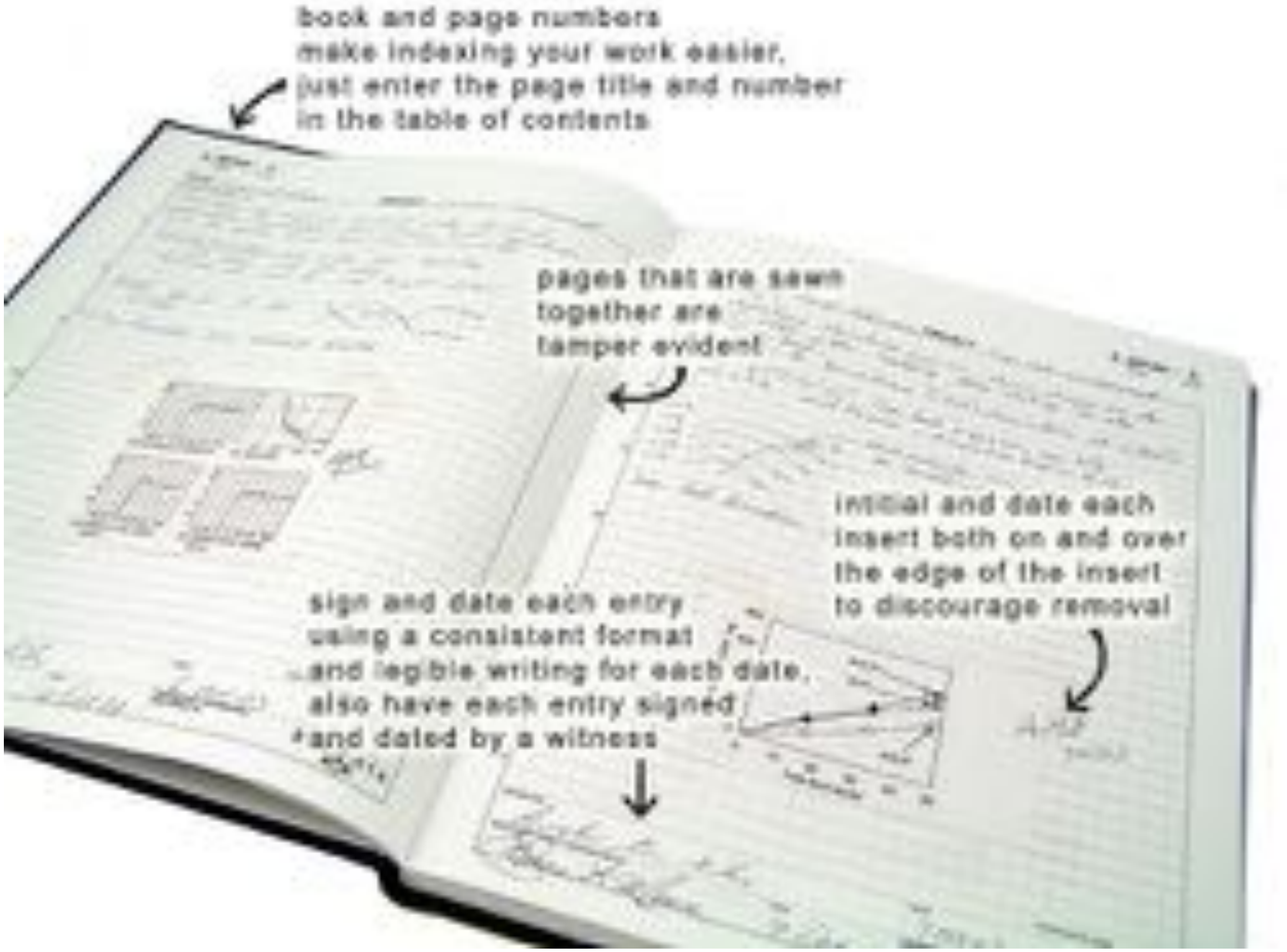
Data Creation

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



book and page numbers make indexing your work easier. just enter the page title and number in the table of contents

pages that are sewn together are tamper evident

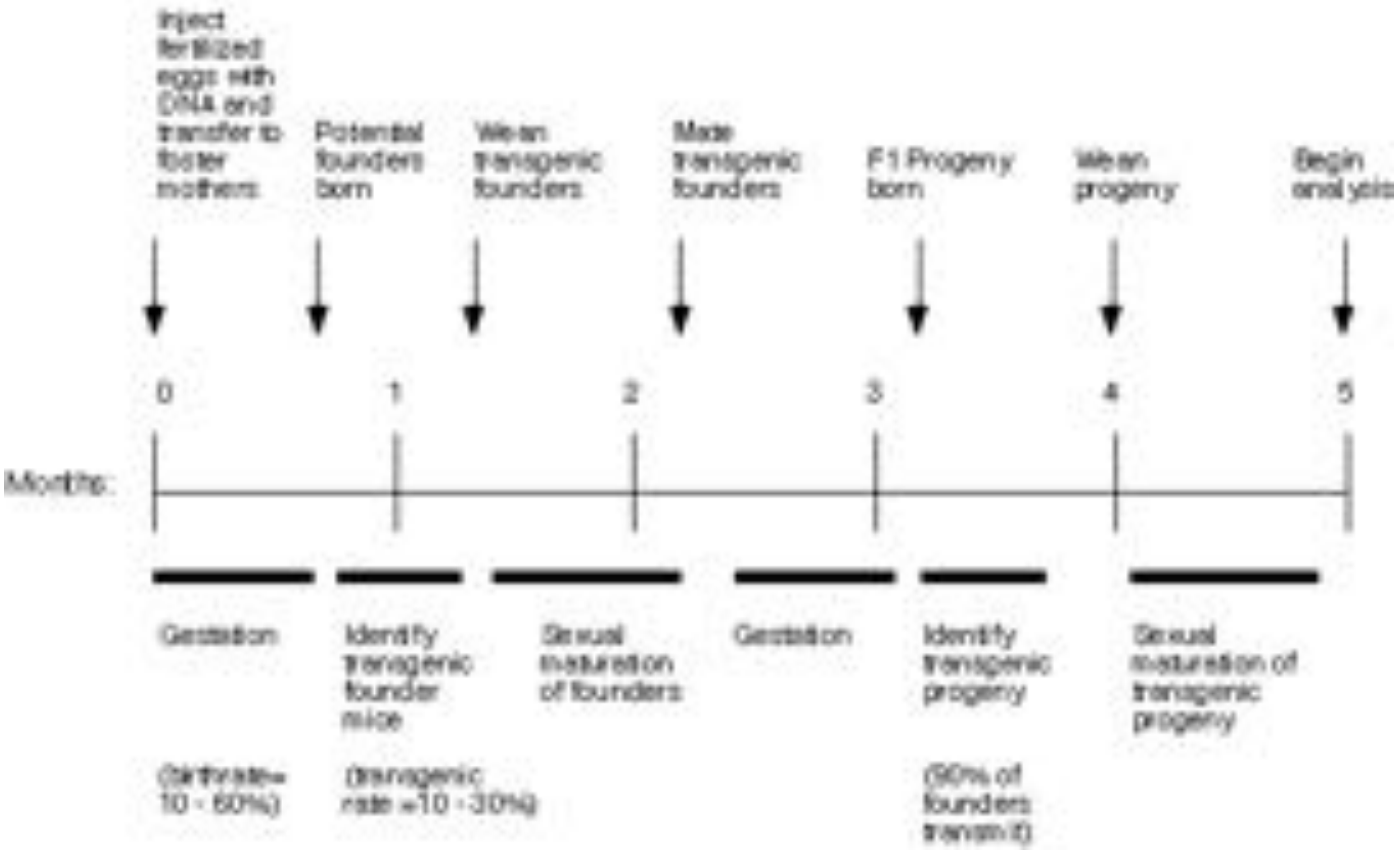
initial and date each insert both on and over the edge of the insert to discourage removal

sign and date each entry using a consistent format and legible writing for each date. also have each entry signed and dated by a witness

- Record of hypotheses
- Record of Protocols
- Second brain

Plan Your Experiment, Experiment With your Plan

Timeline for Transgenic Mouse Analysis


Inject fertilized eggs with DNA and transfer to foster mothers
Potential founders born
Wean transgenic founders
Mate transgenic founders
F1 Progeny born
Wean progeny
Begin analysis
0
1
2
3
4
5
Months:
Gestation
Identify transgenic founder mice
Sexual maturation of founders
Gestation
Identify transgenic progeny
Sexual maturation of transgenic progeny
(birth rate= 10 - 60%)
(transgenic rate =10 - 30%)
(90% of founders transmit)

medicine.umich.edu

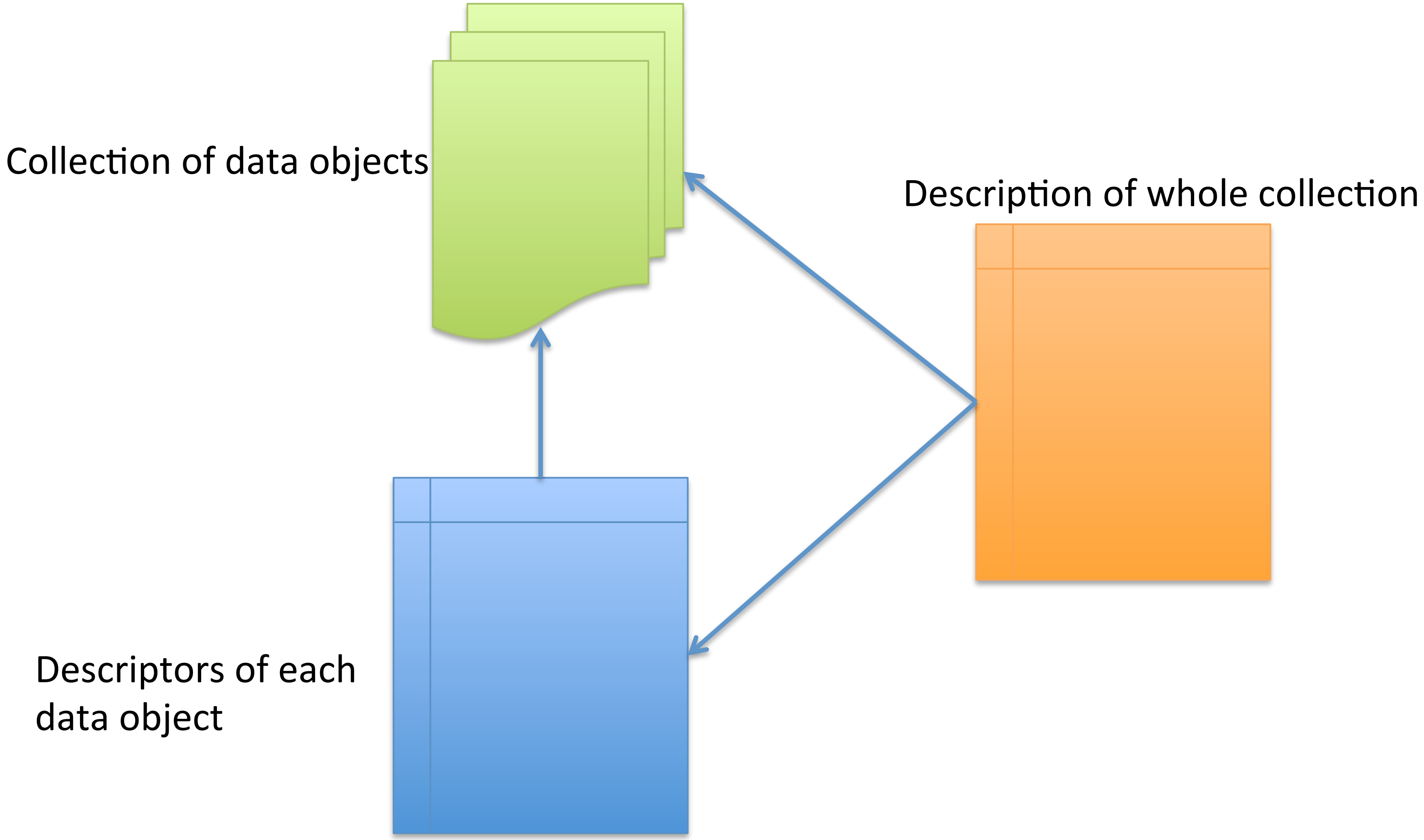
Question to Ask About your Data Collection Activity

- 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

metadata
describe
context
entities
systems
objects
definitions
additional
structured
using
one
code
people
information
Data

What is Metadata?



Collection of data objects

Description of whole collection

Descriptors of each data object

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

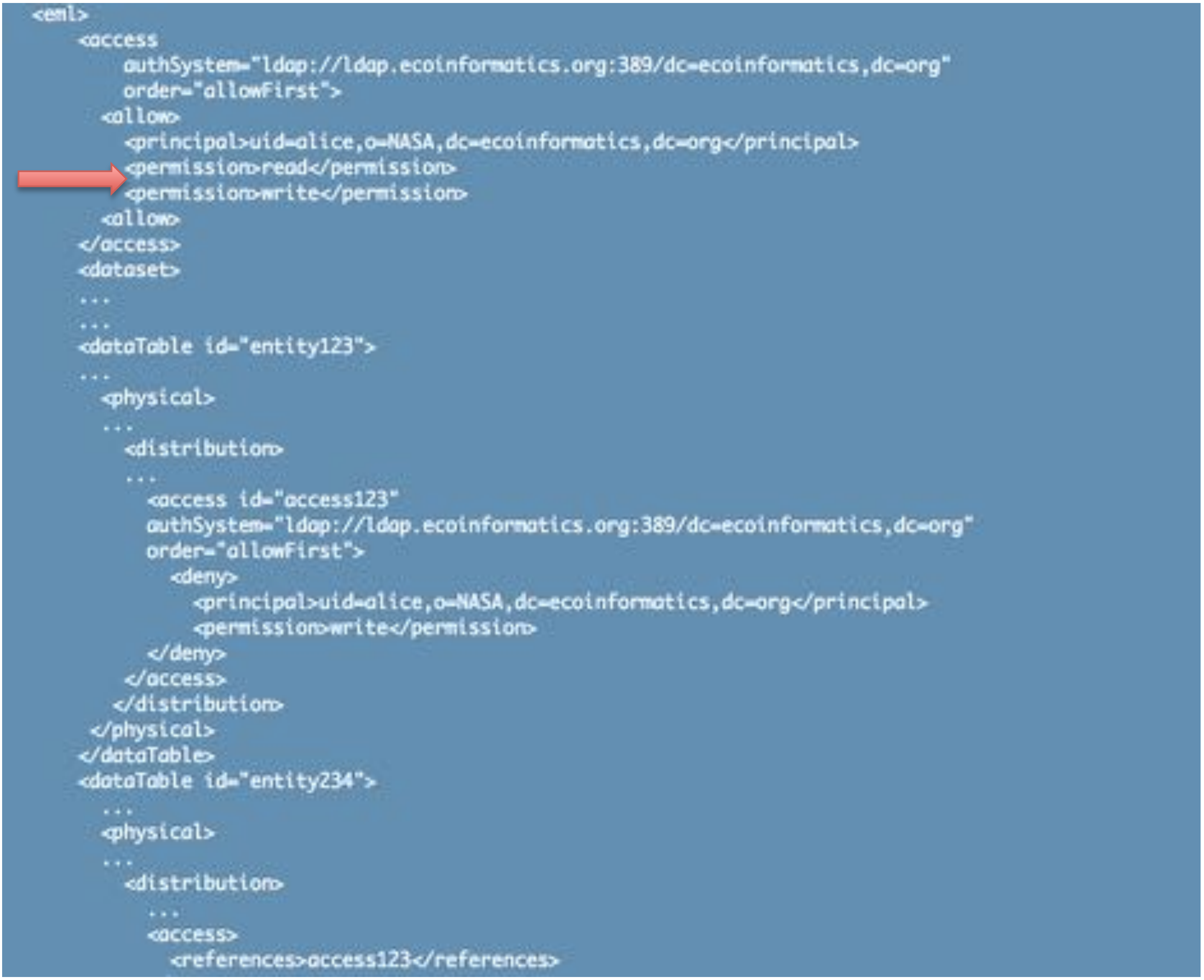
Dublin Core Example

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”

Understanding Metadata: niso.org

Structured Metadata

TAGS



```
<eml>
    <access
        authSystem="ldap://ldap.ecoinformatics.org:389/dc=ecoinformatics,dc=org"
        order="allowFirst">
      <allow>
        <principal>uid=alice,o=NASA,dc=ecoinformatics,dc=org</principal>
        <permission>read</permission>
        <permission>write</permission>
      <allow>
    </access>
    <dataset>
    ...
    ...
    <dataTable id="entity123">
    ...
      <physical>
      ...
        <distribution>
        ...
          <access id="access123"
          authSystem="ldap://ldap.ecoinformatics.org:389/dc=ecoinformatics,dc=org"
          order="allowFirst">
            <deny>
              <principal>uid=alice,o=NASA,dc=ecoinformatics,dc=org</principal>
              <permission>write</permission>
          </deny>
        </access>
       </distribution>
     </physical>
    </dataTable>
    <dataTable id="entity234">
      ...
      <physical>
      ...
        <distribution>
          ...
          <access>
            <references>access123</references>
```

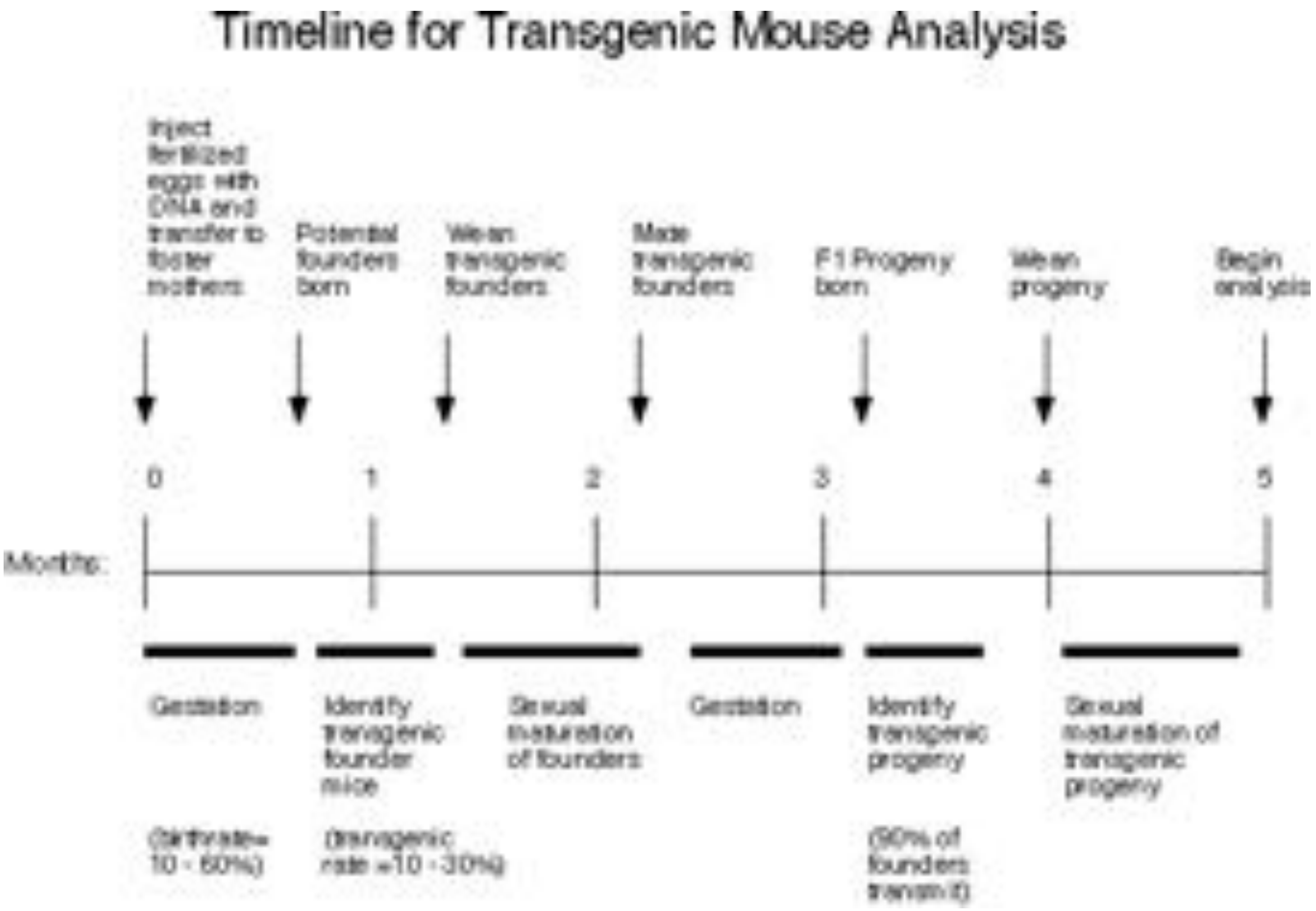
KNB

Britishlibrary.co.uk

Case Study 1

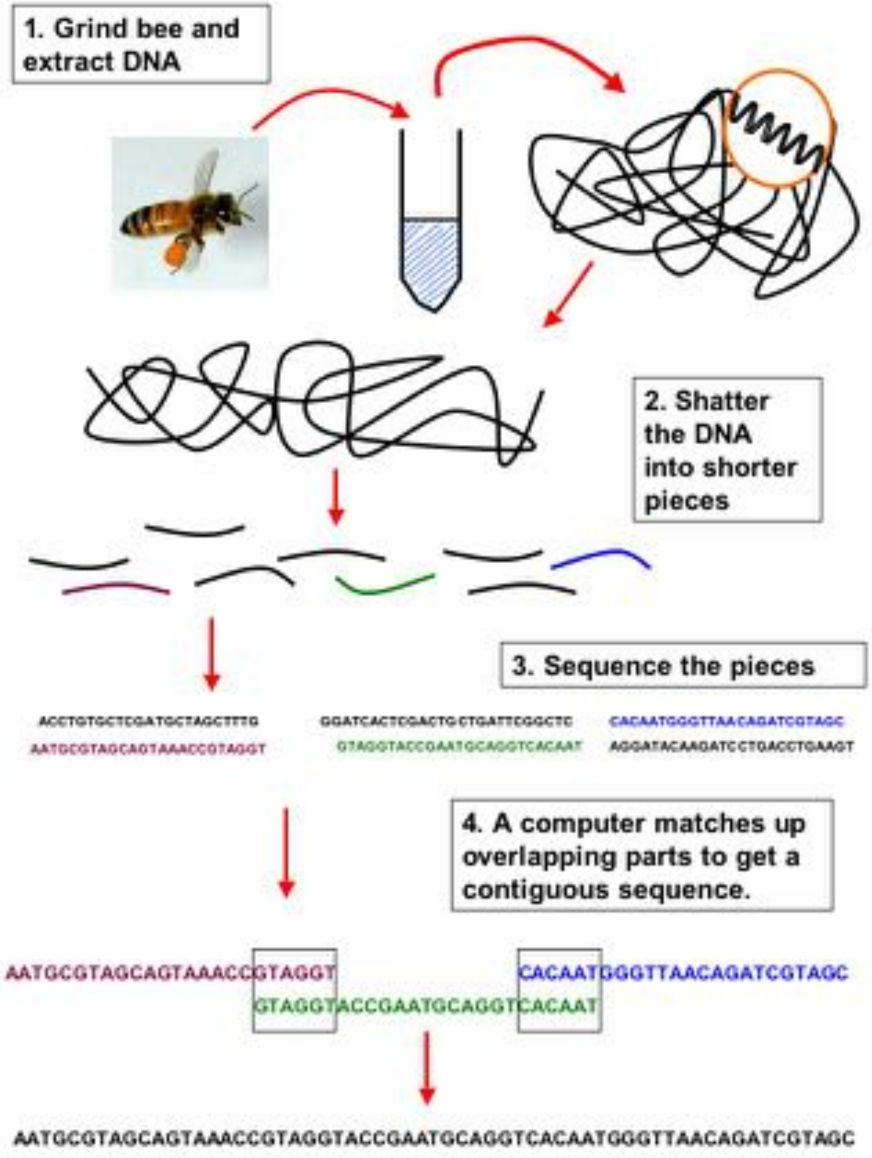
Digital Image © 2002 University of Wisconsin-Milwaukee Libraries

Case Study 2


Timeline for Transgenic Mouse Analysis

Inject fertilized eggs with DNA and transfer to foster mothers
Potential founders born
Wean transgenic founders
Mate transgenic founders
F1 Progeny born
Wean progeny
Begin analysis
0
1
2
3
4
5
Months:
Gestation
Identify transgenic founder mice
Sexual maturation of founders
Gestation
Identify transgenic progeny
Sexual maturation of transgenic progeny
(birthrate= 10 - 60%)
(transgenic rate =10 -30%)
(90% of founders transmit)

Case Study 3



1. Grind bee and extract DNA

2. Shatter the DNA into shorter pieces

3. Sequence the pieces

ACCTGTGCTCGATGCTAGCTTTG

AATGCGTAGCAGTAAACCGTAGGT

GGATCACTCGACTGCTGATTCGGCTC

GTAGGTACCGAATGCAGGTCACAAT

CACAATGGGTTAACAGATCGTAGC

AGGATACAAGATCCTGACCTGAAGT

4. A computer matches up overlapping parts to get a contiguous sequence.

AATGCGTAGCAGTAAACCGTAGGT

GTAGGTACCGAATGCAGGTCACAAT

CACAATGGGTTAACAGATCGTAGC

AATGCGTAGCAGTAAACCGTAGGTACCGAATGCAGGTCACAATGGGTTAACAGATCGTAGC

beespotter.org

Data Collection

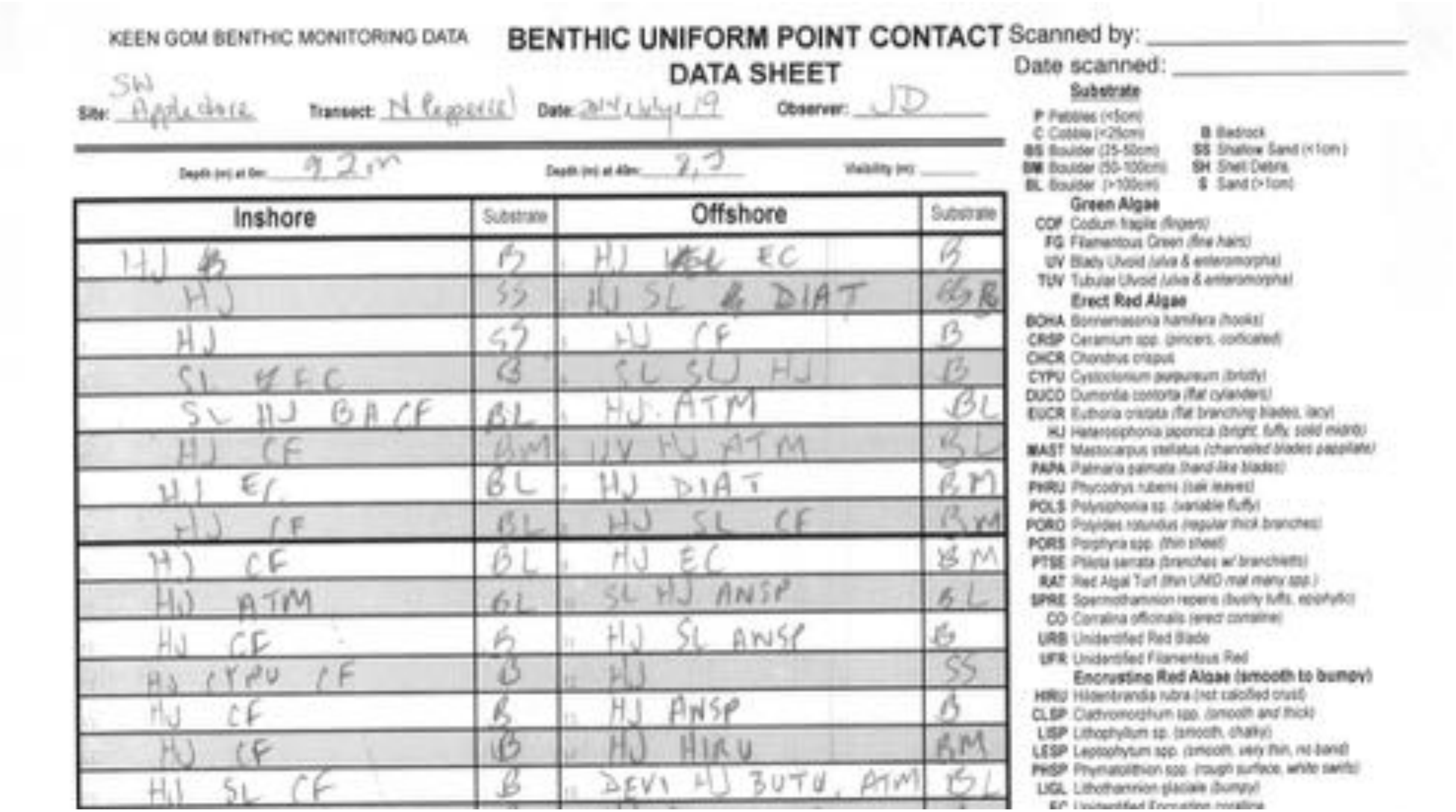
Creating a Good Data Gathering Sheet

- 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 (next time)

Scanning



KEEN GOM BENTHIC MONITORING DATA

BENTHIC UNIFORM POINT CONTACT DATA SHEET

Scanned by: ______________

Date scanned: ______________

Site: SW Appledore Transect: N Pepperell Date: 2014 July 19 Observer: JD

Depth (m) at 0m: 9.2m Depth (m) at 40m: 8.3 Visibility (m): ______

Inshore	Substrate	Offshore	Substrate
HJ ~~B~~	B	HJ ~~VSL~~ EC	B
HJ	SS	HJ SL ~~B~~ DIAT	~~BS~~ B
HJ	SS	HJ CF	B
SL ~~E~~ EC	B	SL SU HJ	B
SL HJ BA CF	BL	HJ ATM	BL
HJ CF	BM	UV 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 CYPU 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)	B	Bedrock
BS	Boulder (25-50cm)	SS	Shallow Sand (<1cm)
BM	Boulder (50-100cm)	SH	Shell Debris
BL	Boulder (>100cm)	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 cylanders)
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	Ptilota serrata (branches w/ branchlets)
RAT	Red Algal Turf (thin UMID mat many spp.)
SPRE	Spermothamnion repens (bushy tufts, epiphytic)
CO	Corralina officinalis (erect corraline)
URB	Unidentified Red Blade
UFR	Unidentified Filamentous Red

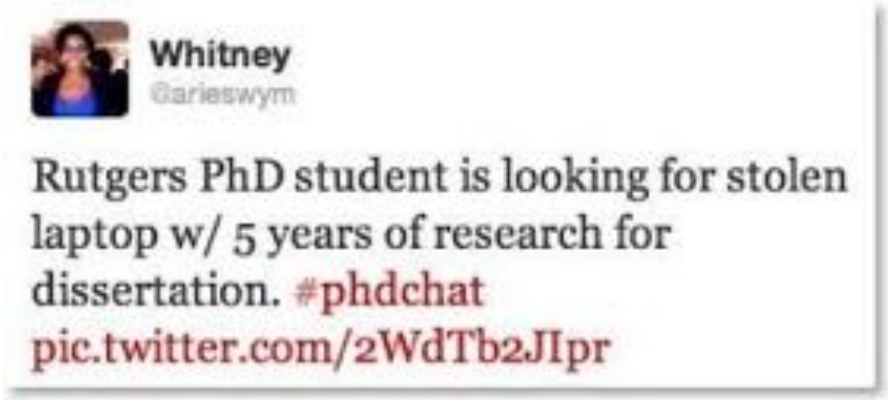
Encrusting Red Algae (smooth to bumpy)

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

EXCEL TIME!

Entry
Fills
Basic Functions
Functions for Error Checking
Controlled Vocabularies

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

Data Sharing

blog.veritythink.com

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

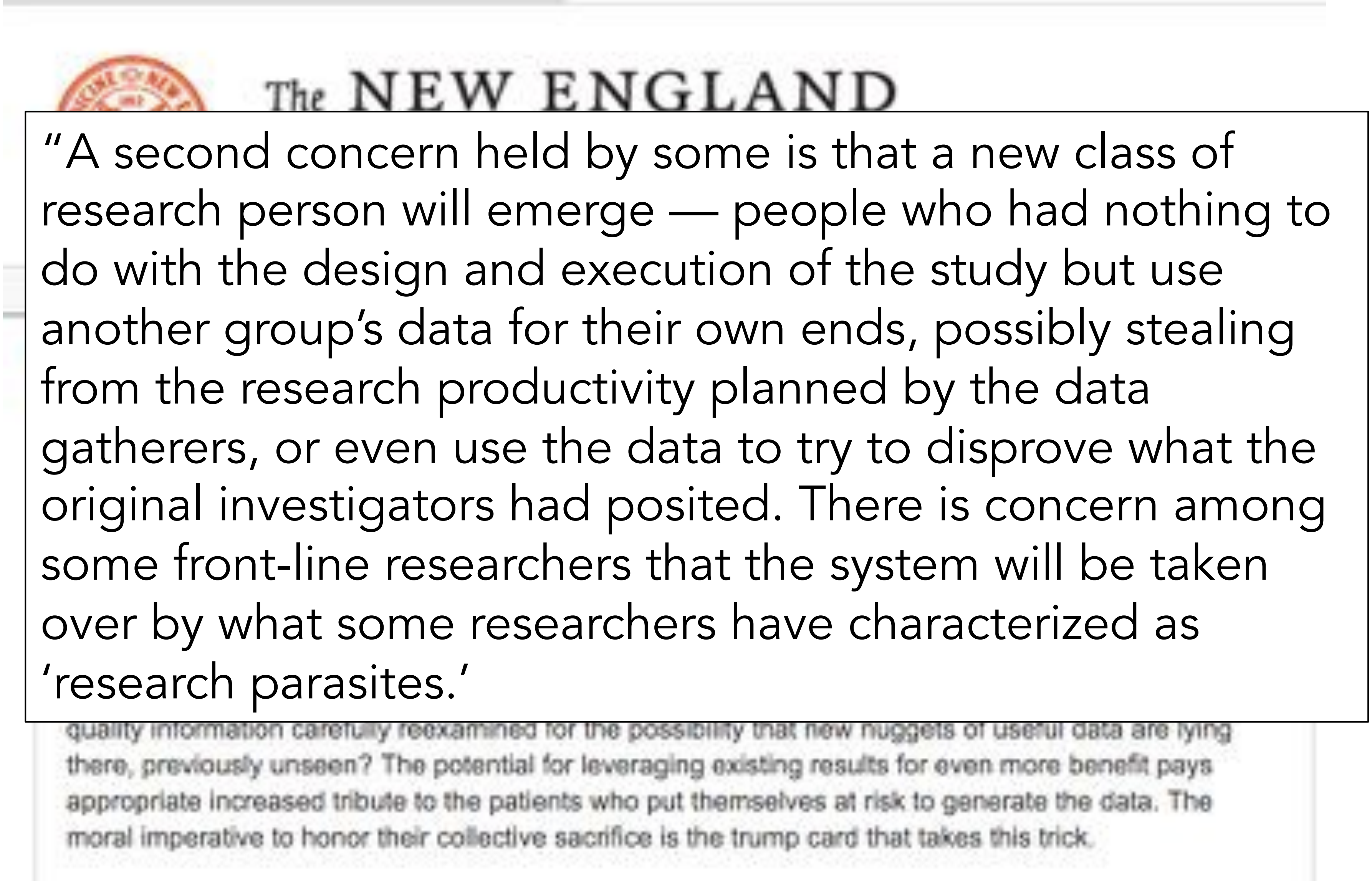
- Science must be reproducible

- 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?

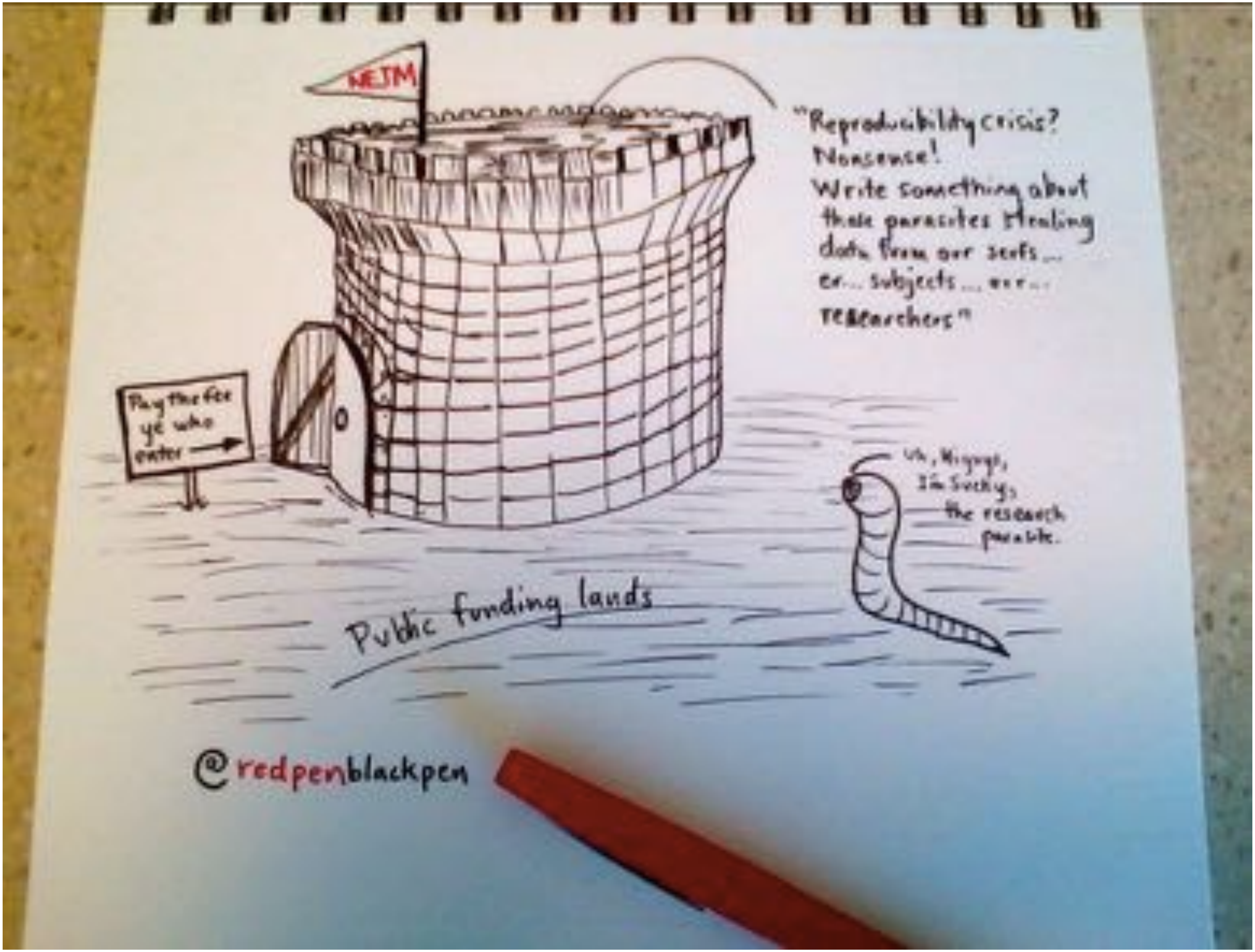


The NEW ENGLAND

“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.

If I have seen further
it is because I am parasitic on research giants


NEJM
"Reproducibility crisis? Nonsense! Write something about these parasites stealing data from our serfs... er... subjects... err... researchers"
Pay the fee ye who enter
Uh, Hiyyyy, I'm Sucky, the research parasite.
Public funding lands
@redpenblackpen

Should all Data be Open?

Sucky, the research parasite, in...

The Parasites of Penzance!!!

♩ I am the very model
of a modern
research parasite, ♫
I build on other research
leaving only here and there a cite.
I know the Nobel laureates
and quote the fights cladistical,
From Felsenstein to
Farris using language
♪ most statistical.
I know our mythic
history, from Sir Galton
to Chetverikov, ♫
And advocate the medicines you cannot get generic of.
I've memorized the sequence of each ENCODE base of DNA, ♪
And I can tell at sight Feng Zhang from Doudna and Charpentier.
But still I need your data like anemics need their ferrocite
I am the very model of a modern research parasite!

Lyrics: @CT_Bergstrom
Costume and choreography: @redpenblackpen
(see, we're <u>co-authors</u>!)