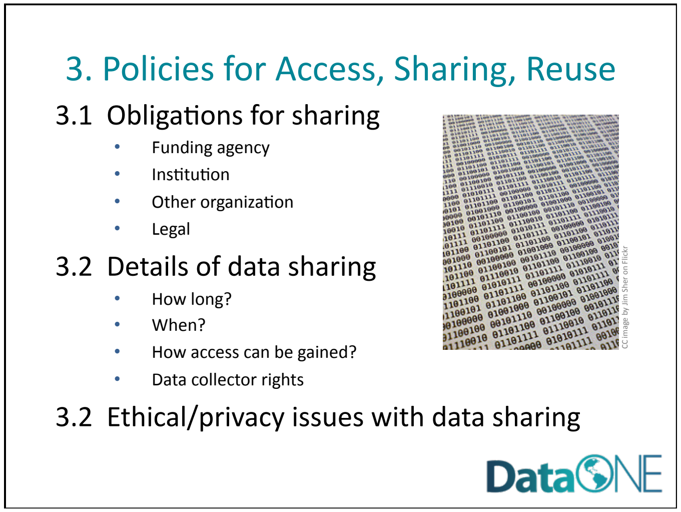
## Tips for Writing Quality Metadata

- Remember: a computer will read your metadata

- Do not use symbols that could be misinterpreted: Examples: ! @ # % { } | / \ < > ~

- Don’t use tabs, indents, or line feeds/carriage returns

- When copying and pasting from other sources, use a text editor (e.g., Notepad) to eliminate hidden characters



## 3. Policies for Access, Sharing, Reuse

3.1 Obligations for sharing

- Funding agency
- Institution
- Other organization
- Legal

3.2 Details of data sharing

- How long?
- When?
- How access can be gained?
- Data collector rights

3.2 Ethical/privacy issues with data sharing

CC image by Jim Sher on Flickr