

# **Academic Libraries and Research Data Management: Trends and Challenges in the U.S.**

---



**Dr. Hsin-liang (Oliver) Chen**  
**Dean of the Library**  
**Missouri University of Science and  
Technology**

# Agenda

- Top Research Universities Ranked by Research Expenditures
- Libraries' Research Data Services at those Universities
- Challenges
- Examples of Data Portals and Library Research Data Services
- Education and Job Duties of Data Management Professionals
- Data Resources at University of California San Francisco

# Top Research Universities Ranked by R&D Expenditures

Ranking	2017 Rankings by total R&D expenditures	2016 Rankings by total federal obligations
1	Johns Hopkins University	Johns Hopkins University
2	University of Michigan	University of Michigan
3	University of California, San Francisco	University of Washington
4	University of Pennsylvania	University of California, San Francisco
5	University of Washington	University of California, San Diego
6	University of Wisconsin-Madison	University of Pennsylvania
7	University of California, San Diego	Columbia University
8	Duke University	Stanford University
9	Harvard University	University of Pittsburgh
10	Stanford University	University of California, Los Angeles

(NSF, <https://ncesdata.nsf.gov/profiles/site?method=rankingBySource&ds=herd>)

# Characteristics of those Universities

**13 universities: 8 public and 5 private**

**7 universities listed in both categories: 4 public and 3 private**

Columbia University

Duke University

Harvard University

**Johns Hopkins University\***

**Stanford University\***

University of California, Los Angeles

**University of California, San Diego\***

**University of California, San Francisco\***

**University of Michigan\***

**University of Pennsylvania\***

University of Pittsburgh

**University of Washington\***

University of Wisconsin-Madison

**\*Ranked in both categories**

# **Libraries' Data Services at top US Research Universities**

(Data collection: August 1-31, 2020)

## **Storage and sharing**

- Independent data portal: 3 (1 private; 2 public)
- Part of institutional repository: 7 (3 private; 4 public)
- Dataverse, an open data consortium: 3 ( 2 private; 1 public)

## **Professional support**

- Dedicated unit: 11 (5 private; 6 public)
- Part of digital service: 2 (public)

## **Service programming**

- Yes, 13
- Emerging tools, discovery & evaluation, process & analysis, share & archive, etc.

# **Libraries' Data Services at top US Research Universities**

## **Emerging metadata/record elements:**

- Citation format
- Impact analysis (i.e., # of downloads, citations, views)
- DOI
- Tags

## **Dataset collections:**

- Over 5,000 datasets: 2 (1 private; 1 public)
- Under 500 datasets: 9 (4 private; 5 public)
- Unknown due to the system structure: 2 (public)

# Challenge #1: Metadata and Search Function Alignment

## Global Impact Study Non-user Survey Data (CSV Format)

No Thumbnail

### View/Open

- Description of New SPSS Variables.docx (106.8Kb)
- Non User Web 2013Jan7.csv (12.05Mb)
- Non\_User Survey Data Readme.docx (113.8Kb)

As part of its commitment to open data, the Global Impact Study is providing the micro-data for the project surveys. This file contains the Global Impact Study non-user survey data in CSV format. The non-user survey data is also available in SPSS/SAV format in this web library. The non-user survey data includes data from over 2,000 non-users in Bangladesh, Brazil, Chile, Ghana, and the Philippines. The non-user survey data follows the non-user survey instrument, which can also be found in this web library. This file also includes a readme document that is intended to provide more information on the non-user survey data, as well as a document with new SPSS variables.

URI  
<http://hdl.handle.net/1773/25292>

Date  
2013

Collections  
TASCHA Repository [320]

Author  
Technology & Social Change Group

Metadata  
[Show full item record](#)

Search

Search ResearchWorks  
 This Collection

BROWSE

- All of ResearchWorks
- Communities & Collections
- By Issue Date
- Authors
- Titles
- Subjects
- This Collection**
- By Issue Date
- Authors
- Titles
- Subjects

<b>dc.contributor.author</b>	Technology & Social Change Group	
<b>dc.date.accessioned</b>	2014-03-27T20:05:16Z	
<b>dc.date.available</b>	2014-03-27T20:05:16Z	
<b>dc.date.issued</b>	2013	
<b>dc.identifier.citation</b>	Technology & Social Change Group (TASCHA). (2013). Global Impact Study non-user survey data. Seattle: Technology & Social Change Group, University of Washington Information School.	en_US
<b>dc.identifier.uri</b>	<a href="http://hdl.handle.net/1773/25292">http://hdl.handle.net/1773/25292</a>	
<b>dc.description.abstract</b>	As part of its commitment to open data, the Global Impact Study is providing the micro-data for the project surveys. This file contains the Global Impact Study non-user survey data in CSV format. The non-user survey data is also available in SPSS/SAV format in this web library. The non-user survey data includes data from over 2,000 non-users in Bangladesh, Brazil, Chile, Ghana, and the Philippines. The non-user survey data follows the non-user survey instrument, which can also be found in this web library. This file also includes a readme document that is intended to provide more information on the non-user survey data, as well as a document with new SPSS variables.	en_US
<b>dc.description.sponsorship</b>	International Development Research Centre and Bill & Melinda Gates Foundation	en_US
<b>dc.language.iso</b>	en_US	en_US
<b>dc.publisher</b>	Technology & Social Change Group	en_US
<b>dc.subject</b>	Bangladesh, Brazil, Chile, datasets, Ghana, ICT4D, ICTD, non-user, open data, Open Research, Philippines, Survey	en_US
<b>dc.title</b>	Global Impact Study Non-user Survey Data (CSV Format)	en_US
<b>dc.title.alternative</b>	2013	en_US
<b>dc.type</b>	Dataset	en_US

# Challenge #2: Search Interface

## Example #1

Advanced Search [?](#)

document\_type:dataset

Document Type  [+](#)

Date range:

Limit search to:

Sort by:  Relevance  Publication Date

Format:

[Basic](#)

## Example #2

### My Account

[Login](#)

### Search

[Advanced Search](#)

### Browse

[by School](#)

[by Research Center](#)

[by Year](#)

[by Document Type](#)

[Latest Additions](#)

## Browse by Document Type

Please select a value to browse from the list below.

- [Article](#) (8937)
- [University of Pittsburgh ETD](#) (9949)
- [Other Thesis, Dissertation, or Long Paper](#) (743)
- [Book Section](#) (983)
- [Book](#) (156)
- [Monograph](#) (225)
- [Conference or Workshop Item](#) (1615)
- [Composition](#) (1)
- [Video](#) (12)
- [Audio](#) (3)
- [Dataset](#) (62)
- [Other](#) (80)



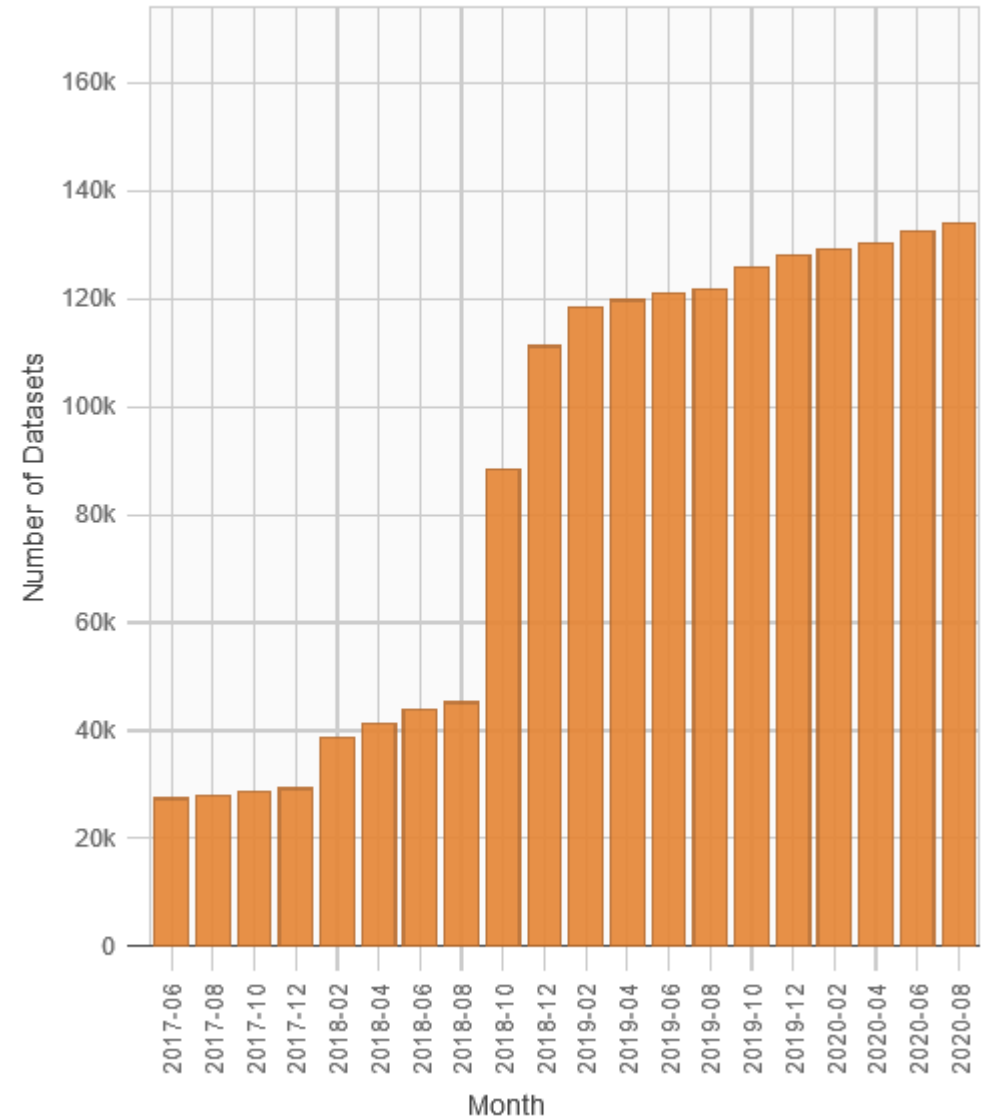
# Dataverse Project (<https://dataverse.org/>)

## DATVERSE REPOSITORIES - A WORLD VIEW

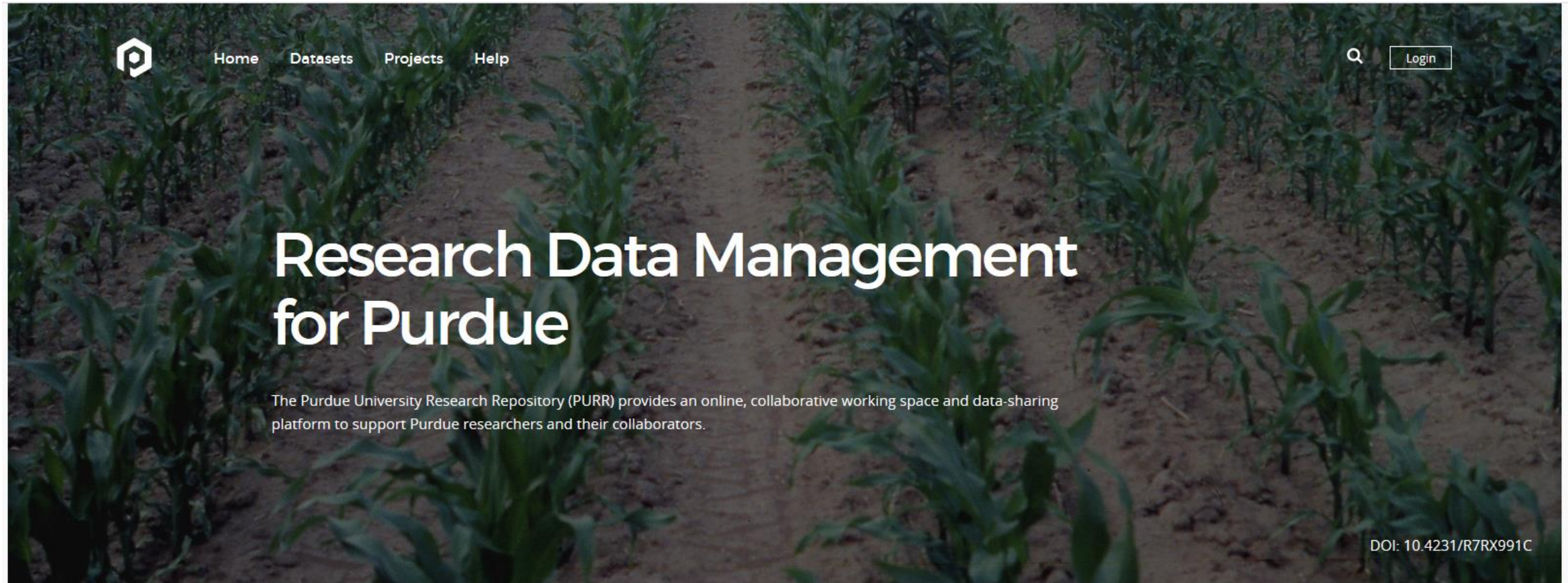
60 Installations



### Total Datasets



# Data Portal Example: Purdue University



A project managed by Purdue Libraries with support of VP-Research and VP-IT

<https://purr.purdue.edu/>

# Data Portal Example: Purdue University

## Archived standard tests to characterize stress tolerance and agronomic performance of alfalfa cultivars

Listed in [Datasets](#)

[About](#) [Supporting Docs](#) [Versions](#) [Citations](#) [Usage](#)

By North American Alfalfa Improvement Conference, [Stanislav Pejša](#)<sup>1</sup>, [Jeffrey Volenec](#)<sup>2</sup>

<sup>1</sup>. *Purdue Libraries* <sup>2</sup>. *Dept. of Agronomy, Purdue University*

The dataset contains standard test protocols used to characterize alfalfa cultivars compiled by the The North American Alfalfa Improvement Conference.

→ [Go to data](#)

↳ [Additional materials available](#)

Version **1.0** - published on 08 Jul 2020

doi:10.4231/N248-MV16 - [cite this](#)

🕒 Archived on 08 Aug 2020

📄 Licensed under [CC0 1.0 Universal](#)

📄 [0 citation\(s\)](#)

👁️ 59 total view(s), 9 download(s)

→ Share: [f](#) [t](#) [s](#) ...



### Description

The dataset contains archived copies of the peer-reviewed standard test protocols used to characterize Insect, disease, and agronomic responses of alfalfa cultivars, 3rd edition (amended 2004) published by the North American Alfalfa Improvement Conference and made available at conference's site <https://www.naaic.org/resource/stdtests.php>. The dataset was supplemented with additional metadata found on the NAAIC webpage. These protocols were used to characterize alfalfa varieties and experimental lines included in the Alfalfa Variety Database (DOI: 10.4231/PHKH-4334), and are provided here as detailed supplemental information supporting the metadata document for Alfalfa Varieties and Experimental Lines 1986 to 1999 (doi:10.4231/FMY9-6966).

All standardized tests followed the format of the Third Edition of the Standardized Test to Characterize Alfalfa Cultivars, edited by Cheryl Fox, et al. and were published by the NAAIC. The standardized test included check cultivars, a list of scientists with expertise, references and in the case of diseases, insects, and nematodes, a distribution and severity U.S. map.

### Cite this work

Researchers should cite this work as follows:

North American Alfalfa Improvement Conference., [Pejša, S.](#), [Volenec, J.](#) (2020). **Archived standard tests to characterize stress tolerance and agronomic performance of alfalfa cultivars.** Purdue University Research Repository. doi:10.4231/N248-MV16

[BibTex](#) [EndNote](#)

### Tags

[Abiotic stress](#) [Agriculture](#) [Agronomy](#) [Alfalfa](#) [Alfalfa breeding](#) [alfalfa\\_db](#) [Bacterial pathogen](#) [Biotic resistance](#) [Biotic stress](#) [Forage quality](#) [Fungal pathogen](#) [Grazing tolerance](#) [Insect stress](#)  
[Lucerne](#) [Medicago](#) [Plant breeding](#) [Plant Pathology](#) [Resistance score](#) [resistance tests](#) [Stress tolerance](#) [Winter hardiness](#)

### Notes

The dataset contains forty five pdf files.

# Instruction Example: University of Houston

What does the Library do for your data needs: A Conversation with UH Libraries research services

**Wenli Gao**  
Data Services Librarian  
**Andrea Malone**  
Coordinator of Research Services



**Communicating**  
Sharing  
Archiving  
Preservation  
Metadata



**Planning**  
Write your plan  
Organize & structure  
Document Workflow

**Conducting**  
Documentation  
Storage & Backup  
Security & Compliance

<https://hdl.handle.net/10657/6792>

# Library and Information Science Education for Data Management Professionals

**Table 3**

Most common job responsibilities, basic and preferred qualifications ( $N = 70$ ).

Responsibilities <sup>a</sup>	Basic qualifications <sup>a</sup>	Preferred qualifications <sup>a</sup>
1. Data collection ( $n = 8$ )	1. Data management ( $n = 7$ )	1. Advanced degree ( $n = 4$ )
Data management ( $n = 8$ )	2. Social sciences ( $n = 6$ )	Data management ( $n = 4$ )
3. Research data ( $n = 7$ )	3. Research data ( $n = 5$ )	Related field ( $n = 4$ )
4. Faculty and students ( $n = 6$ )	4. Higher education ( $n = 4$ )	Social sciences ( $n = 4$ )
5. Data analyst ( $n = 5$ )	Information science ( $n = 4$ )	5. Experience with institutional repositories ( $n = 3$ )
Data services ( $n = 5$ )	Related field ( $n = 4$ )	Experience with metadata ( $n = 3$ )
7. Data analysis ( $n = 4$ )		Research library ( $n = 3$ )

<sup>a</sup> Ranked by frequency; alphabetically for ties.

Chen, H., & Zhang, Y. (2017). Educating Data Management Professionals: A Content Analysis of Job Descriptions. *Journal of Academic Librarianship*, 43(1), 18–24.



Brown University  
GIS and Data Librarian

Apply Now

### Education and Experience

- Required education: ALA-accredited master's or other advanced degree [including ABD] in a discipline working with GIS, geography, or spatial analysis.
- Successful experience in staff supervision and training in an academic and/or library environment.
- Experience or demonstrated potential supporting researchers with data services, including data discovery, visualization, and data storage preferred
- Demonstrated experience with using GIS applications including the Esri suite of ArcGIS applications.
- Ability to represent the Library to University and external audiences.

### Job Competencies

- Knowledge of the current landscape for open data, statistical and geospatial analysis, GIS, data visualization and research data management, and will have hands-on experience or a willingness to learn software and techniques in support of this work.
- Excellence in the following skills: verbal and written communication, interpersonal, planning, organizational, and analytical.
- Demonstrated commitment to diversity and understanding of the contributions a diverse workforce brings to the workplace.

<https://joblist.ala.org/job/gis-and-data-librarian/54399928/> (posted July 31, 2020)



## Los Alamos National Laboratory Data Management Librarian

Apply Now

### Requirements:

- Bachelor's degree, MLS/MLIS degree (ALA-accredited preferred) and a minimum of two years of related experience, or an equivalent combination of education and experience; five years of experience, including two years as a science, engineering or research data service librarian in a research library, required for Level 2 position
- Knowledge of the following:
  - Data services needs across the research lifecycle and data best practices, such as the FAIR principles
  - General and discipline-specific metadata standards and schemas, such as DCMI and CSDGM
  - Data repositories and public data sets
  - Data management plan requirements for U.S. government agencies and leading science funding organizations
  - Languages and tools such as Tableau, R, Python or OpenRefine
- Ability to provide reference and research services, instruction and outreach
- Ability to work both independently and as part of a collaborative team with researchers, staff and a growing, culturally diverse laboratory community
- Experience providing consultation to a range of audiences, including scientists, computer engineers and information technologists
- Ability to obtain DOE Q clearance (usually requires U.S. citizenship)

### Desired Skills:

- Development and assessment of data models
- STEM degree or experience working in a research-intensive environment
- Demonstrated experience with APIs
- Knowledge of emerging data management technological advances and trends, such as utilization of Machine Learning (ML) and Artificial Intelligence (AI)
- Knowledge of data analysis and/or visualization practices

<https://joblist.ala.org/job/data-management-librarian/54525041/> (posted August 20, 2020)

**Required Qualifications**

- MLIS from an ALA Accredited institution
- Working knowledge of the data management practices and requirements of researchers and external funding bodies (e.g. NSF, NIH, etc.)
- Knowledge of the research data life cycle
- Familiarity with a digital repository platform
- Knowledge of metadata standards and formats used in research (e.g. Dublin Core, DDI, DOI, ORCID)
- Familiarity with the rich variety of physical and digital types commonly found in an academic research institution.
- Evidence of a successful ability to manage and respond effectively to changing needs and priorities
- Excellent interpersonal skills
- Excellent communication skills, including oral and written and presentation skills
- Ability to work independently and initiate needs-based projects with little guidance
- Demonstrated ability to carry out complex long-term and short-term project-based work with little guidance
- Ability to organize project-based information and data
- Ability to meet the university's requirements for promotion and tenure, including maintaining excellence in librarianship, developing a record of scholarship, and participating in university and professional service.
- Working knowledge of diversity, equity, and inclusion issues

**Preferred Qualifications**

- Academic background in the sciences or social sciences
- Research data management training or work experience
- Experience working with datasets, including data management, curation, and retrieval
- Experience creating and evaluating data management plans
- Knowledge of data science tools (e.g. Jupyter Notebook, GitHub) and research data management resources
- Experience with data preservation and curation
- Experience in successfully managing technology, staff, and teams
- Experience with project management and assessment
- Experience developing and delivering instructional content and consultations in an academic environment



# For Medical University Libraries UCSF as an example:

<https://guides.ucsf.edu/c.php?g=101037&p=2704613>

## Reproducible Data Management

Information and resources for reproducible data management for the UCSF research community

Home

Make a Data Management Plan

Find Research Data

[UCSF Research Data](#)

[Other Sources of Research Data](#)

[Clinical Trial Data \(non-UCSF\)](#)

Organize and Document Data

Store Data

Visualize Data

Publish Data

### UCSF Research Data

- [Research Data Browser](#)

A tool to explore de-identified UCSF patient records for creating cohorts or identifying potential subjects for studies.

- [UCSF Clinical Data](#)

Request counts, de-identified, or identified data from UCSF electronic medical record

- [UCSF Information Commons](#)

A large database of UCSF clinical data and related basic science and genomic information. Currently in beta.

- [Dryad](#)

This public data repository contains over 20,000 openly accessible research datasets and is free to use for UCSF researchers

# UCSF Data Resources: <https://data.ucsf.edu/>

## UCSF Data Resources

Research ▼ Clinical & Quality Education Administration Self-service Analytics (SSA) ▼ General Data Assets ▼ Process & Policy ▼ Events & Community ▼



Data powers UCSF work in healthcare,  
research, education and administration

# UCSF Data Resources

[Research](#) ▼ [Clinical & Quality](#) [Education](#) [Administration](#) [Self-service Analytics \(SSA\)](#) ▼ [General Data Assets](#) ▼ [Process & Policy](#) ▼ [Events & Community](#) ▼

[Home](#) > [Our Vision](#)

## Our Vision

Data.ucsf.edu will be a centralized resource that will enable UCSF staff, faculty and students to understand the available data and analytics tools and services available to support their clinical, research, educational and administrative efforts.

Data.ucsf.edu will enable you to:

- **Discover UCSF data assets and how to gain access to them.** You will have the ability to learn about the clinical, educational, and administrative data assets available for use at UCSF, including data model definitions, data dictionaries, and comprehensive definitions of available databases. These metadata will be searchable, which will ease the efforts in finding data assets that will suit a particular user's needs. Information will be provided on how to gain access for each data asset. The types of data assets published on the website include:
  - Data Sources
  - Reports and dashboards
  - Applications
  - Measure and metric definitions
  - Learn about how to register a new data asset

# UCSF Information Commons:

<https://informationcommons.ucsf.edu/>

Information Commons


Data

Tools

Infrastructure

Getting Started

Support



Harnessing the power of machine intelligence to advance precision medicine.

# Information Commons

## **Governing Campus Units**

Center for Digital Health Information (CDHI)

Clinical and Translational Science Institute (CTSI)

Information Technology (IT)

Baker Computational Health Sciences Institute (BCHSI) — *Owner organization*

Precision Medicine Initiative

Radiology & Biomedical Imaging

SOM Tech

# Information Privacy and Open Data: A Lesson from 2006

The New York Times

## *A Face Is Exposed for AOL Searcher No. 4417749*

By [Michael Barbaro](#) and [Tom Zeller Jr.](#)

Aug. 9, 2006



Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”



Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Erik S. Lesser for The New York Times

<https://www.nytimes.com/2006/08/09/technology/09aol.html>

## **Discussion:**

- **Campus infrastructure and administration**
- **Technical issues: data curation, management, retrieval, manipulation, analysis, etc.**
- **Education and professional development**
- **On-campus and off-campus collaboration**

**Q&A**