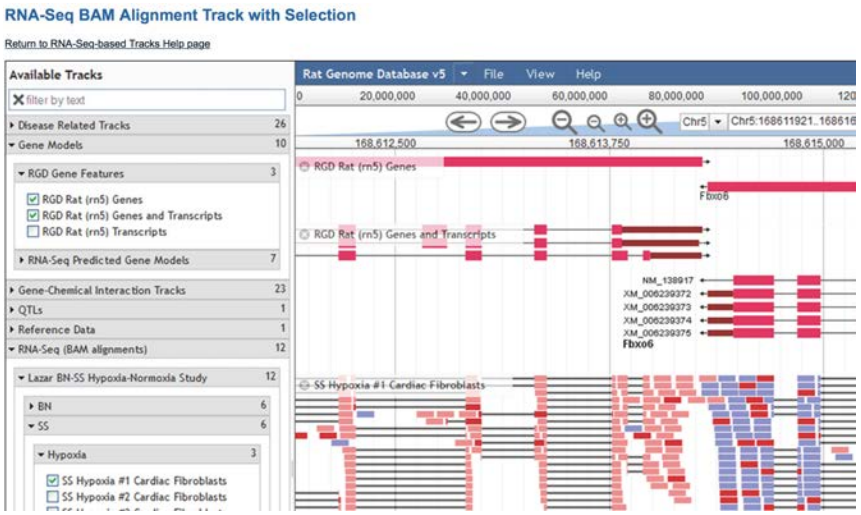
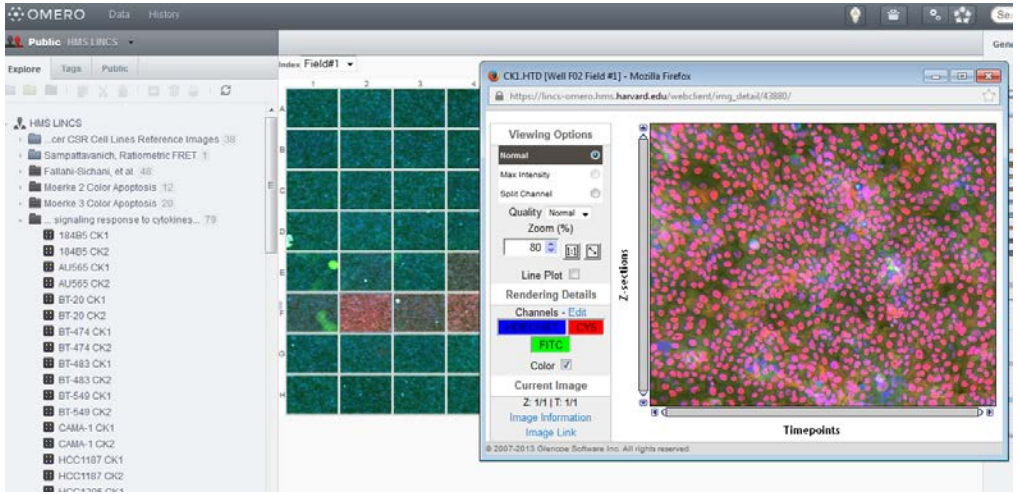
Data Management Resources for Biomedical Research

Caroline Shamu, Ph.D.
Assistant Professor, Harvard Medical School





A new resource for information about data management relevant to biomedical research

http://datamanagement.hms.harvard.edu

Harvard Biomedical Data Management

Best practices & support services for research data lifecycles

About ▾ | Best Practices ▾ | Planning ▾ | Data Repositories ▾ | Storage ▾ | Policies ▾ | Harvard Open Access

Data Management

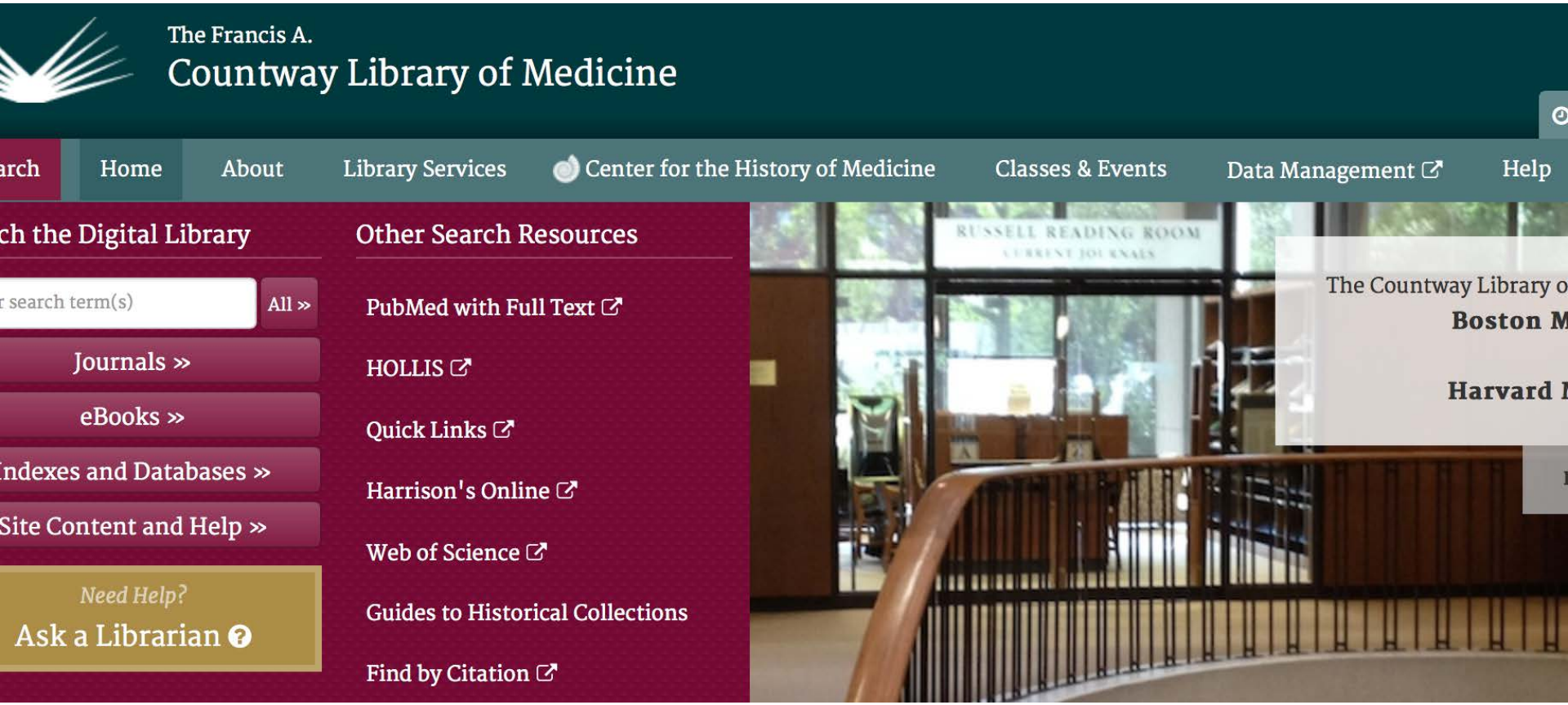
Data Management is the process of providing the appropriate labeling, storage, and access for data at all stages of a research project. We recognize that best practices for each of these aspects of data management can and often do change over time, and are different for different stages in the data lifecycle.

Early and attentive management at each step of the data lifecycle will ensure the discoverability and longevity of your research.

Ask us your biomedical research data management questions!

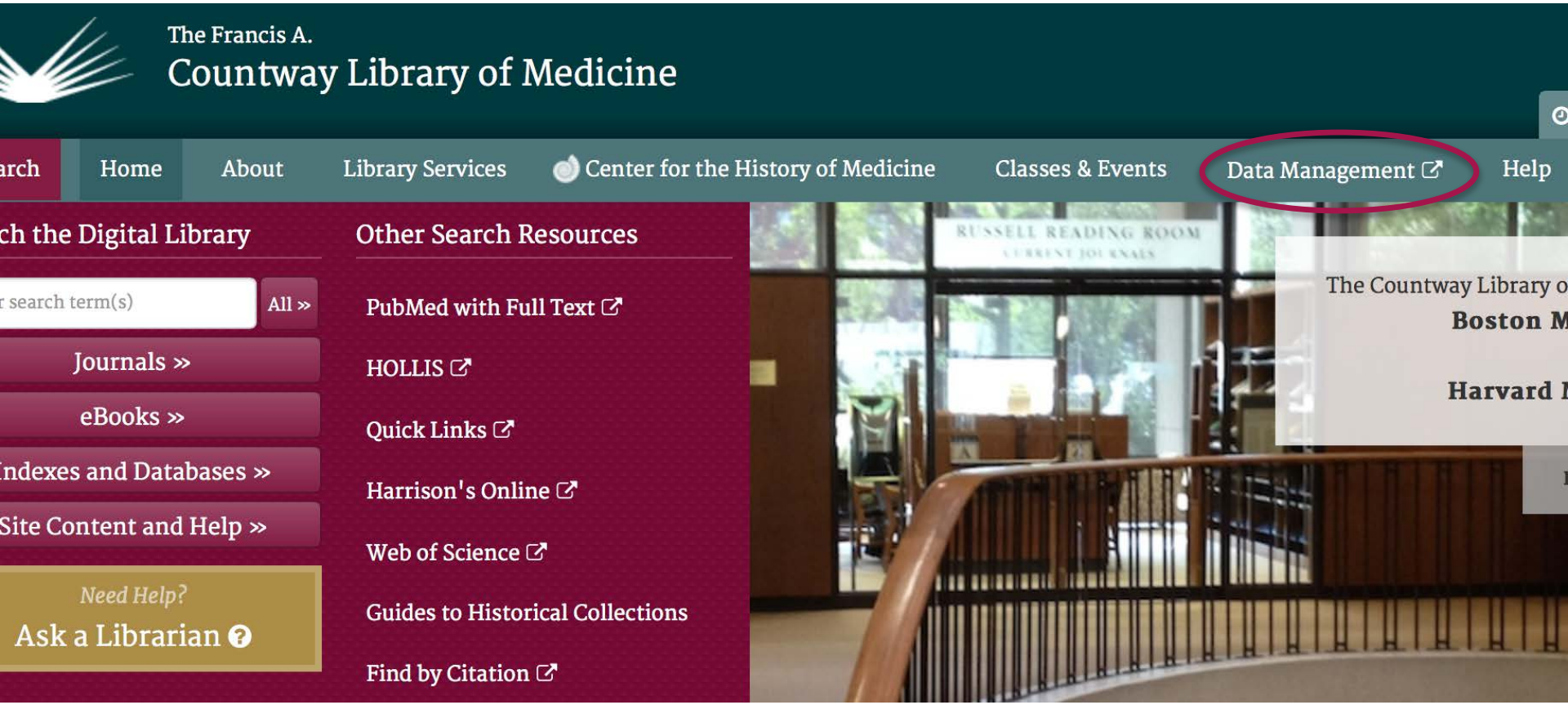
HARVARD MEDICAL SCHOOL | The Francis A. Countway Library of Medicine

Direct link from the Countway Library homepage



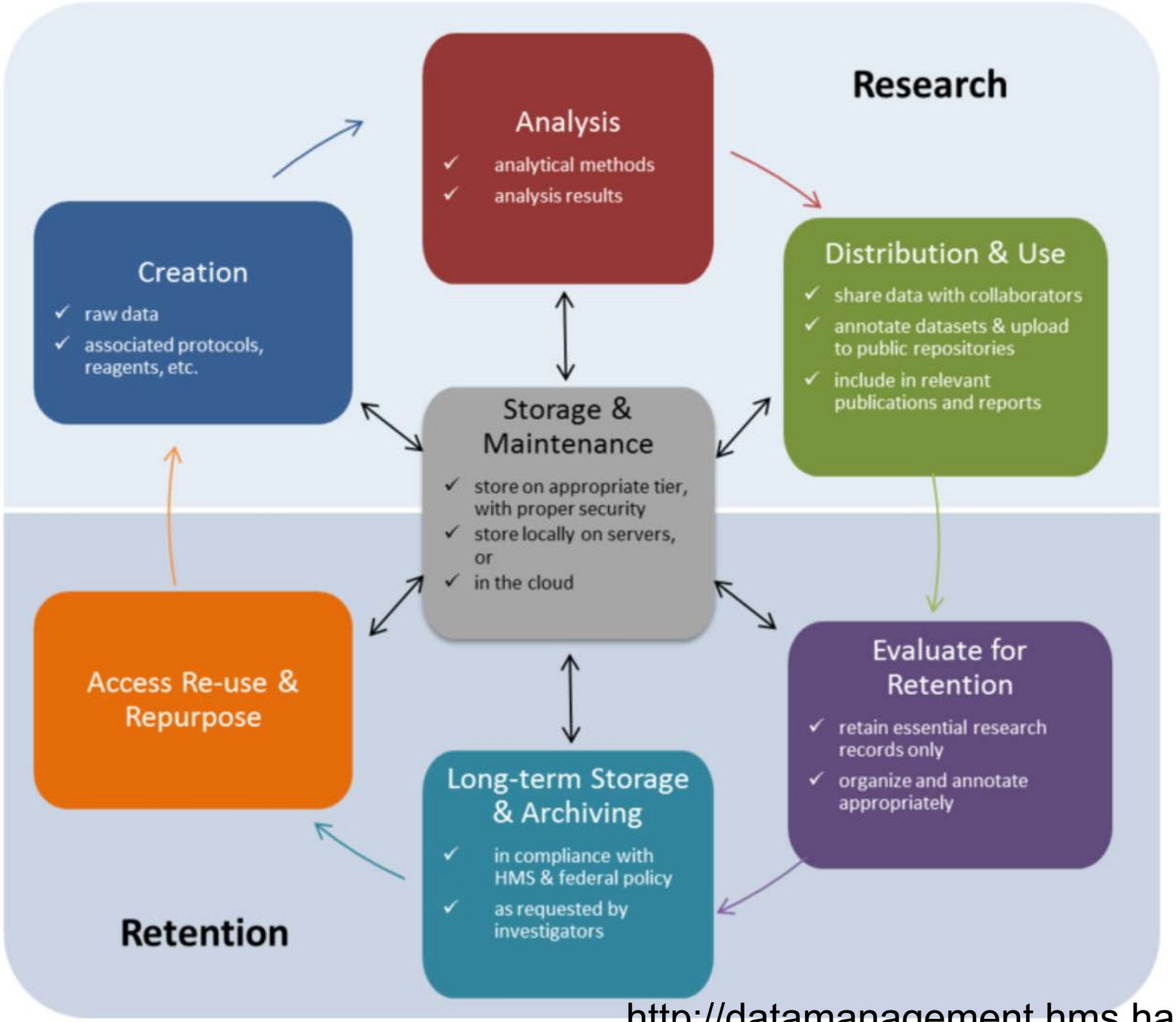
https://www.countway.harvard.edu

Direct link from the Countway Library homepage



https://www.countway.harvard.edu

Data lifecycle for biomedical research



http://datamanagement.hms.harvard.edu

Focus

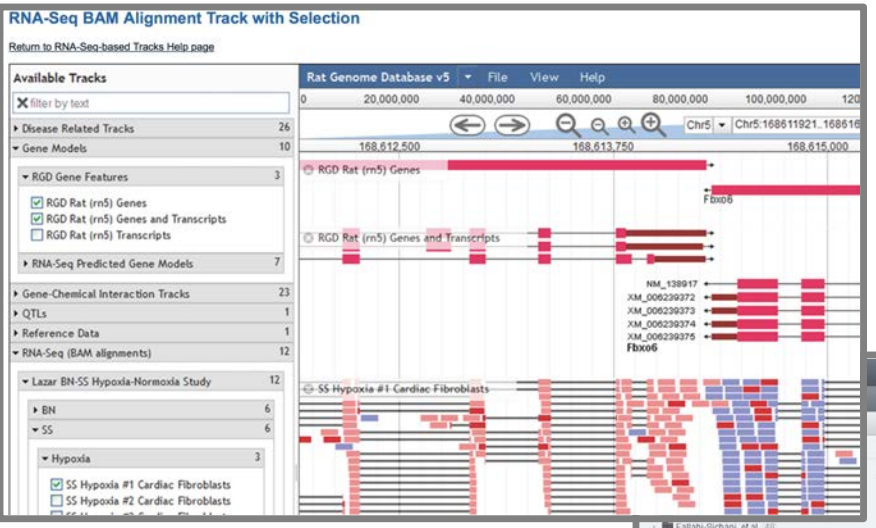
- Documenting biomedical experiments and data analyses: relevant metadata

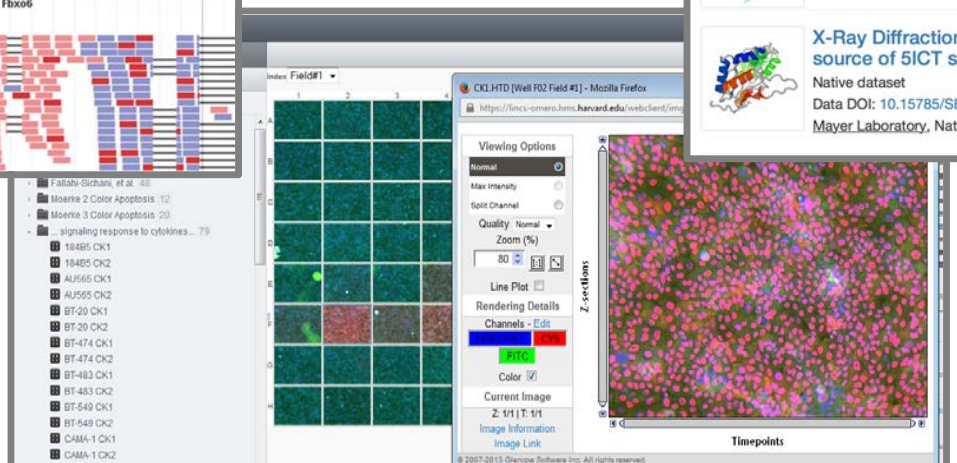
- Data Sharing

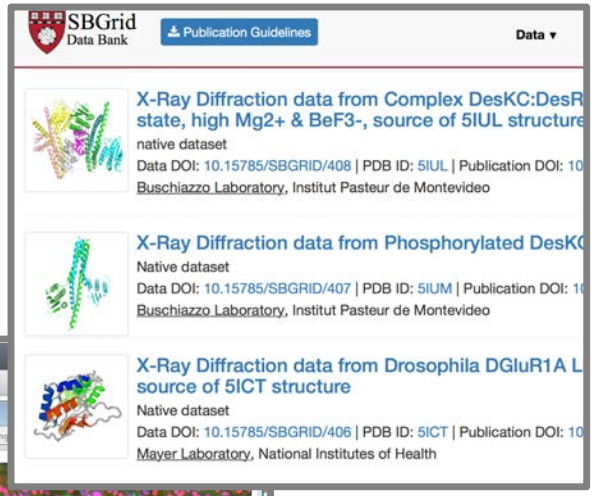
Examples of biomedical research data

Experimental data

- Genomic —DNA sequences, RNA sequences
- Images—of cells, tissues, organisms via light microscopy or EM
- Structural biology datasets – crystallography, NMR, cryoEM



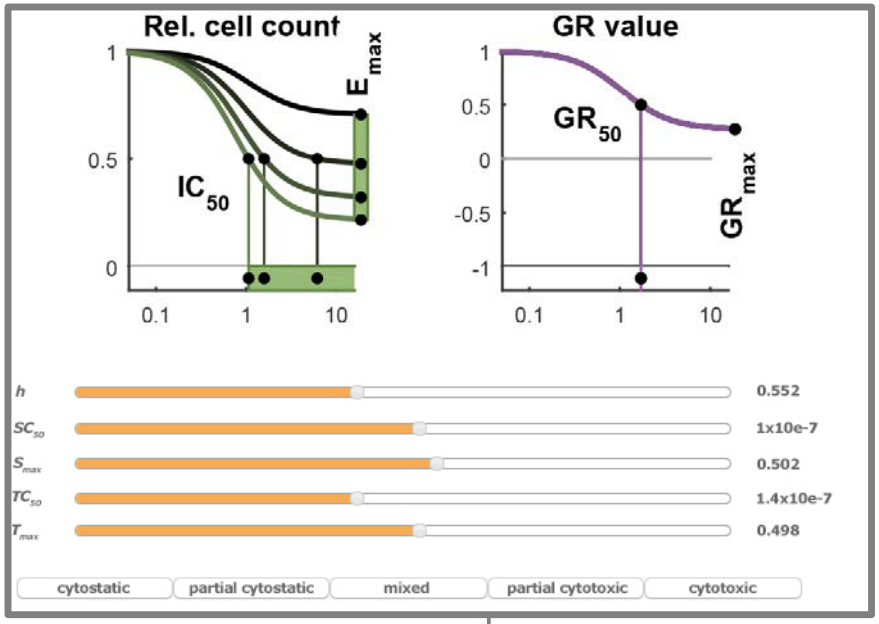


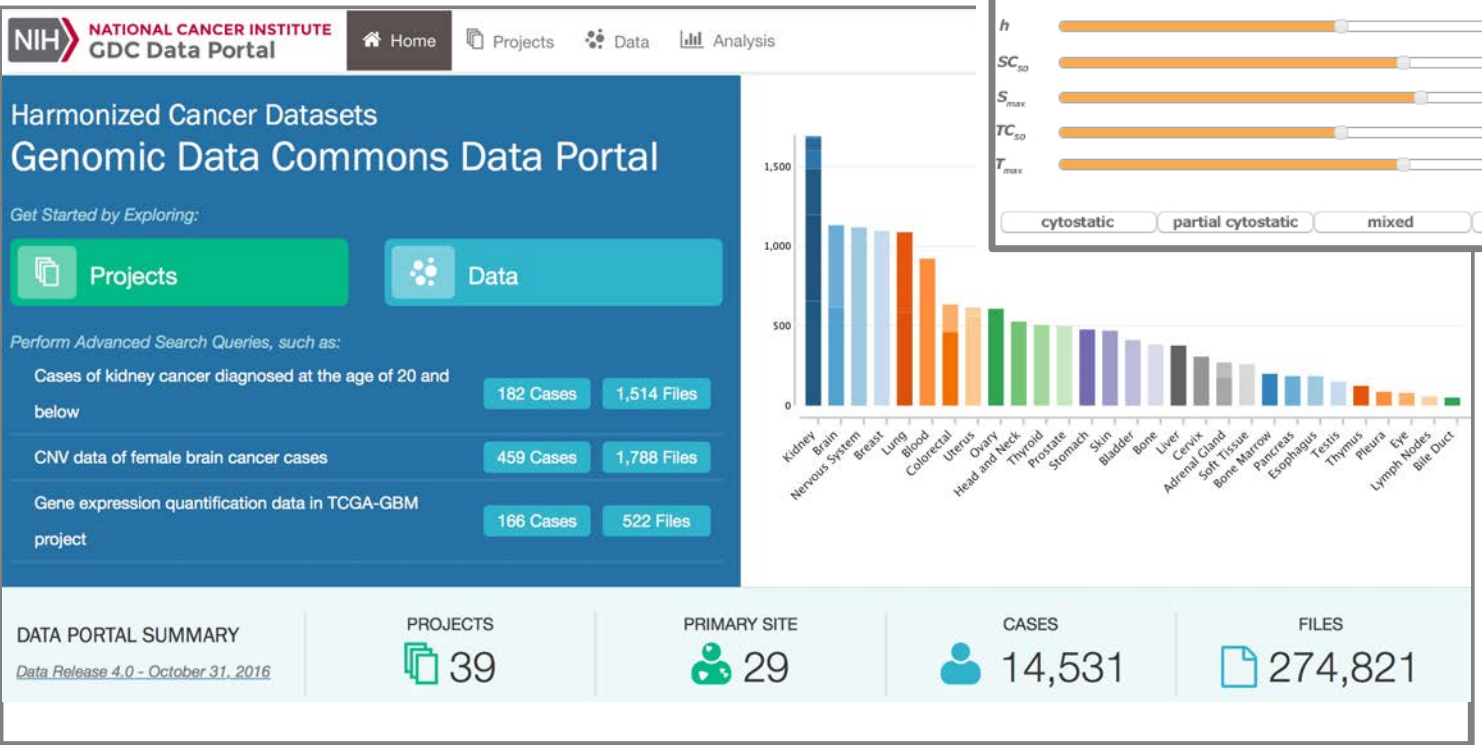


Examples of biomedical research data

Results of data analysis

- Individual experiments
- Large compiled datasets





Some challenges in managing biomedical research data

- Quantity
  - -large file sizes
  - -some datasets comprised of many smaller files
- Many different file types, many different experiment/analysis workflows
- Moving data
- Appropriate data storage solutions
  - -local vs. cloud
  - -short term vs. long term
  - -affordable
  - -appropriate security levels, especially for patient data
  - -public/non-public repositories
- Annotating and curating datasets
- Sharing data—finding appropriate public repositories

Motivations for robust data annotation workflows

1. *Enable continuity of research projects*

   -Easier for data producer, PI, and collaborators to find data.

2. *Promote data reproducibility*

   -Well-documented reagents, protocols, & datasets are available.
   -Granting agencies and journals are increasingly requiring this documentation.

3. *Facilitate data sharing and re-use*

   -Increases visibility of research.

4. *Reduce research and data storage costs*

   -Minimize storage of duplicate files, increase ability to re-use data.

Metadata are important!

Metadata for biomedical research may include:

Reagent Metadata: Information about the clinical samples, biological reagents (e.g. cell lines, antibodies, siRNAs), chemical reagents (e.g. drugs), etc. used to generate the data.

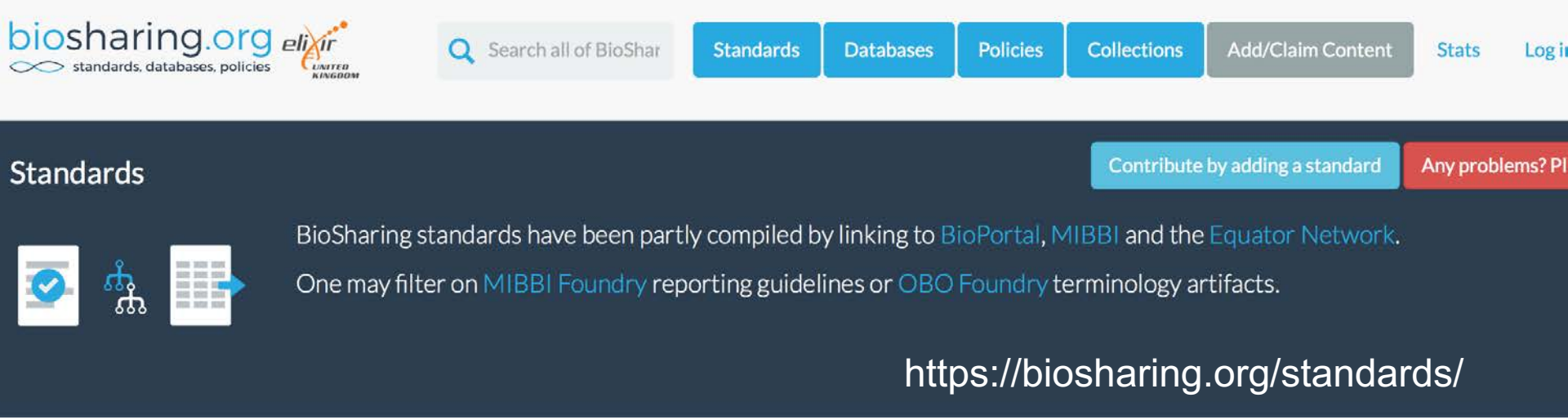
Technical Metadata: Information automatically generated by research instruments and associated software.

Experimental Metadata: Information about the experimental conditions (e.g. assay type, time points), the experimental protocol, and the equipment used to generate the data.

Analytical Metadata: Information about data analysis methods including software name and version, quality control parameters, and output file type details.

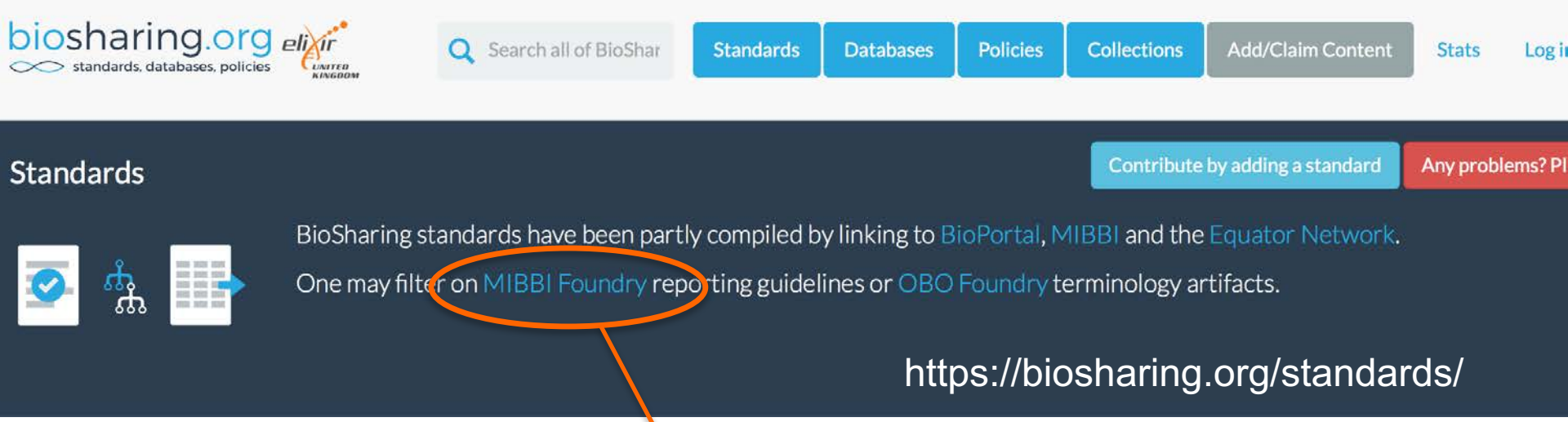
Dataset Level Metadata: Information about the objectives of the research project, participating investigators, relevant publications, and funding sources.

Metadata: Various schema exist or are being developed



https://biosharing.org/standards/

Metadata: Various schema exist or are being developed



Minimum Information for Biological and Biomedical Investigations

Metadata: Various schema exist or are being developed

biosharing.org elixir
standards, databases, policies

Search all of BioShar | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in o

standards > reporting guideline > bsg-s000177

Ask Question | Claim

R Minimum Information About a Microarray Experiment

Abbreviation:MIAME

mibbi REPORTING GUIDELINE

General Information

MIAME is intended to specify all the information necessary for an unambiguous interpretation of a microarray experiment, and potentially to reproduce it. MIAME defines the content but not the format for this information.

Homepage http://www.fged.org/projects/miame/

Prioritize collection of *reagent* metadata

- Reagent information is key when comparing/integrating datasets:

  -Are results from the same or different cell line being compared?
  -Were they treated with the same drug?
  -Is the same protein or subcellular feature being monitored?

Data Standards

LINCS Phase II Extended Metadata Standards

In LINCS Production Phase II, the **LINCS Data Working Group (DWG) is currently developing extended metadata specifications** describing LINCS reagents, assays and experiments. Annotations for the perturbagens (small molecules, siRNA, growth factors and other ligands), cells, and some elements of experimental metadata should be common between all LINCS Centers. This will facilitate development of data analysis, formatting, and visualization strategies by LINCS investigators, and also the development of databases and data repositories in which to store and share LINCS data.

Current Versions of Standards Released: 5-13-2016

- Antibody reagents
- Cell lines
- Differentiated cells
- Embryonic stem cells
- iPSCs
- Nucleic acid reagents
- Other reagents
- Primary cells
- Proteins
- Small molecules

http://www.lincsproject.org/LINCS/data/standards

NIH LINCS
PROGRAM

LIBRARY OF INTEGRATED NETWORK-BASED CELLULAR SIGNATURES

Home About Centers Data Tools Community Publications News

Data Standards

LINCS Phase II Extended Metadata Standards

In LINCS Production Phase II, the **LINCS Data Working Group (DWG) is currently developing extended metadata specifications** describing LINCS reagents, assays and experiments. Annotations for the perturbagens (small molecules, siRNA, growth factors and other ligands), cells, and some elements of experimental metadata should be common between all LINCS Centers. This will facilitate development of data analysis, formatting, and visualization strategies by LINCS investigators, and also the development of databases and data repositories in which to store and share LINCS data.

Overview
Releases
Release Policy
Standards

Current Versions of Standards Released: 5-13-2016

- Antibody reagents
- Cell lines
- Differentiated cells
- Embryonic stem cells
- iPSCs
- Nucleic acid reagents
- Other reagents
- Primary cells
- Proteins
- Small molecules

Goal:
-Minimal set of descriptors
-Not onerous to implement

Developed as a collaboration between experimentalists and database/informatics experts

http://www.lincsproject.org/LINCS/data/standards

Prioritize collection of *reagent* metadata

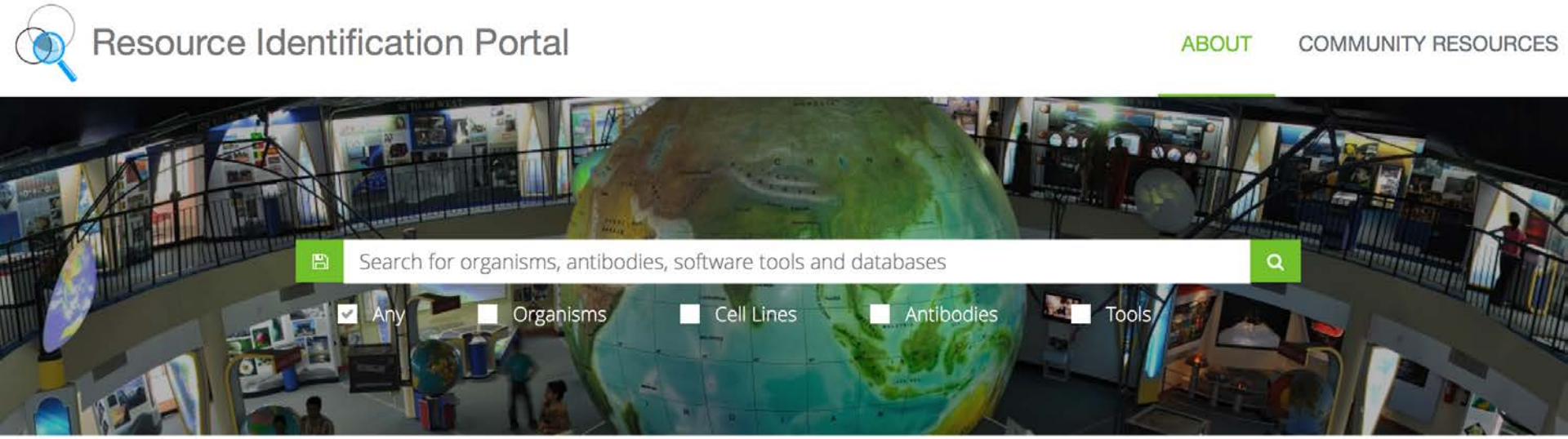
- Reagent information is key when comparing/integrating datasets:

  -Are results from the same or different cell line being compared?
  -Were they treated with the same drug?
  -Is the same protein or subcellular feature being monitored?

- It's easiest to track reagent metadata from the beginning of a project.

- Well-maintained reagent registries facilitate reagent identification.
  -local databases (e.g. *Reagent Tracker* at the Lab for Systems Pharmacology)
  -public databases (e.g. *Resource Identification Portal*)

Reagent Registration through the Resource Identification Initiative



Welcome

This is the Resource Identification Portal, supporting NIH's new guidelines for Rigor and Transparency in biomedical publications. Authors are instructed to authenticate key biological resources: Antibodies, Model Organisms, and Tools (software, databases, services), by finding or generating stable unique identifiers. We appreciate your patience and any feedback. If you experience any difficulties, please contact us at rii-help at scicrunch.org or just click on 'report an issue ' below and we will help you obtain the appropriate identifiers.

- Registry and information aggregator for research resources
- So far, it includes organisms, antibodies, software, databases, and other tools
- Assigns unique RRIDs (Research Resource Identifiers) to each item for inclusion in publications

Antibody: RRID:AB_2140114
Organism: RRID:MGI_MGI:3840442
Tool: RRID:nif-0000-00280

- Also provides standardized text for citation of each resource

https://scicrunch.org/resources

Reagent Registration through the Resource Identification Initiative

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Abstract

Send to:

J Comp Neurol. 2015 Feb 3. doi: 10.1002/cne.23755. [Epub ahead of print]

The gyri of the octopus vertical lobe have distinct neurochemical identities.

Shigeno S[1], Ragsdale CW.

Author information

Abstract

The cephalopod vertical lobe is the largest learning and memory structure known in invertebrate nervous systems. It is part of the visual learning circuit of central brain, which also includes the superior frontal and subvertical lobes. Despite the well-established functional importance of this system, little is known about neuropil organization of these structures and there is to date no evidence that the five longitudinal gyri of the vertical lobe, perhaps the most distinctive morphological feature of the octopus brain, differ in their connections or molecular identities. We studied the histochemical organization of these structures in hatchling and adult Octopus bimaculoides brains with immunostaining for serotonin, octopus gonadotropin-releasing hormone (oGNRH) and octopressin-neurophysin (OP-NP). Our major finding is that the five lobules forming the vertical lobe gyri have distinct neurochemical signatures. This is most prominent in the hatchling brain, where the median and medio-lateral lobules are enriched in OP-NP fibers, the lateral lobule is marked by oGNRH innervation, and serotonin immunostaining labels the median and lateral lobules heavily. A major source of input to the vertical lobe is the superior frontal lobe, which is dominated by a neuropil of interweaving fiber bundles. We have found that this neuropil also has an intrinsic neurochemical organization: it is partitioned into territories alternately enriched or impoverished in oGNRH-containing fascicles. Our findings establish that the constituent lobes of the octopus superior frontal-vertical system have an intricate internal anatomy, one likely to reflect the presence of functional subsystems within cephalopod learning circuitry.

KEYWORDS: GNRH; Octopus bimaculoides; RRID: AB_10562367; RRID: AB_141372; RRID: AB_143165; RRID: AB_2341084; RRID: AB_2341085; RRID: AB_477522; RRID: AB_477585; RRID:SciRes_000111; RRID:SciRes_000161; RRID:SciRes_000164; RRID:SciRes_000165; brain; cephalopod; octopressin; serotonin

Sharing Data

- Ensures scientific process is transparent and reproducible.
- Promotes re-use of data (reduces repeated work).
- Increases citation of publications.
- Data sharing required for many NIH projects and by a growing number of journals.

Challenges

- There are many biological data types for which no data standards or public repositories have been established.
- Annotating datasets is very time-consuming—more automated systems and robust reagent registries are needed.
- Adequate resources for data curation are usually not provided by funding agencies.
- Incentives are needed to support robust data curation and dataset publication.

Data Sharing

EDITORIAL

DATA SHARING

Reproducibility will only come with data liberation

IMPROVEMENTS IN HUMAN HEALTH—MADE POSSIBLE BY, FOR EXAMPLE, INNOVATIVE new medicines—are highly dependent on an ecosystem in which academic laboratories publish provocative proof-of-concept studies and in which industrial scientists use these studies to de-

www.ScienceTranslationalMedicine.org 18 May 2016 Vol 8 Issue 339 339ed7

Share Data

- Deposit published datasets and software tools into established, public repositories whenever possible!
  - *e.g.* NCBI TraceArchive or NCBI SRA for DNA and RNA sequencing data
  - *e.g.* PubChem, GenomeRNAi for HTS data
  - figshare, Dryad if necessary
  - *e.g.* github for code