Data Management and Resource Sharing

Rigor & Reproducibility Workshop
27 May 2021

Melissa Eitzen, MT(ASCP), MS, RQAP-GLP
Director, Regulatory Operations
UTMB Institutional Office of Regulated Nonclinical Studies
mmeitzen@utmb.edu

utmb Health

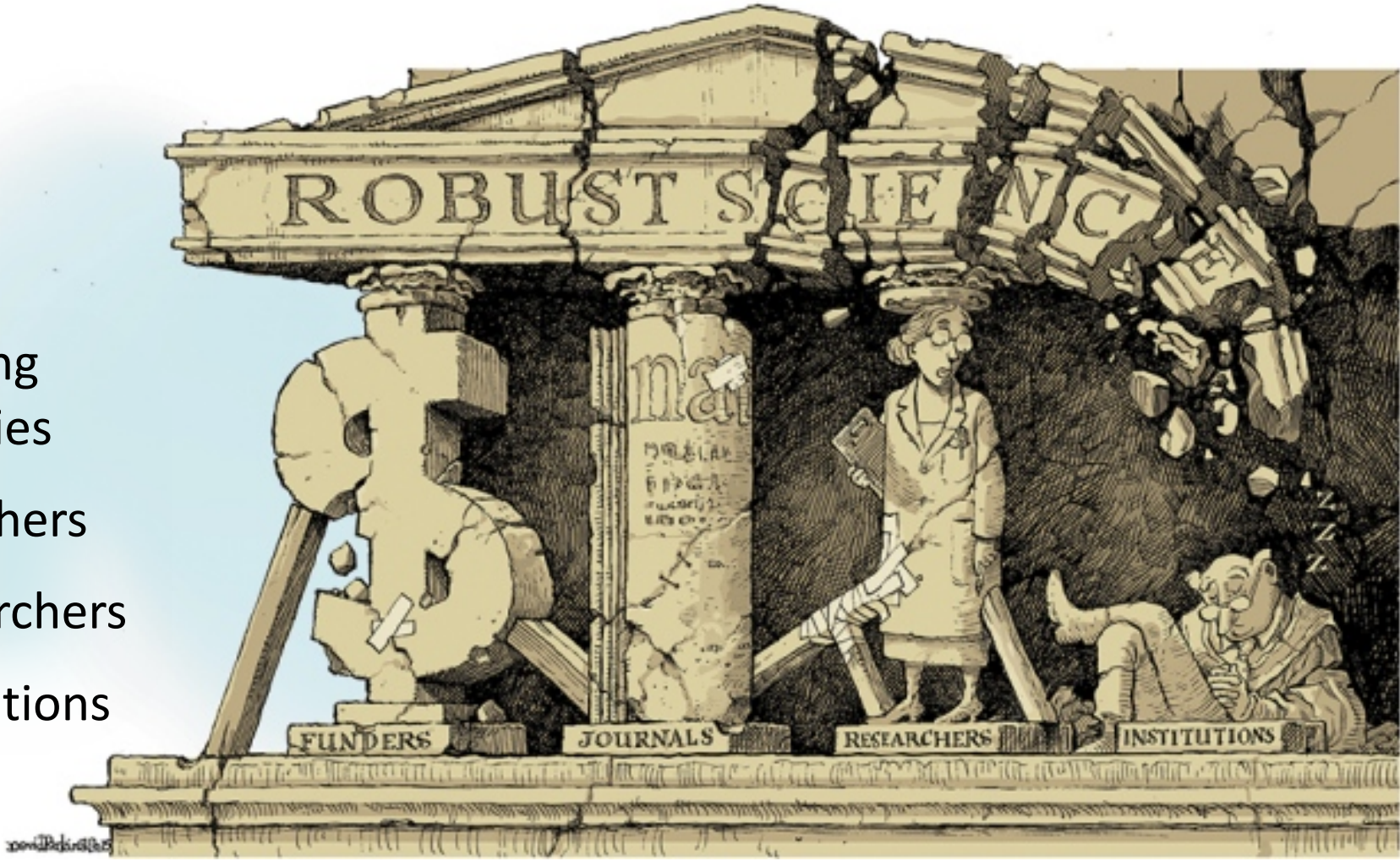
Topics

- Principles & Guidelines
- Data Lifecycle
- Data Quality & Integrity
- Case Study—Break out session

References provided on slides

Stakeholders of Robust Science

- Funding agencies
- Publishers
- Researchers
- Institutions



https://www.nature.com/news/robust-research-institutions-must-do-their-part-for-reproducibility-1.18259

NIH Public Workshop (2014)

- **Funding agencies**
- **Publishers**
- Researchers
- Institutions

- **Sponsors:** NIH + Nature Publishing Group + Science
- **Issue:** Reproducibility, Rigor of research findings
- **Attendees:** Journal editors (>30 basic/preclinical science journals where NIH-funded investigators publish)
- **Goals:** Identify common opportunities in the scientific publishing arena to ***enhance rigor and further support research that is reproducible, robust, and transparent***
- **Outcome:** set of principles to facilitate these goals, which a considerable number of journals have agreed to endorse

https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

NIH Principles and Guidelines

- **Funding agencies**
- Publishers
- Researchers
- Institutions

Principles and Guidelines for Reporting Preclinical Research:

- Rigorous statistical analysis
- Transparency in reporting
- Data and material sharing
- Consider establishing best practice guidelines for:
  - Images
  - Biological materials (antibodies, cell lines, etc.)
  - Animals
- Endorsements (journals, associations, societies)
- Adapted Guidelines

https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

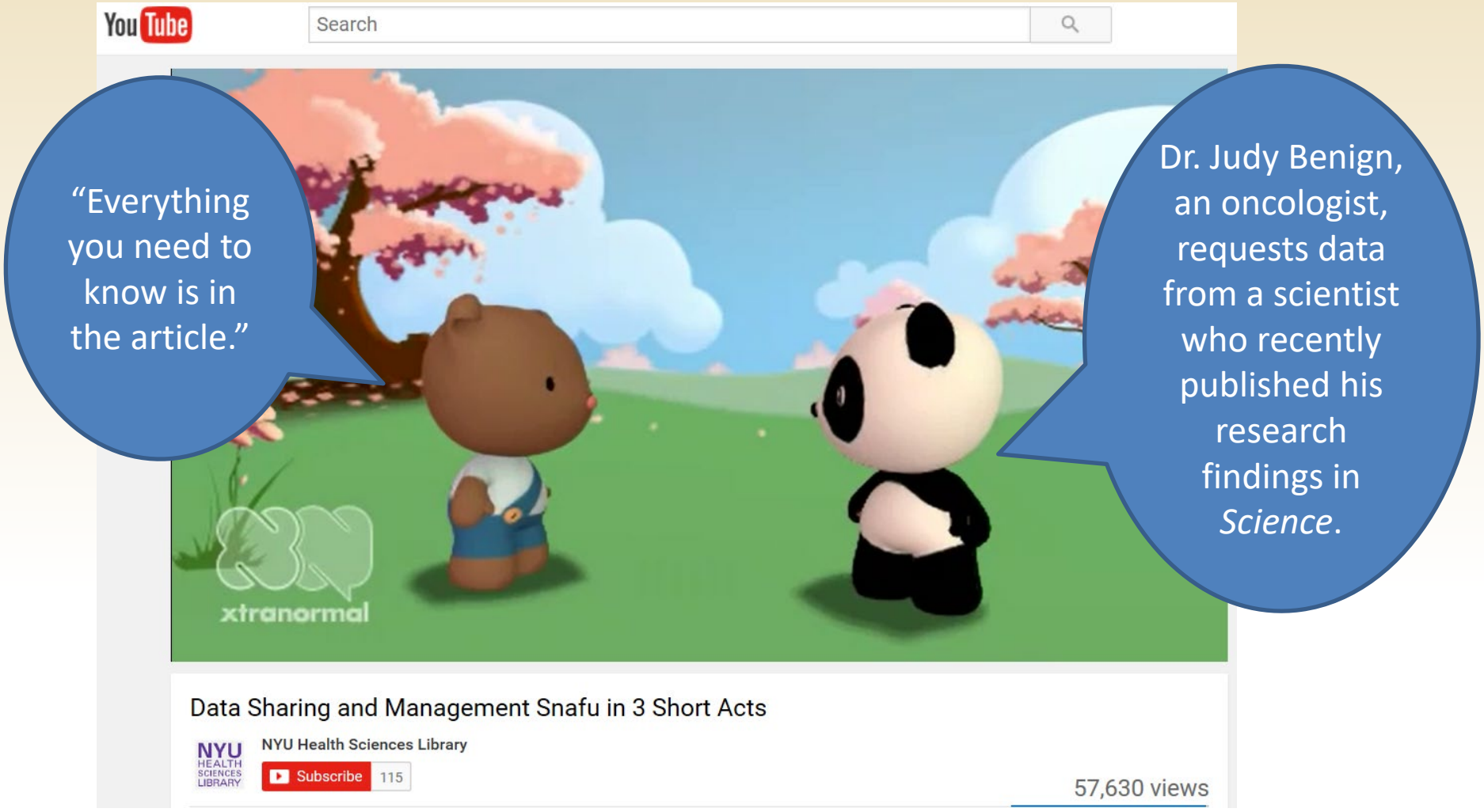
Data and Material Sharing

- **Funding agencies**
- **Publishers**
- Researchers
- Institutions

- Require datasets be made available (where ethically appropriate) upon request
  - during manuscript review
  - upon publication
- Recommend datasets in public repositories, where available
- Encourage presentation of all other data values in machine readable format in the paper (or supplementary information)
- Require materials sharing after publication
- Encourage sharing of software
- Require a statement in the manuscript describing if software is available and how it can be obtained

https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

Why is Data Management and Resource Sharing Important?



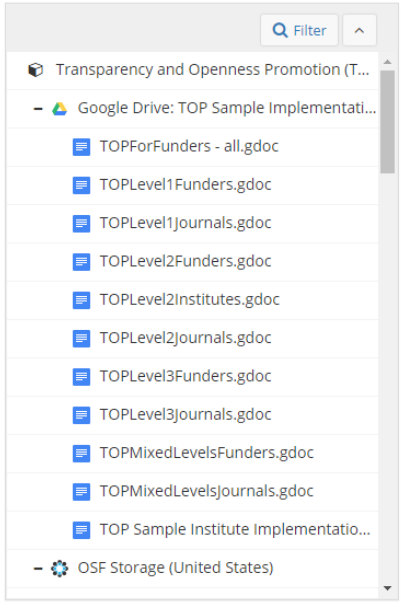
https://www.youtube.com/watch?v=N2zK3sAtr-4

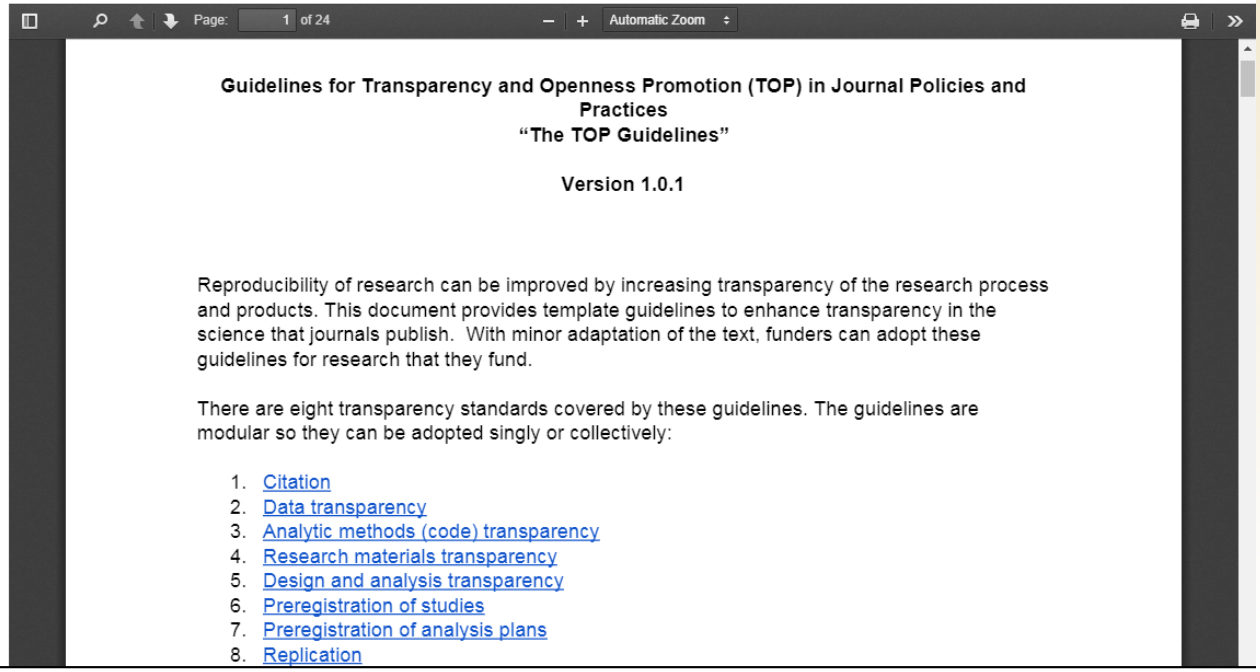
Guidelines for Transparency & Openness Promotion (TOP)

- Funding agencies
- **Publishers**
- Researchers
- Institutions

TOPGuidelines.pdf (Version: 5)

Download | Share | View | Revisions





- Adapted Guidelines

https://osf.io/ud578/?_ga=1.211230620.829898984.1435325845

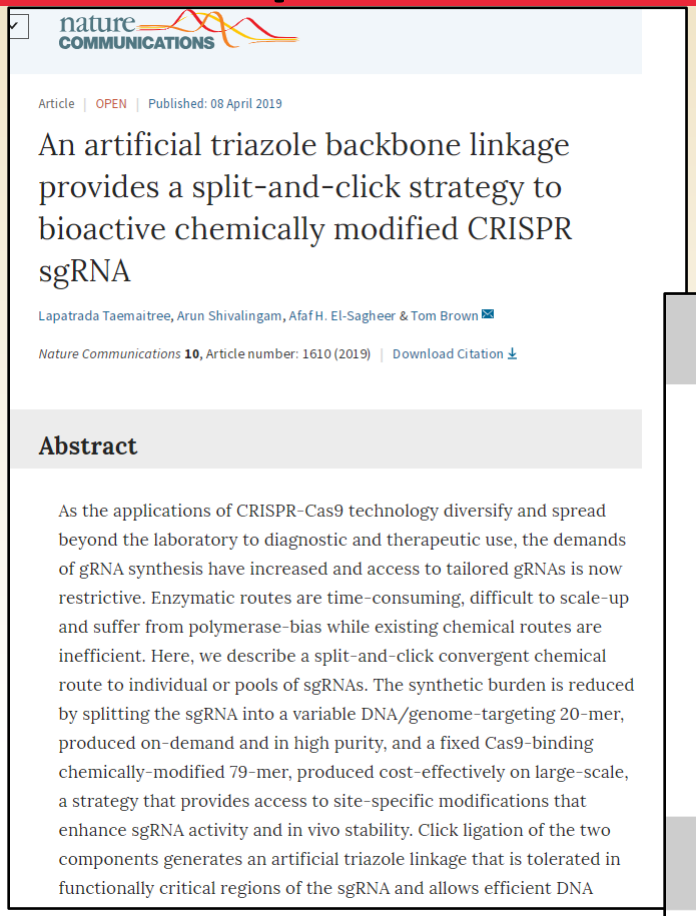
TOP Guidelines—Data Transparency

- Funding agencies
- **Publishers**
- Researchers
- Institutions

- Disclosure—are your data available?
- Share your data (exception: legal/ethical restraints)
- Can results be replicated using your data (pre-publication)

Data Transparency—Example

- Funding agencies
- **Publishers**
- **Researchers**
- Institutions



nature COMMUNICATIONS

Article | OPEN | Published: 08 April 2019

An artificial triazole backbone linkage provides a split-and-click strategy to bioactive chemically modified CRISPR sgRNA

Lapatrada Taemaitree, Arun Shivalingam, Afaf H. El-Sagheer & Tom Brown

Nature Communications **10**, Article number: 1610 (2019) | Download Citation

Abstract

As the applications of CRISPR-Cas9 technology diversify and spread beyond the laboratory to diagnostic and therapeutic use, the demands of gRNA synthesis have increased and access to tailored gRNAs is now restrictive. Enzymatic routes are time-consuming, difficult to scale-up and suffer from polymerase-bias while existing chemical routes are inefficient. Here, we describe a split-and-click convergent chemical route to individual or pools of sgRNAs. The synthetic burden is reduced by splitting the sgRNA into a variable DNA/genome-targeting 20-mer, produced on-demand and in high purity, and a fixed Cas9-binding chemically-modified 79-mer, produced cost-effectively on large-scale, a strategy that provides access to site-specific modifications that enhance sgRNA activity and in vivo stability. Click ligation of the two components generates an artificial triazole linkage that is tolerated in functionally critical regions of the sgRNA and allows efficient DNA

Data availability

Sequencing data that support the findings of this study have been deposited in the NCBI Sequencing Read Archive with the accession code PRJNA512007. The source data underlying Figs. 1B–D, 2C–E, 3B–C, 4A–C; Supplementary Figs. 1, 2, 3B–C, 4, 5, 6, 7, 8, 9, 10; Supplementary Tables 1, 3, 4, 5, 6 and Supplementary Data 1 and 2 are provided in the Source Data file. All data for gels, graphs and mass spectrometry are provided as a Source Data file.

Code availability

Post-CIRCLE-seq data plotting code is available upon request.

Resource Sharing—NIH

- **Funding agencies**
- Publishers
- Researchers
- Institutions

NIH considers the sharing of unique research resources developed through NIH-sponsored research an important means to **enhance the value and further the advancement of research**.

When resources have been developed with NIH funds and the associated research findings published or provided to NIH, it is important that the **results be made readily available** for research purposes to qualified individuals within the scientific community.

https://grants.nih.gov/grants/peer/guidelines_general/Resource_sharing_plans.pdf

Resource Sharing—NIH

- **Funding agencies**
- Publishers
- Researchers
- Institutions

- Samples
- Reagents
- Model organism (e.g., transgenic mouse strain)
- Data

Data—Definition

Definition of Data

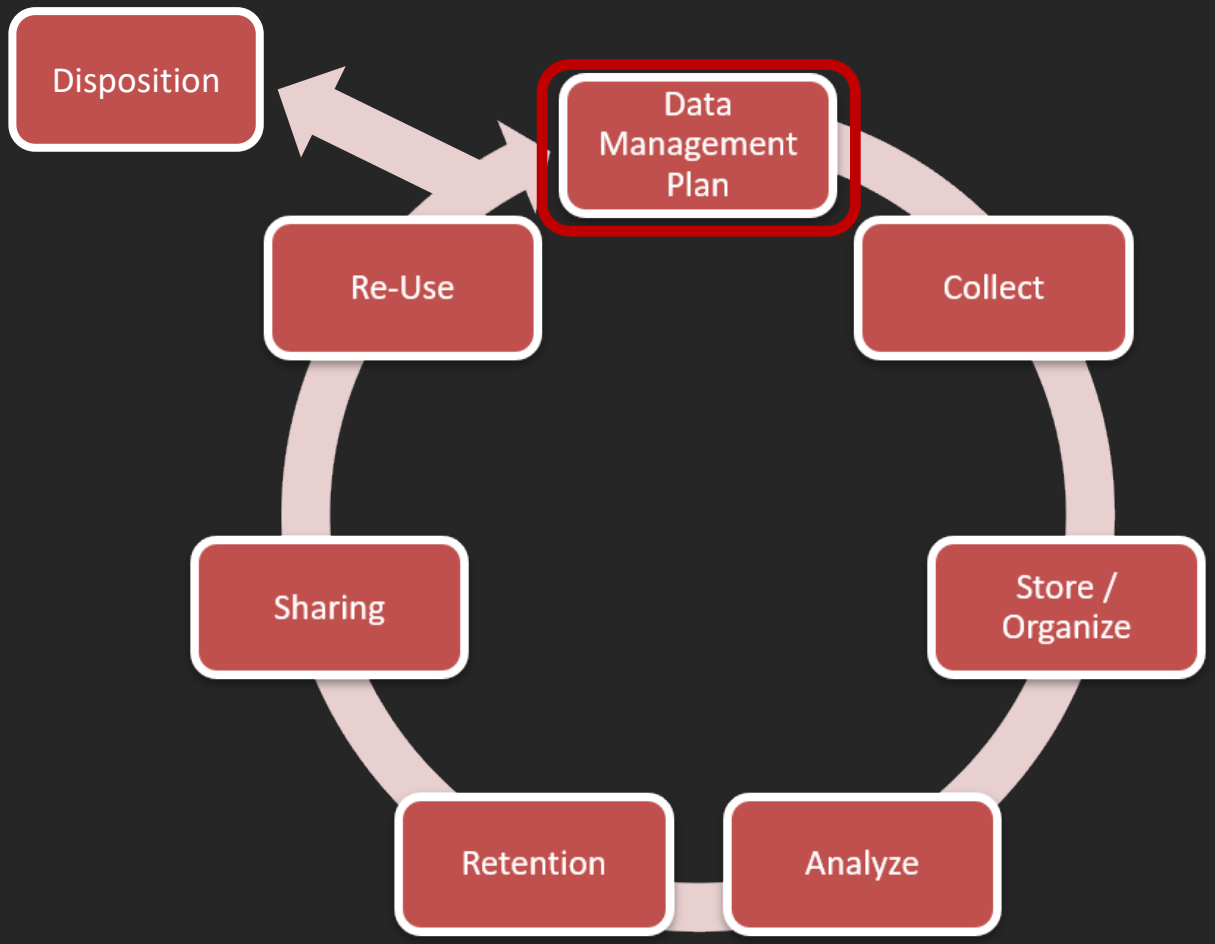
Data means recorded information, regardless of form or the media on which it may be recorded. The term includes computer software (computer programs, computer databases, and documentation thereof), and records of scientific or technical nature. The term does not include information incidental to award administration, such as financial, administrative, cost or pricing, or management information. In practice, scientific data include both intangible data (statistics, findings, conclusions, etc.) and tangible data. Tangible data include, but are not limited to notebooks, printouts, electronic storage, photographs, slides, negatives, films, scans, images, autoradiograms, electrophysiological recordings, gels, blots, spectra, cell lines, reagents, modified organisms, specimens, IRB consent forms, case report forms, drilling cores, collected organisms, and other materials that are relevant to the research project.

Where Do We Begin?

Topics

- Principles & Guidelines
- **Data Lifecycle**
- Data Quality & Integrity
- Case Study

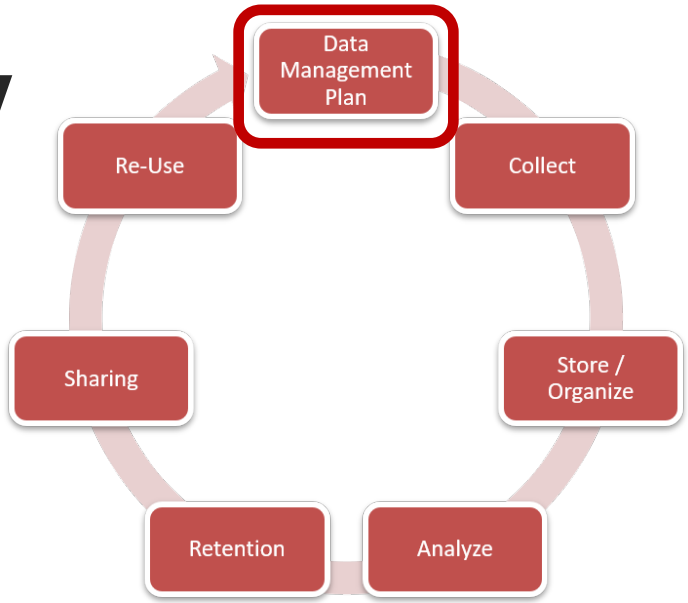
Data Lifecycle



Data Management

Definition…

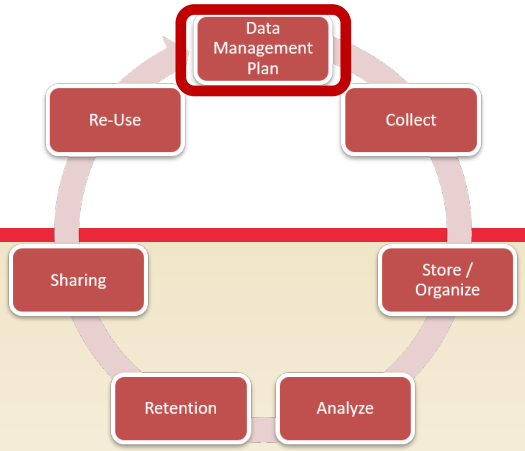
"Research data management is a term that describes the organization, storage, preservation, and sharing of data collected and used in a research project."



University of Pittsburgh | Library System

https://pitt.libguides.com/managedata

Data Management



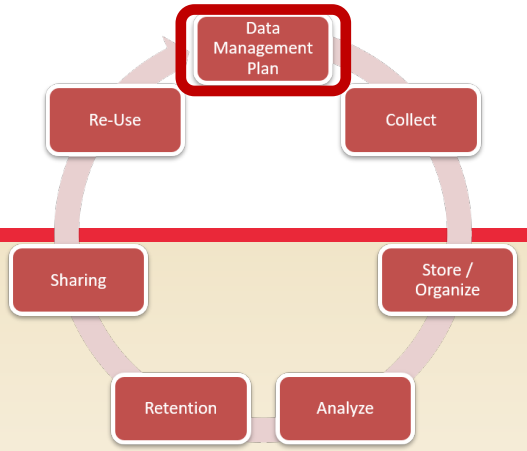
- Data is (are) a scholarly product
- Data are fragile and easily lost
- Growing research data requirements
- Good management helps prevent errors and increases the quality of your analysis
- Well-managed and accessible data allows others to validate and replicate findings
- Research data management facilitates sharing of research data and, when shared, data can lead to valuable discoveries by others outside of the original research team

University of Pittsburgh | Library System

utmb Health

https://pitt.libguides.com/managedata

Data Management Plan

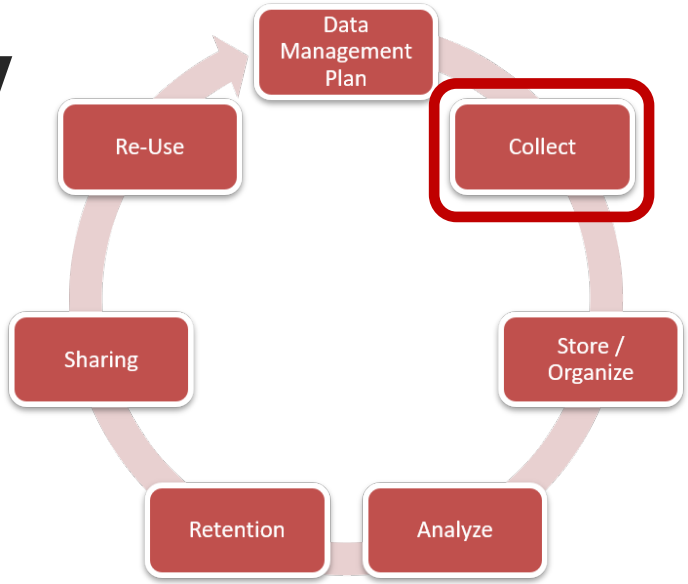


- Data Description / Types
- Data Standards for Format and Content
- Mechanisms for Access and Sharing (Provisions, Privacy Protection, Confidentiality, Security, Intellectual Property, etc.)
- Provisions for Data Reuse and Redistribution

Data Collection

Perform an inventory…

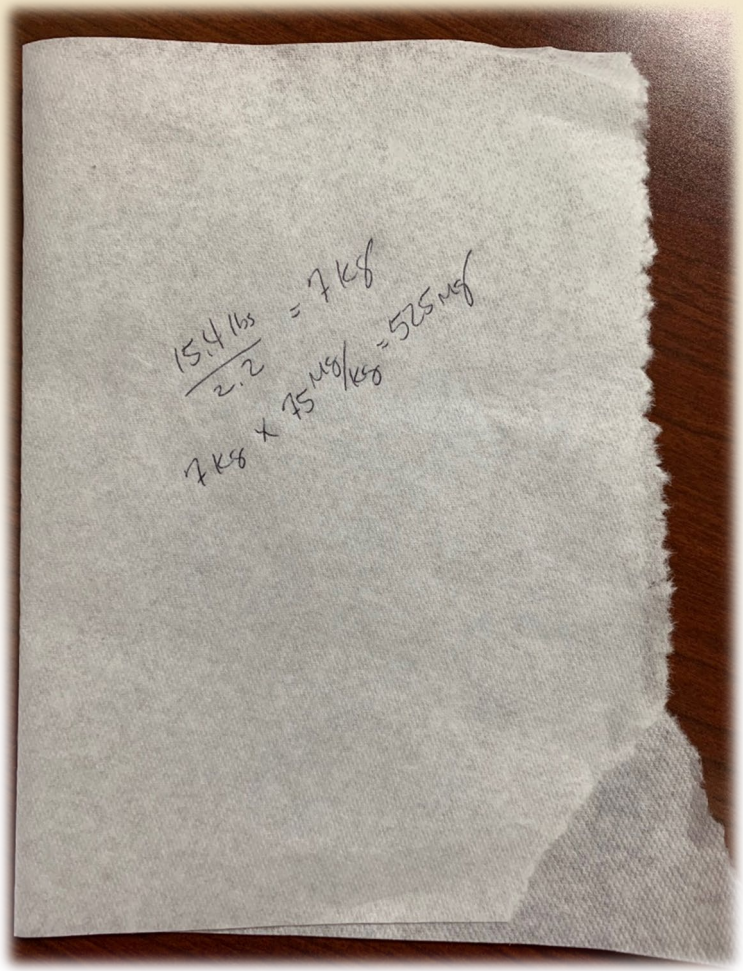
- Source (Raw) Data
- Types (observational, derived, etc.)
- Format (text, numeric, modeling, images, etc.)
- Quantity
- Regulatory (e.g., HIPAA)
- Proprietary
- Owner

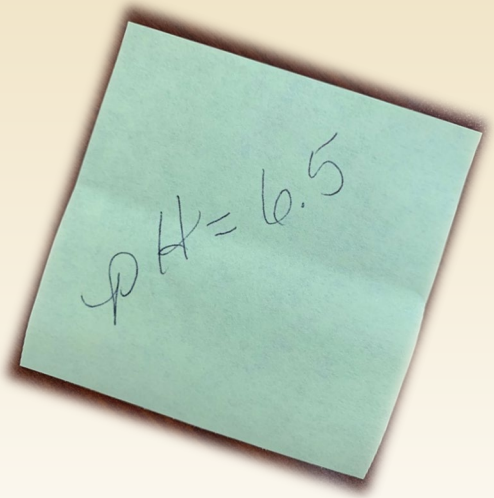


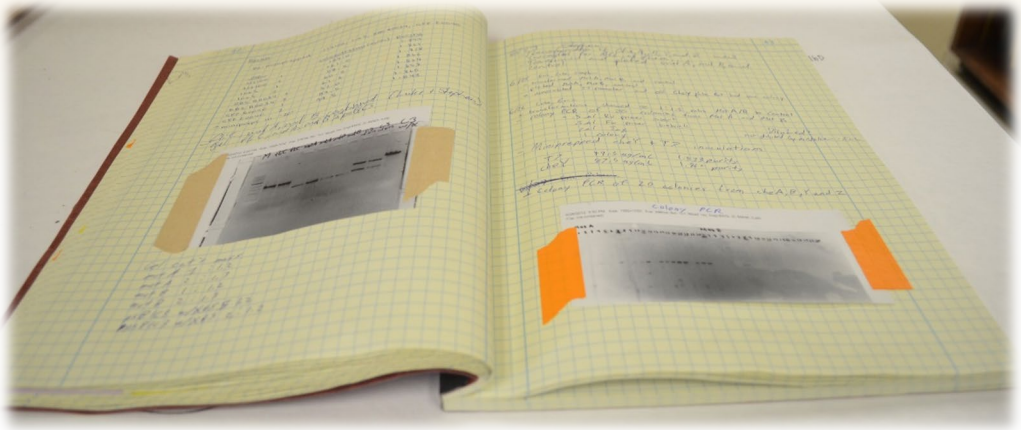
Source Data (Original)

Images courtesy of Bing

Source Data / Transcription

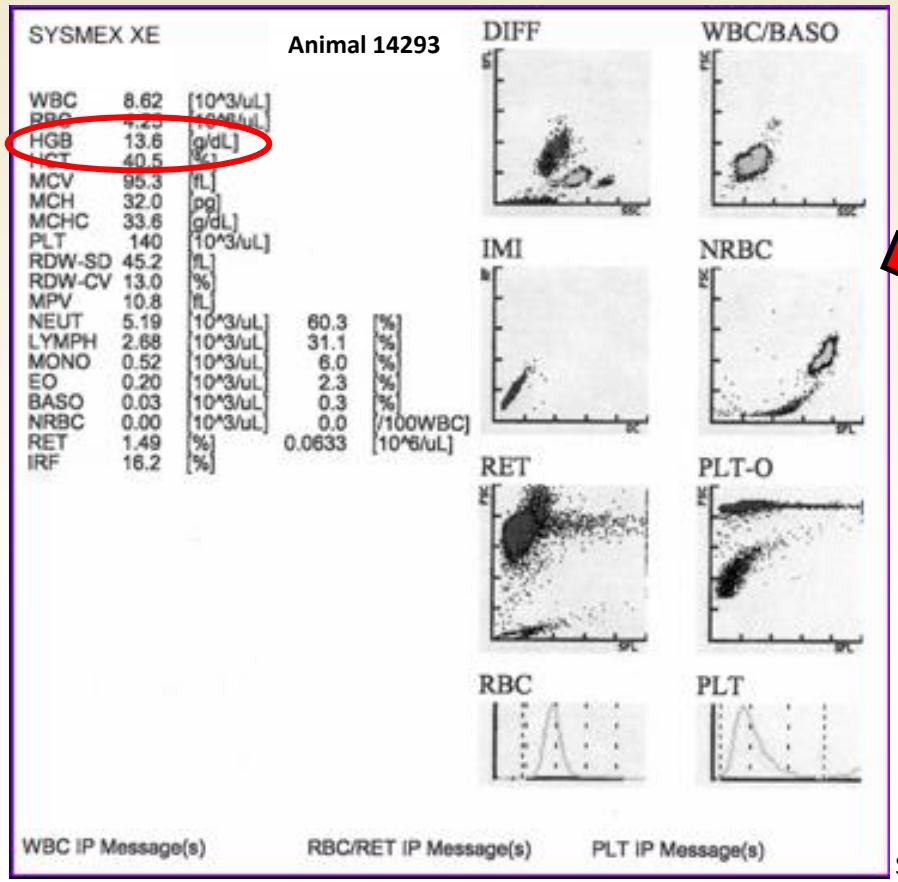






Transcription Errors

Hemoglobin Value



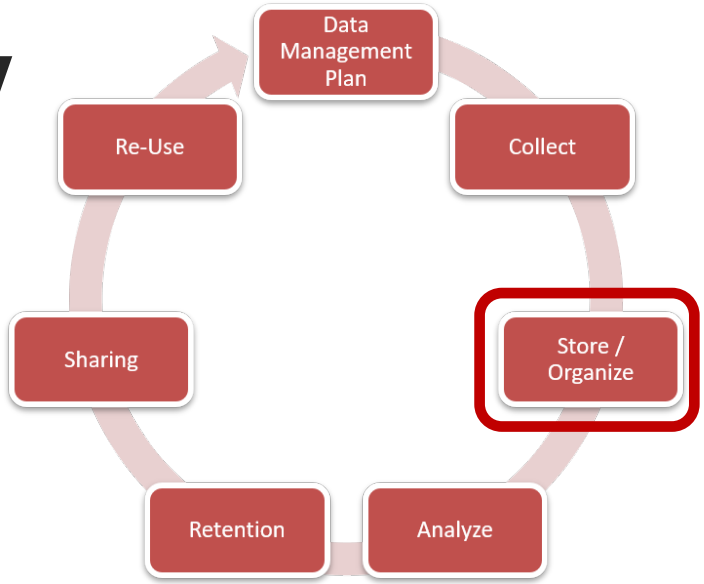
Animal	HGB
12938	12.2
14039	8.9
14293	3.6
14980	13.8
15209	12.5
15490	9.5
15560	14.0

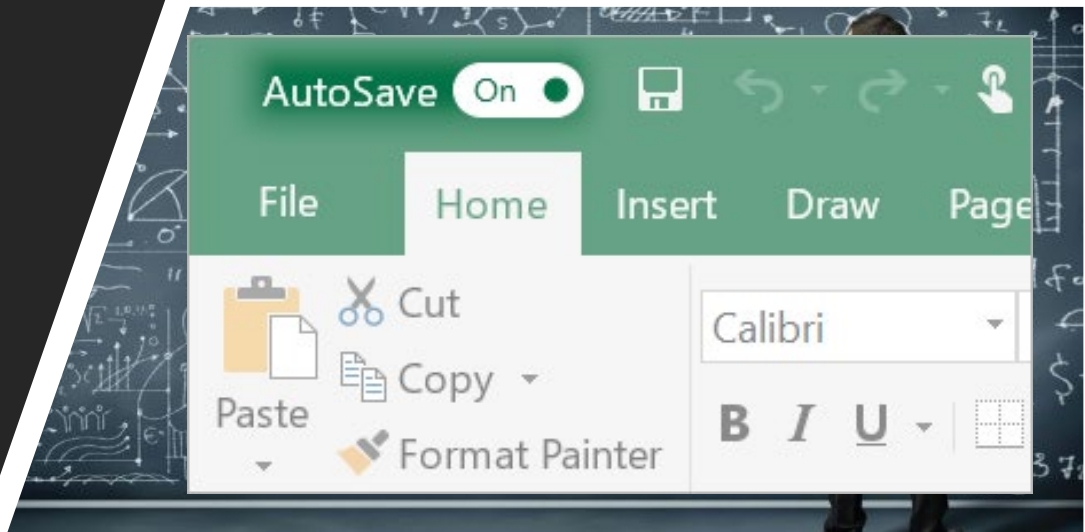
Source: Google Images

Organization and Storage

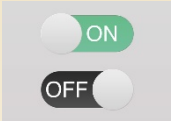
Things to think about prospectively…

- Location (physical / electronic)
- Transcription of source data
- Accessibility (limited)
- Security
- Change control
- Protection





Managing Electronic Data



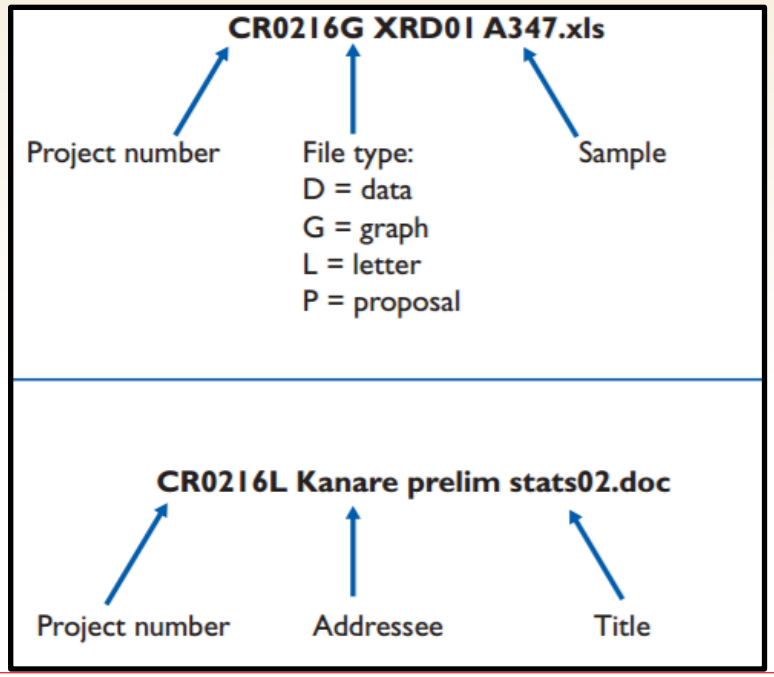
- Audit Trails / Meta Data
- Security / Encryption
- Software Compatibility
- Back-up
- Program Updates
  - Automatic
  - Impact to significant digits
- Data Migration
- Checksums

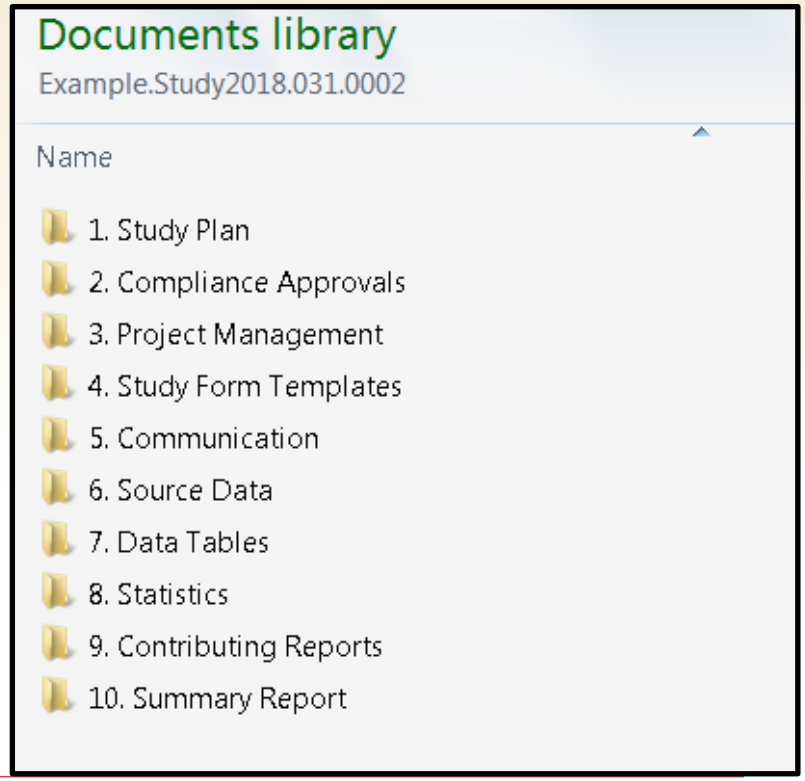
Electronic Documents / Files

Get organized!

Standard File Naming System

- Brief, descriptive, consistent, dated
- Plans for edits and changes





Data Analysis (Data Manipulation)

Prospective thinking…

- Methods to reduce transcription errors
- Define inclusion / exclusion criteria
- Develop statistical plan (study plan)
- Retain *methods* to allow for study reconstruction

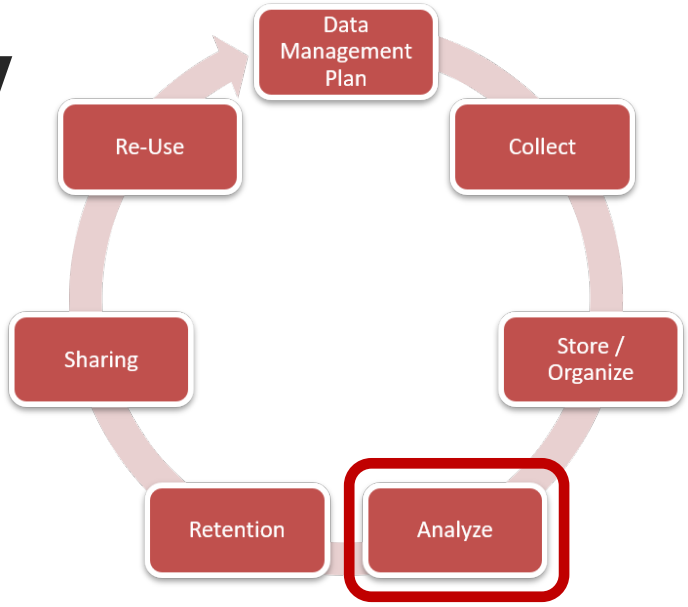


Image Manipulation

FIGURE 1. COMET ASSAY

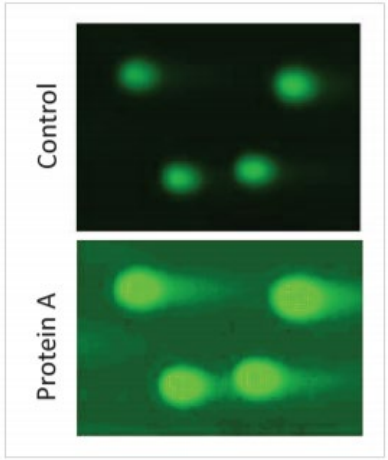


FIGURE 2. IMMUNOFLUORESCENCE COLOCALIZATION ASSAY

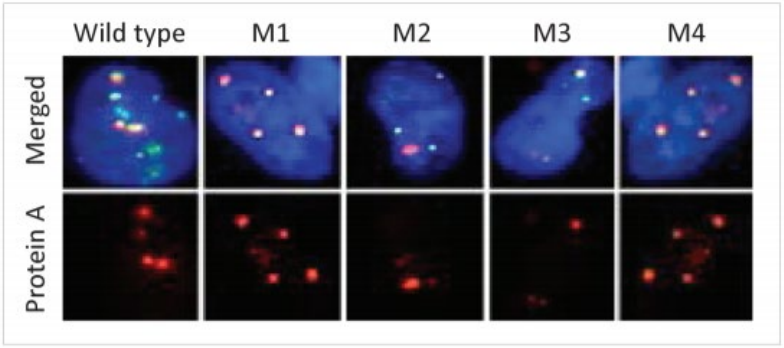


FIGURE 3. WESTERN BLOT

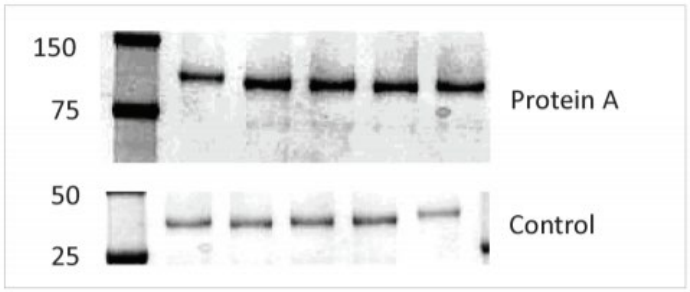
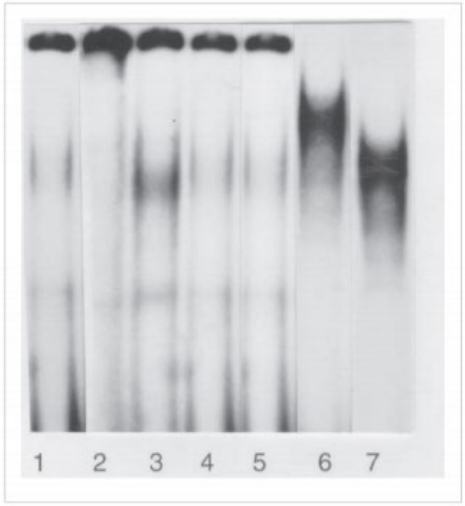


FIGURE 4. GEL SHIFT ASSAY



https://ori.hhs.gov/infographics

Image Manipulation

DISCUSSION

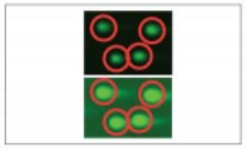


FIGURE 1. COMET ASSAY
The control image was cropped and relabeled as the image for Protein A. It was also intentionally lightened to make the "tails" appear longer.

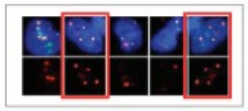


FIGURE 2. IMMUNOFLUORESCENCE COLOCALIZATION ASSAY
M1 and M4 are the same image but flipped vertically.

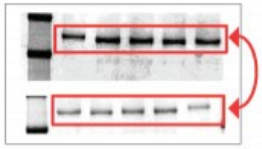


FIGURE 3. WESTERN BLOT
The top panel and bottom panel of Figure 3 are from the same source image. The Protein A blot image has been flipped horizontally and represented as the control blot image.

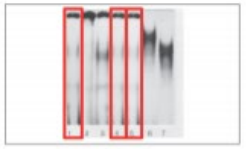


FIGURE 4. GEL SHIFT ASSAY
Lanes 1, 4, and 5 are from the same image source and were relabeled and reused to represent different experimental conditions.

FIGURE 1. COMET ASSAY

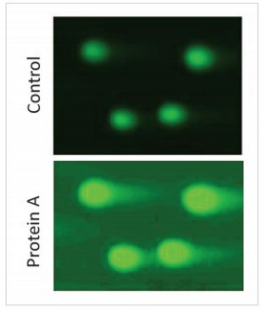


FIGURE 2. IMMUNOFLUORESCENCE COLOCALIZATION ASSAY

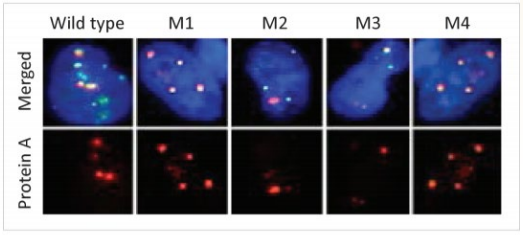


FIGURE 3. WESTERN BLOT

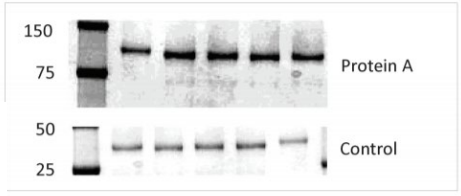
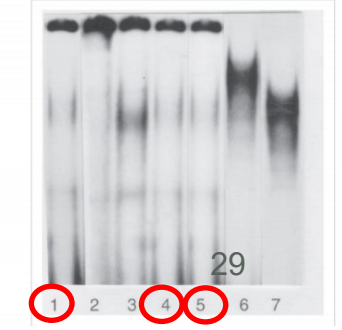


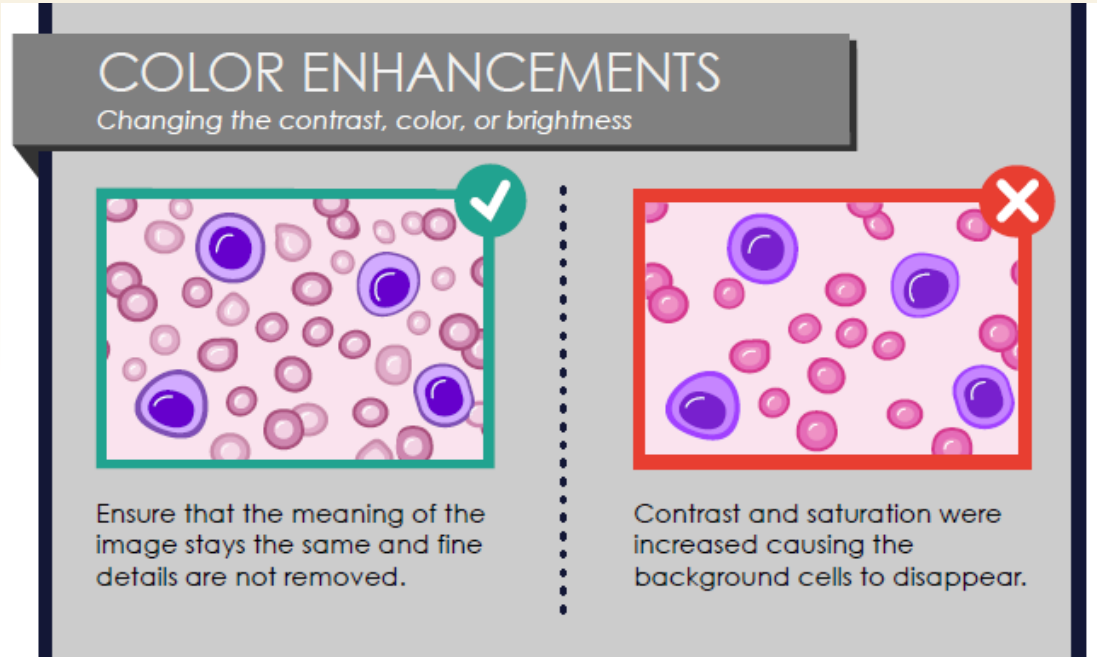
FIGURE 4. GEL SHIFT ASSAY



utmb Health

Image Manipulation

- Document all changes
- Retain unprocessed image
- Follow journal guidelines for permissible processing

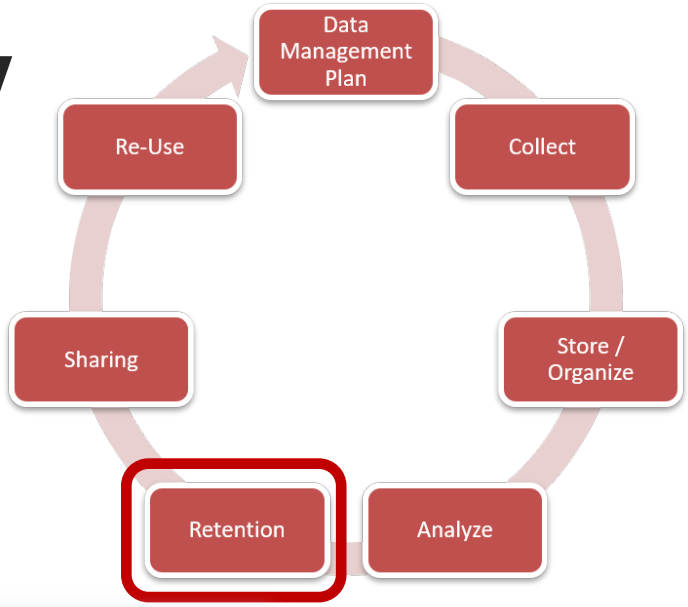


https://ori.hhs.gov/sites/default/files/2017-12/6_Image_Manipulation_scalable.pdf

Retention

Protect the data!

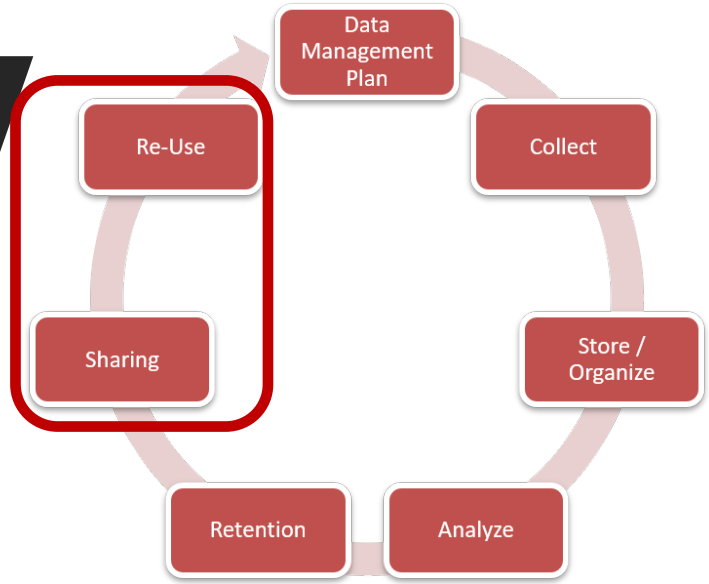
- Location
- Accessibility
- Security (encryption)
- Back-up
- System sustainability
- Data migration



Mechanisms / Conditions for Sharing

Define conditions…

- Email
- Online repositories
- Supplemental to publication
- Sharing agreements / plans
- Mixed media
- Conditions / Exclusions
- Instructions



Data Sharing Plan—Template

- What?
- Who?
- Where?
- When?
- How?

Example Plan addressing Key Elements for a Data Sharing Plan under NIH Extramural Support
(For questions, contact the NIH Office of Extramural Research (OER), Email Sharing@nih.gov)

Example Data Sharing Plan for FOA-XX-XXXX

What data that will be shared:
I will share phenotypic data associated with the collected samples by depositing these data at _____ which is an NIH-funded repository. Genotype data will be shared by depositing these data at _____. Additional data documentation and de-identified data will be deposited for sharing along with phenotypic data, which includes demographics, family history of XXXXXX disease, and diagnosis, consistent with applicable laws and regulations. I will comply with the NIH GWAS Policy and the funding IC's existing policies on sharing data on XXXXXX disease genetics to include secondary analysis of data resulting from a genome wide association study through the repository. Meta-analysis data and associated phenotypic data, along with data content, format, and organization, will be available at _____. Submitted data will confirm with relevant data and terminology standards.

Who will have access to the data:
I agree that data will be deposited and made available through _____ which is an NIH-funded repository, and that these data will be shared with investigators working under an institution with a Federal Wide Assurance (FWA) and could be used for secondary study purposes such as finding genes that contribute to process of XXXXXX. I agree that the names and Institutions of persons either given or denied access to the data, and the bases for such decisions, will be summarized in the annual progress report. Meta-analysis data and associated phenotypic data, along with data content, format, and organization, will be made available to investigators through _____.

Where will the data be available:
I agree to deposit and maintain the phenotypic data, and secondary analysis of data (if any) at _____, which is an NIH-funded repository and that the repository has data access policies and procedures consistent with NIH data sharing policies.

When will the data be shared:
I agree to deposit genetic outcome data into _____ repository as soon as possible but no later than within one year of the completion of the funded project period for the parent award or upon acceptance of the data for publication, or public disclosure of a submitted patent application, whichever is earlier.

How will researchers locate and access the data:
I agree that I will identify where the data will be available and how to access the data in any publications and presentations that I author or co-author about these data, as well as acknowledge the repository and funding source in any publications and presentations. As I will be using _____, which is an NIH-funded repository, this repository has policies and procedures in place that will provide data access to qualified researchers, fully consistent with NIH data sharing policies and applicable laws and regulations.

Rev. 20100831

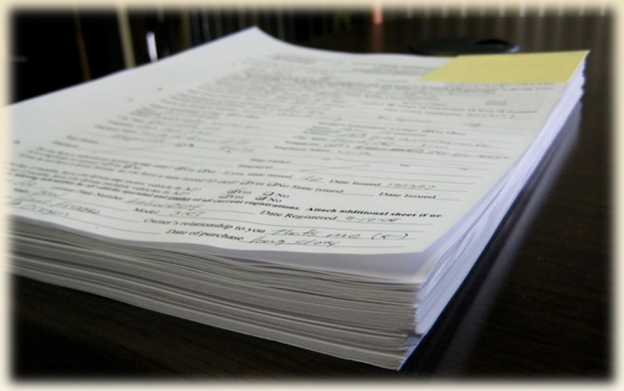
https://www.niaid.nih.gov/research/sample-data-sharing-plan

utmb Health

Topics

- Principles & Guidelines
- Data Lifecycle
- **Data Quality & Integrity**
- Case Study

Data Quality: ALCOA Principles



Data Quality

- **A**ttributable
- **L**egible
- **C**ontemporaneous
- **O**riginal
- **A**ccurate

Data Integrity

- Complete, Consistent, Enduring, Readily Available

Data Quality: Exercise

- **A**ttributable
- **L**egible
- **C**ontemporaneous
- **O**riginal
- **A**ccurate

Data Quality: Exercise

- ❑ Month / Day / Year
- ❑ Day / Month / Year
- ❑ Year / Month / Day

Electronic Laboratory Notebooks

Pros

- Project organization
- Collaboration
- Custom forms/fields to assure all data are captured
- Searchable
- Audit trail
- Data exportable

Cons

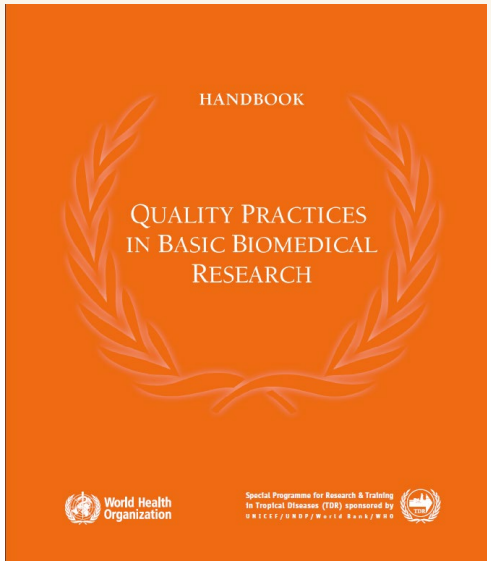
- Cost
- Sustainability ($)
- System administration
- Compatibility with other systems
- Software updates/data migration verification
- Discontinued (or support discontinued)

Reconstructability

Documentation:

- Full record of all information
- Enables study reconstruction
- Demonstrates what occurred at the time

"Without documentation the process is meaningless; essentially there has been no study."



https://www.who.int/tdr/publications/documents/quality_practices.pdf?ua=1

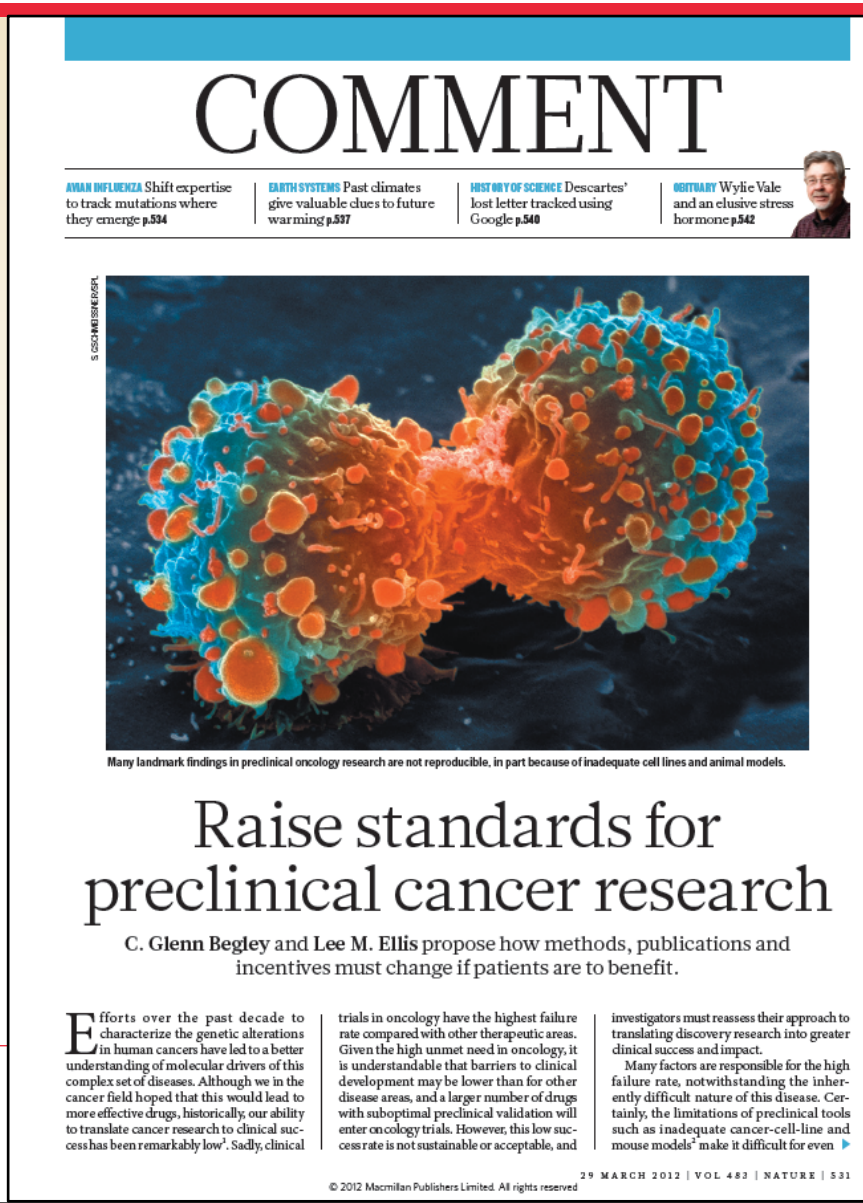
REPRISE: Data—Definition

Definition of Data

Data means recorded information, regardless of form or the media on which it may be recorded. The term includes computer software (computer programs, computer databases, and documentation thereof), and records of scientific or technical nature. The term does not include information incidental to award administration, such as financial, administrative, cost or pricing, or management information. In practice, scientific data include both intangible data (statistics, findings, conclusions, etc.) and tangible data. Tangible data include, but are not limited to notebooks, printouts, electronic storage, photographs, slides, negatives, films, scans, images, autoradiograms, electrophysiological recordings, gels, blots, spectra, cell lines, reagents, modified organisms, specimens, IRB consent forms, case report forms, drilling cores, collected organisms, and other materials that are relevant to the research project.

Reconstructability

- 53 landmark studies
- 6 confirmed (11%)
  - Controls
  - Reagents
  - Investigator bias
  - **Described complete data set**



COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542

SCIENCE PHOTO LIBRARY

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even ▶

utmb Health

Reconstructability—Communication

I am going to miss the March 3rd Friday call in…

March 2019						
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

Closing Thoughts…

Where to start…

- Get organized!
- Data stewardship throughout the data lifecycle
- Data management plan
- Prospectively plan
- Implement the ALCOA principles

Topics

- Principles & Guidelines
- Data Lifecycle
- Data Quality & Integrity
- **Case Study**

Case Study—Data Sharing

Take note of Room Number !

Your research study will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner-city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, you will be collecting identifying information.

Even though the final dataset will be stripped of identifiers prior to release for sharing, there remains the possibility of deductive disclosure of subjects with unusual characteristics.

Identify options (i.e., conditions) for sharing the data.

Case Study—Data Sharing

The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner-city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics.

Thus, we will make the data and associated documentation available to users only under a *data-sharing agreement* that provides for:

(1) a commitment to using the data only for research purposes and not to identify any individual participant;

(2) a commitment to securing the data using appropriate computer technology; and

(3) a commitment to destroying or returning the data after analyses are completed.

https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#ex



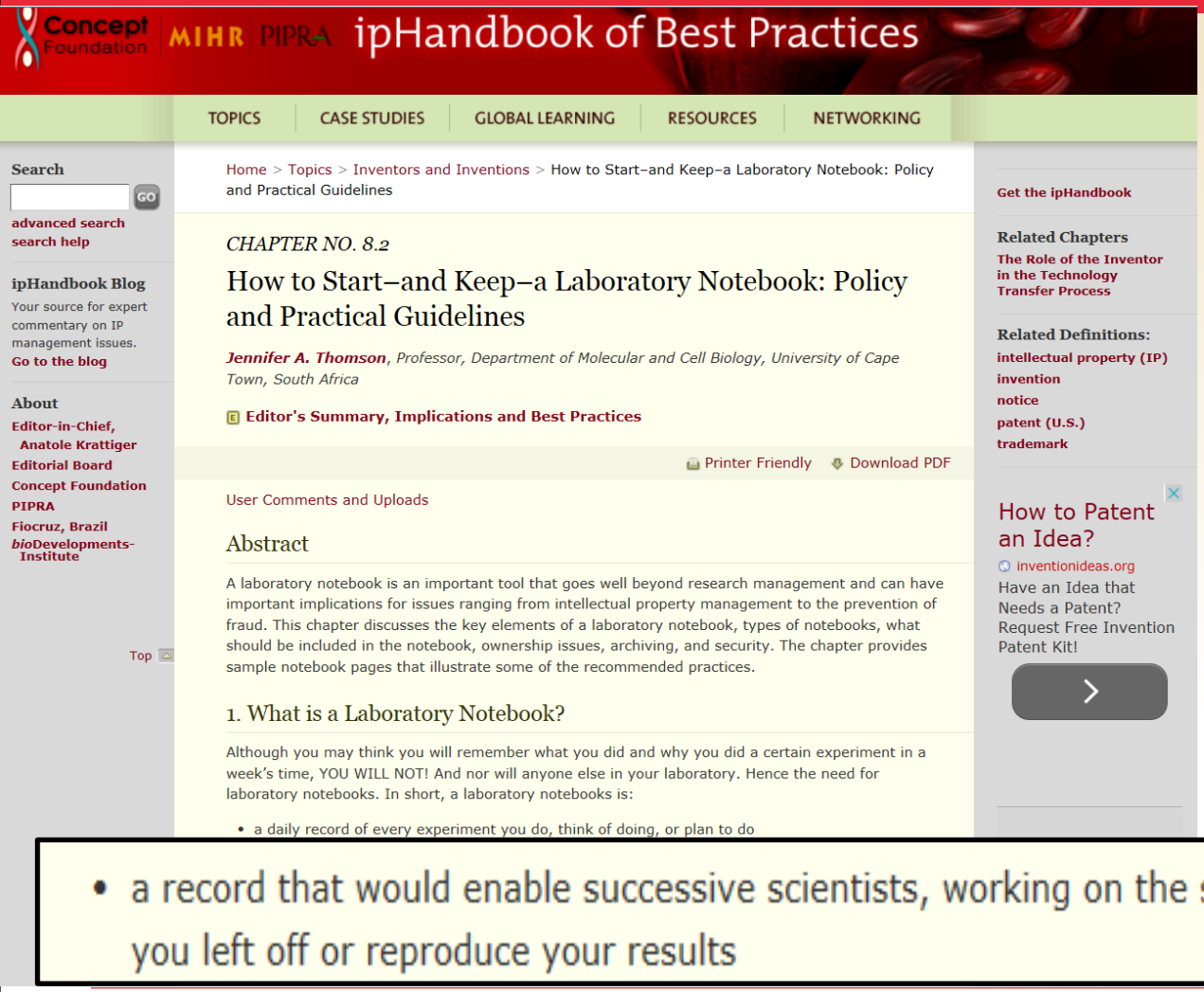
Photo: Suzanne & Walter Scott Chihuly Sanctuary, Univ. of Nebraska Medical Center

Thank you!

Melissa Eitzen, MT(ASCP), MS, RQAP-GLP
Director, Regulatory Operations
Institutional Office of Regulated Nonclinical Studies (ORNcS)
University of Texas Medical Branch at Galveston
2.810 Rebecca Sealy, Mail Stop 0184
301 University Boulevard
Galveston, TX 77555-0184
Phone: (409) 266-9422
Email: mmeitzen@utmb.edu
Web: http://www.utmb.edu/orncs

utmb Health

Reference: Laboratory Notebooks



- a record that would enable successive scientists, working on the same project, to pick up where you left off or reproduce your results

http://www.iphandbook.org/handbook/ch08/p02/