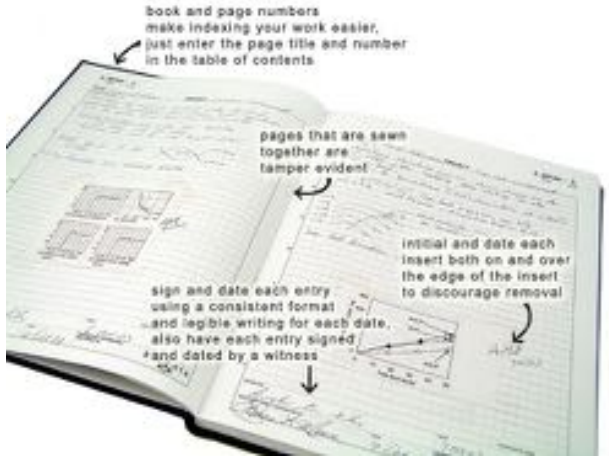
Chain of Data Creation

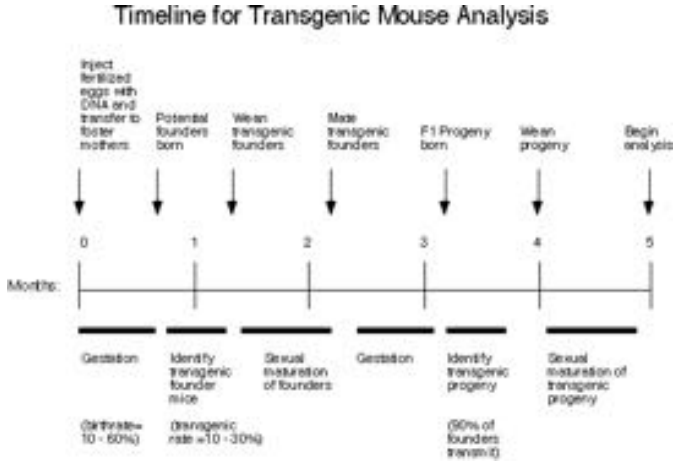
1. Preparation
2. Creation of Metadata
3. Acquisition
4. Building a Permanent Record
5. Data Management
6. Storage
7. Data Sharing

Lab Notebook



- Record of hypotheses
- Record of Protocols
- Second brain

Plan Your Experiment, Experiment With your Plan



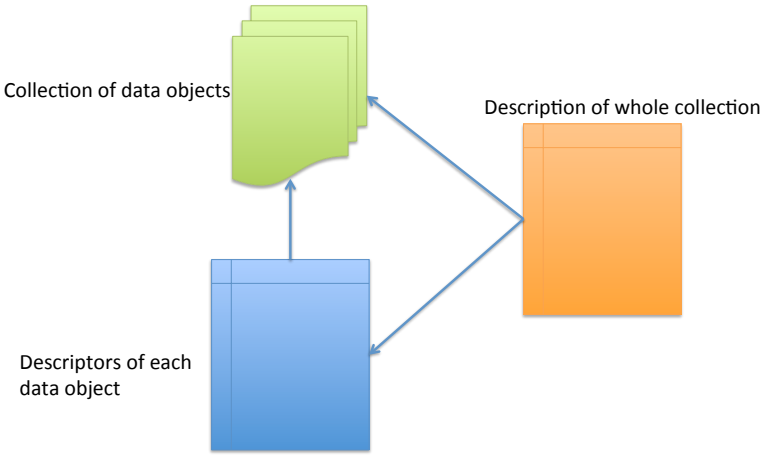
medicine.umich.edu

Question to Ask About your Data Collection Activity

- What am I measuring?
- When am I measuring it?
- How am I measuring it?
- What are the tools I am using?
- What about the lab/field environment do I need to know?
- Is my protocol reproducible?

Meta-Data

What is Metadata?



What is Metadata?

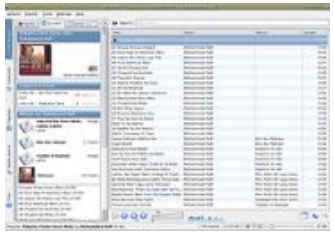
Metadata is: Data 'reporting'

- 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?

DataONE

Metadata in Real Life

- Metadata is all around…



Author(s)	Boullosa, Carmen.
Title(s)	They're cows, we're pigs / by Carmen Boullosa
Place	New York : Grove Press, 1997.
Physical Descr	viii, 180 p ; 22 cm.
Subject(s)	Pirates Caribbean Area Fiction.
Format	Fiction

Nutrition Facts

Serving Size 4 OZ. SERVING (112g)
Servings Per Container VARIED

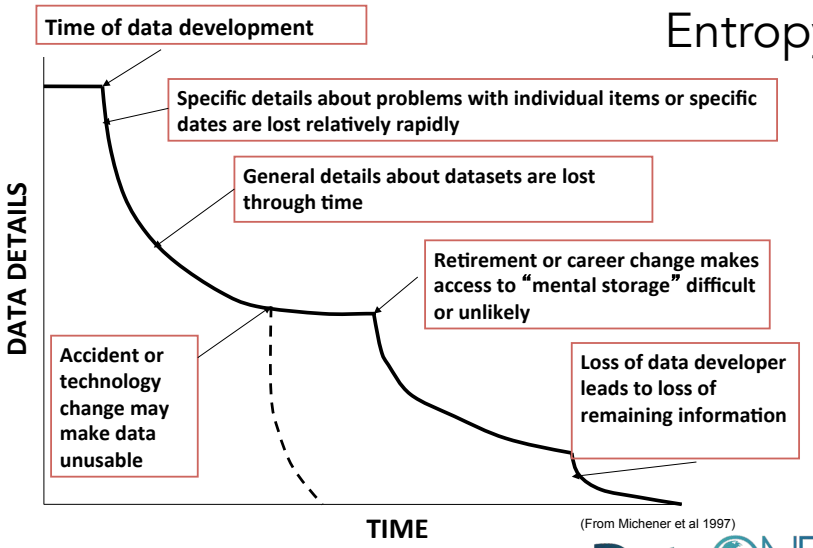
Amount Per Serving

Calories 170	Calories from Fat 70
	% Daily Value*
Total Fat 8g	12%
Saturated Fat 3g	15%
Cholesterol 65mg	22%
Sodium 70mg	3%
Total Carbohydrate 0g	0%
Dietary Fiber 0g	0%
Sugars 0g	
Protein 23g	
Vitamin A 0% • Vitamin C 0%	
Calcium 0% • Iron 15%	

*Percent Daily Values are based on a 2,000 calorie diet

DataONE

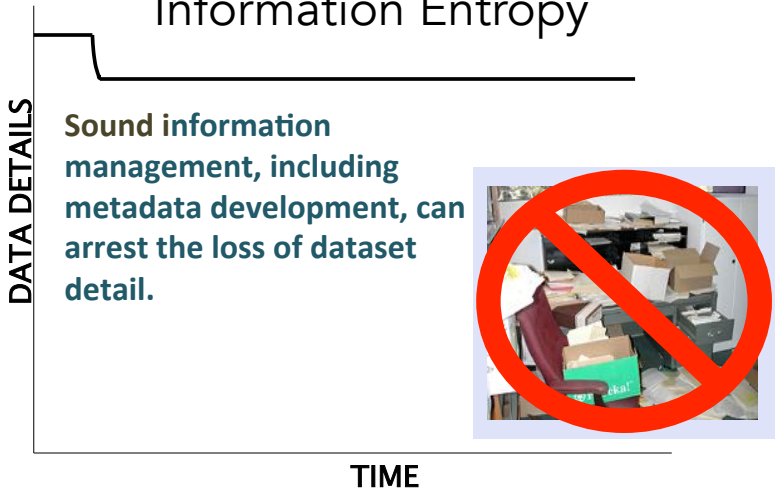
Information Entropy



(From Michener et al 1997)

DataONE

Information Entropy



DataONE

Data Management via Metadata

DataONE

NYR Daily

EMAIL PRINT

Tweet Share

'We Kill People Based on Metadata'

David Cole

Rick Bowmer/AP Photo

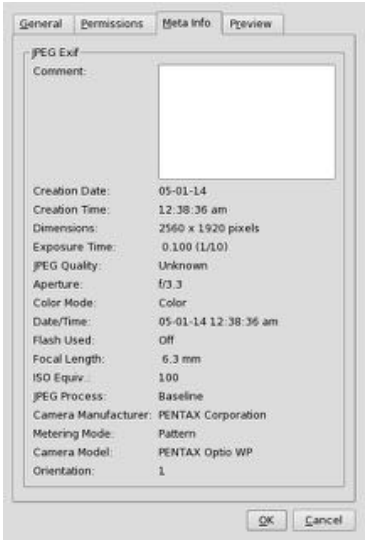
The National Security Agency's $1.5 billion data storage facility in Bluffdale, Utah, June 2013

Supporters of the National Security Agency inevitably defend its sweeping collection of phone and Internet records on the ground that it is only collecting so-called "metadata"—who you call, when you call, how long you talk. Since this does not

Cellphone Metadata

http://www.zeit.de/datenschutz/malte-spitz-data-retention

EXIF Metadata



What Meta-Data Do You Need?

- **Descriptive metadata** describes a resource for purposes such as discovery and identification
- **Administrative metadata** provides information to help manage a resource, such as when and how it was created
- **Rights management metadata**, which deals with intellectual property rights
- **Preservation metadata**, which contains information needed to archive and preserve a resource

Understanding Metadata: niso.org

Structured Metadata

Dublin Core Example

Title=”Metadata Demystified”
Creator=”Brand, Amy”
Creator=”Daly, Frank”
Creator=”Meyers, Barbara”
Subject=”metadata”
Description=”Presents an overview of metadata conventions in publishing.”
Publisher=”NISO Press”
Publisher=”The Sheridan Press”
Date=”2003-07"
Type=”Text”
Format=”application/pdf”
Identifier=”http://www.niso.org/standards/resources/Metadata_Demystified.pdf”
Language=”en”

Structured Metadata

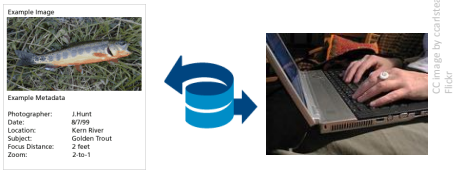
```
<mcp:parameterName>
    <mcp:DP_Term>
TAGS <mcp:term>
            <gco:CharacterString>tile_lower_right_y_LL</gco:CharacterString>
        </mcp:term>
        <mcp:type>
            <mcp:DP_TypeCode codeList="http://schemas.aodn.org.au/mcp-2.0/
schema/resources/Codelist/gmxCodelists.xml#DP_TypeCode"
codeListValue="shortName">shortName</mcp:DP_TypeCode>
        </mcp:type>
        <mcp:usedInDataset>
            <gco:Boolean>true</gco:Boolean>
        </mcp:usedInDataset>
    </mcp:DP_Term>
</mcp:parameterName>
```


A Simple Typology of Data

- Numerical
  - Continuous
  - Integers
  - Ordinal
- Controlled Vocabulary
  - Certain defined words with defined meanings
  - Has a reference ‘dictionary’
- Dates/Times
  - Many formats – POSIX
- Raw Text
- Other Media

What is a Metadata Standard?

- A Standard provides a structure to describe data with:
  - Common terms to allow consistency between records
  - Common definitions for easier interpretation
  - Common language for ease of communication
  - Common structure to quickly locate information
- In search and retrieval, standards provide:
  - Documentation structure in a reliable and predictable format for computer interpretation
  - A uniform summary description of the dataset



Metadata Standards

Metadata Standards

ABCD - Access to Biological Collection Data
A standard for the access to and exchange of primary biodiversity data, including specimens and observations.

Darwin Core
A body of standards, including a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries.

EML - Ecological Metadata Language
Ecological Metadata Language (EML) is a metadata specification particularly developed for the ecology discipline.

Genome Metadata
Descriptive data about single genomes within the Pathosystems Resource Integration Center.

ISA-Tab
A general purpose framework with which to capture and communicate metadata for data files from 'omics-based' experiments employing combinations of technologies.

MIBBI - Minimum Information for Biological and Biomedical Investigations
A common portal to a group of checklists of Minimum Information in nearly 40 biological disciplines.

Observ-OM
Used to integrate and compare observation data across experimental projects, disease databases, and clinical biobanks.

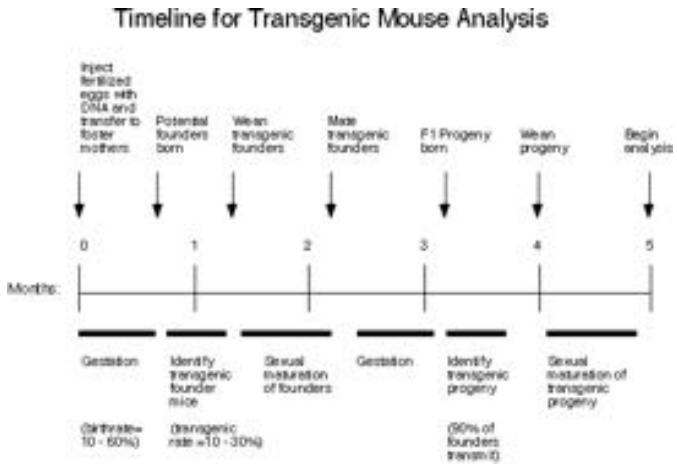
OME-XML - Open Microscopy Environment XML
A metadata standard and data file format for biological light microscopy data.

http://www.dcc.ac.uk/resources/subject-areas/biology

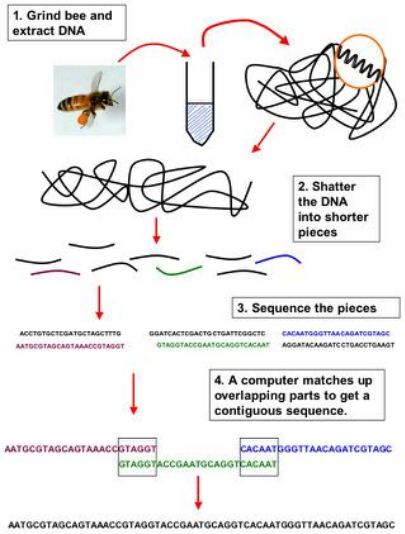
Britishlibrary.co.uk

Case Study 1

Case Study 2



Case Study 3



beespotter.org

Data Collection

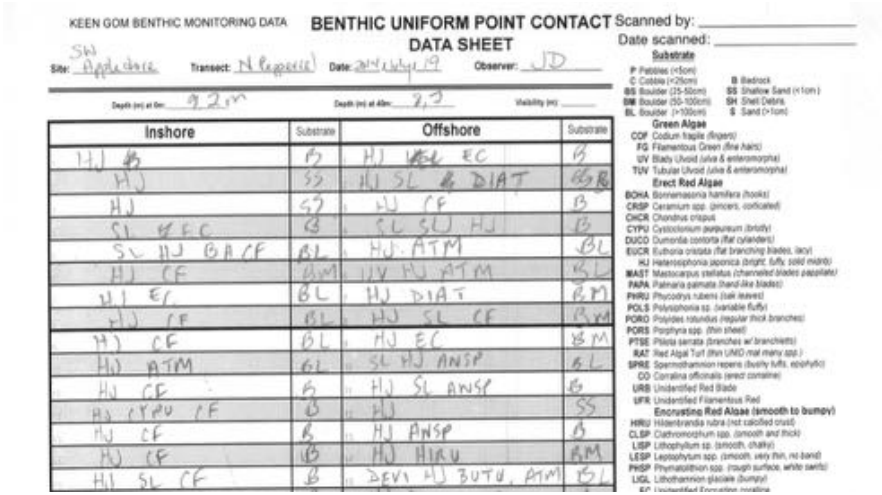
Creating a Good Data Gathering Sheet

- How easy is it to read?
- Are column and row definitions clear?
- Is there metadata?
- How similar is it to your digital data entry form?
- Can you use it at 4am?

After the Collection…

- Preserve original data
- Created digital archive of raw data
- Implement robust storage strategy
- Quality Control

Scanning



KEEN GOM BENTHIC MONITORING DATA

BENTHIC UNIFORM POINT CONTACT DATA SHEET

Scanned by: _____

Date scanned: _____

Site: SW Applachee Transect: N Peppered Date: 2/4/2019 Observer: JD

Depth (m) at 0m: 9.2m Depth (m) at 40m: 2.3 Visibility (m): _____

Inshore	Substrate	Offshore	Substrate
HJ	B	HJ EC	B
HJ	SS	HJ SL DIAT	GR
HJ	SS	HJ CF	B
SL EC	B	SL SLI HJ	B
SL HJ BACF	BL	HJ ATM	BL
HJ CF	BM	HJ ATM	BL
HJ EC	BL	HJ DIAT	BM
HJ CF	BL	HJ SL CF	BM
HJ CF	BL	HJ EC	BM
HJ ATM	BL	SL HJ ANSP	BL
HJ CF	B	HJ SL ANSP	B
HJ CYAN CF	B	HJ	SS
HJ CF	B	HJ ANSP	B
HJ CF	B	HJ HIRU	BM
HJ SL CF	B	DEVI HJ BUTU, ATM	BL

Substrate

P Pebbles (<5cm)
C Cobble (>25cm)
BS Boulder (25-50cm)
BM Boulder (50-100cm)
BL Boulder (>100cm)
B Bedrock
SS Shallow Sand (<1cm)
SH Shell Debris
S Sand (>1cm)

Green Algae
COF Codium fragile (fingers)
FG Filamentous Green (fine hairs)
UV Blady Ulvoid (ulva & enteromorpha)
TUV Tubular Ulvoid (ulva & enteromorpha)

Erect Red Algae
BOHA Bonnemaisonia hamifera (hooks)
CRSP Ceramium spp. (pincers, corticated)
CHCR Chondrus crispus
CYPU Cystoclonium purpureum (bristly)
DUCO Dumontia contorta (flat cylinders)
EUCR Euthora cristata (flat branching blades, lacy)
HJ Heterosiphonia japonica (bright, fluffy, solid midrib)
MAST Mastocarpus stellatus (channeled blades papillate)
PAPA Palmaria palmata (hand-like blades)
PHRU Phycodrys rubens (oak leaves)
POLS Polysiphonia sp. (variable fluffy)
PORO Polyides rotundus (regular thick branches)
PORS Porphyra spp. (thin sheet)
PTSE Pterosiphonia sp. (branches w/ branchlets)
RAT Red Algal Turf (thin UNID mat many spp.)
SPRE Spermothamnion repens (bushy tufts, epiphytic)
CO Corralina officinalis (erect corraline)
URB Unidentified Red Blade
URF Unidentified Filamentous Red

Encrusting Red Algae (smooth to bumpy)
HIRU Hildenbrandia rubra (not calcified crust)
CLSP Clathromorphum spp. (smooth and thick)
LISP Lithophyllum sp. (smooth, chalky)
LESP Leptophyllum spp. (smooth, very thin, no band)
PHSP Phymatolithon spp. (rough surface, white swirls)
LIGL Lithothamnion glaciale (bumpy)

Storage: Physical

Whitney
@arieswym

Rutgers PhD student is looking for stolen laptop w/ 5 years of research for dissertation. #phdchat pic.twitter.com/2WdTb2JIpr

DO NOT LET THIS BE YOU

Storage: Physical

Storage: The Cloud

Dropbox

GitHub

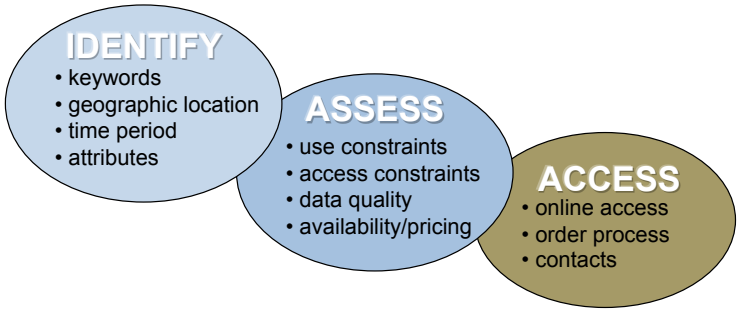
Google Drive

Data Sharing

blog.veritythink.com

Distribution: Data Discovery

- The descriptive content of the metadata file can be used to identify, assess, and access available data resources.



Things to Consider when Data Sharing

1. Is what you did understandable?
2. How do you want your work credited?
3. Will your data sharing service be around in 50 years?

Why Share Data?

- One scientist can only do so much
  - More data = more Power
- Science must be reproducible
- Who paid for this data collection?

Examples

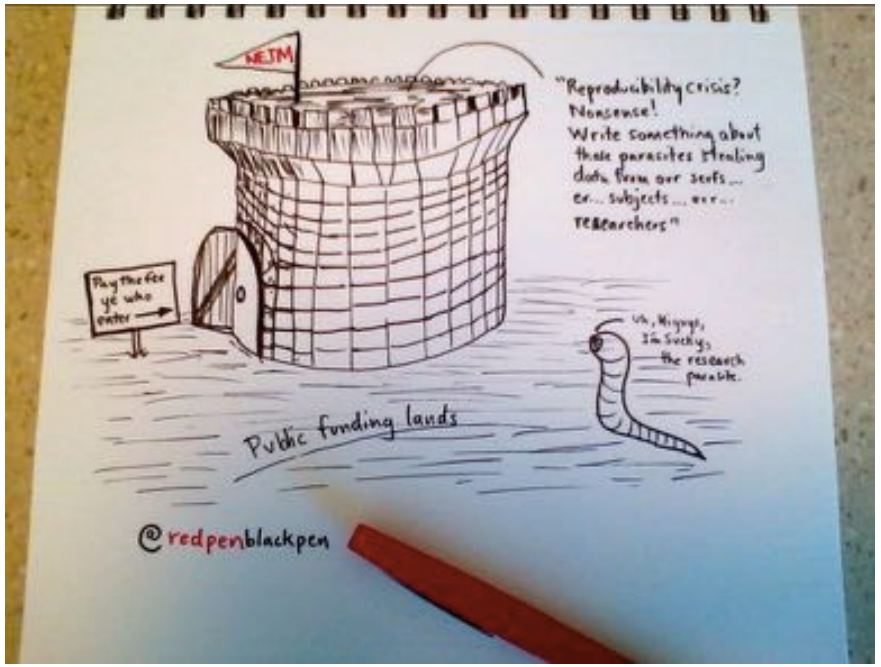
- https://www.dataone.org/
- http://blast.ncbi.nlm.nih.gov/Blast.cgi
- http://datadryad.org/
- http://www.oceandataportal.org/

Backlash?

The NEW ENGLAND

"A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as 'research parasites.'

quality information carefully reexamined for the possibility that new nuggets of useful data are lying there, previously unseen? The potential for leveraging existing results for even more benefit pays appropriate increased tribute to the patients who put themselves at risk to generate the data. The moral imperative to honor their collective sacrifice is the trump card that takes this trick.



Should all Data be Open?

Let's Look at Data & Metadata Examples

https://biol355.github.io/datasets.html