Situating Discovery in the Data Sharing Environment

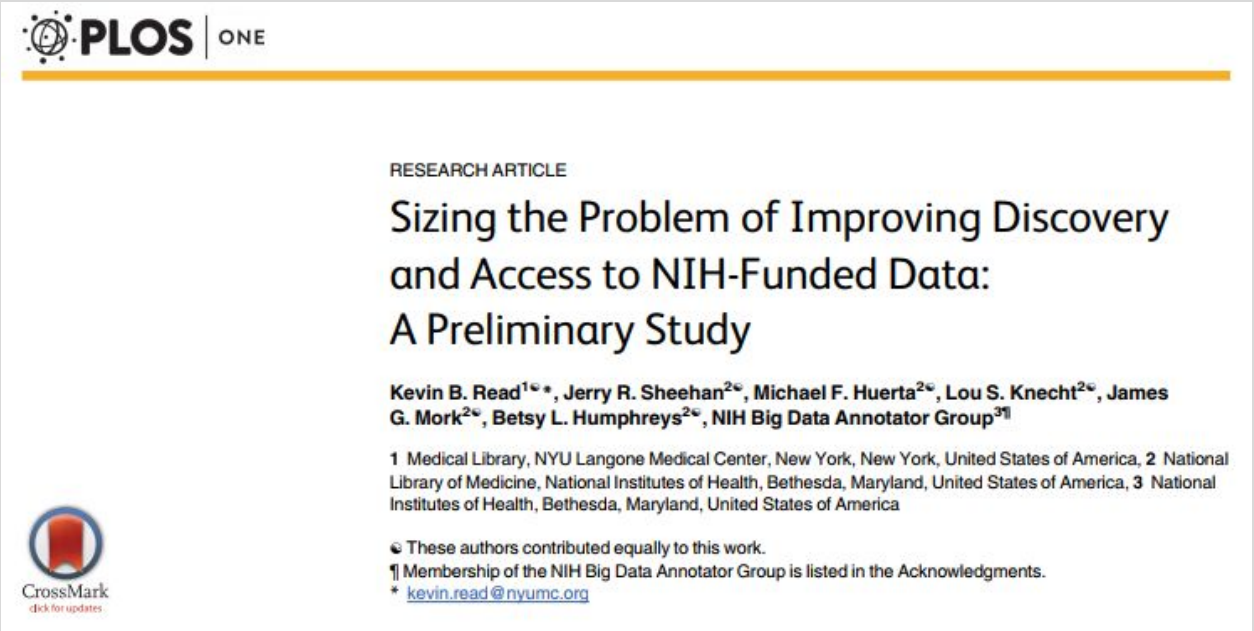
The Efforts and Goals of the Data Discovery Collaboration Project

Nicole Contaxis, MLIS
Lead, NYU Data Catalog
nicole.contaxis@nyulangone.org

The Problem

Data Discovery, ca. 2011

88% of datasets from NIH-funded publications were ‘invisible’ - between 200,000 and 235,000 datasets



PLOS ONE

RESEARCH ARTICLE

Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study

Kevin B. Read[1☯*], Jerry R. Sheehan[2☯], Michael F. Huerta[2☯], Lou S. Knecht[2☯], James G. Mork[2☯], Betsy L. Humphreys[2☯], NIH Big Data Annotator Group[3¶]

1 Medical Library, NYU Langone Medical Center, New York, New York, United States of America, 2 National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, 3 National Institutes of Health, Bethesda, Maryland, United States of America

☯ These authors contributed equally to this work.
¶ Membership of the NIH Big Data Annotator Group is listed in the Acknowledgments.
* kevin.read@nyumc.org

CrossMark
click for updates

Data Discovery, Now

- More repositories and other tools available to help researchers share data
- More requirements and policies that encourage data sharing
- Yet...

Data Discovery, Now

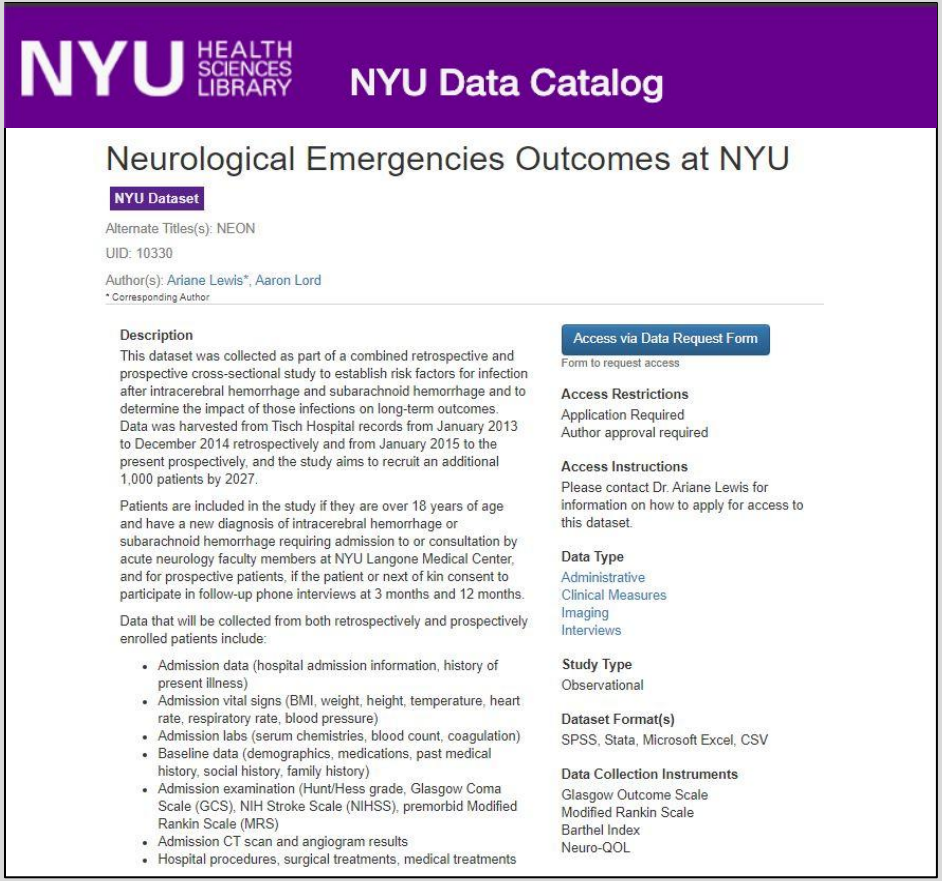
- More repositories and other tools available to help researchers share data
- More requirements and policies that encourage data sharing
- Yet…

- Complex environment with thousands of repositories (2,433 listed in re3data)
- Difficult to decide where to deposit one’s data or look for data for re-use
- Some data (e.g., EHR data) may be useful but will not be included in repositories

The Technology

The NYU Data Catalog: An Overview

- Provides a standardized metadata schema to describe data



NYU HEALTH SCIENCES LIBRARY NYU Data Catalog

Neurological Emergencies Outcomes at NYU

NYU Dataset

Alternate Titles(s): NEON

UID: 10330

Author(s): Ariane Lewis*, Aaron Lord

* Corresponding Author

Description

This dataset was collected as part of a combined retrospective and prospective cross-sectional study to establish risk factors for infection after intracerebral hemorrhage and subarachnoid hemorrhage and to determine the impact of those infections on long-term outcomes. Data was harvested from Tisch Hospital records from January 2013 to December 2014 retrospectively and from January 2015 to the present prospectively, and the study aims to recruit an additional 1,000 patients by 2027.

Patients are included in the study if they are over 18 years of age and have a new diagnosis of intracerebral hemorrhage or subarachnoid hemorrhage requiring admission to or consultation by acute neurology faculty members at NYU Langone Medical Center, and for prospective patients, if the patient or next of kin consent to participate in follow-up phone interviews at 3 months and 12 months.

Data that will be collected from both retrospectively and prospectively enrolled patients include:

- Admission data (hospital admission information, history of present illness)
- Admission vital signs (BMI, weight, height, temperature, heart rate, respiratory rate, blood pressure)
- Admission labs (serum chemistries, blood count, coagulation)
- Baseline data (demographics, medications, past medical history, social history, family history)
- Admission examination (Hunt/Hess grade, Glasgow Coma Scale (GCS), NIH Stroke Scale (NIHSS), premorbid Modified Rankin Scale (MRS)
- Admission CT scan and angiogram results
- Hospital procedures, surgical treatments, medical treatments

Access via Data Request Form

Form to request access

Access Restrictions

Application Required
Author approval required

Access Instructions

Please contact Dr. Ariane Lewis for information on how to apply for access to this dataset.

Data Type

Administrative
Clinical Measures
Imaging
Interviews

Study Type

Observational

Dataset Format(s)

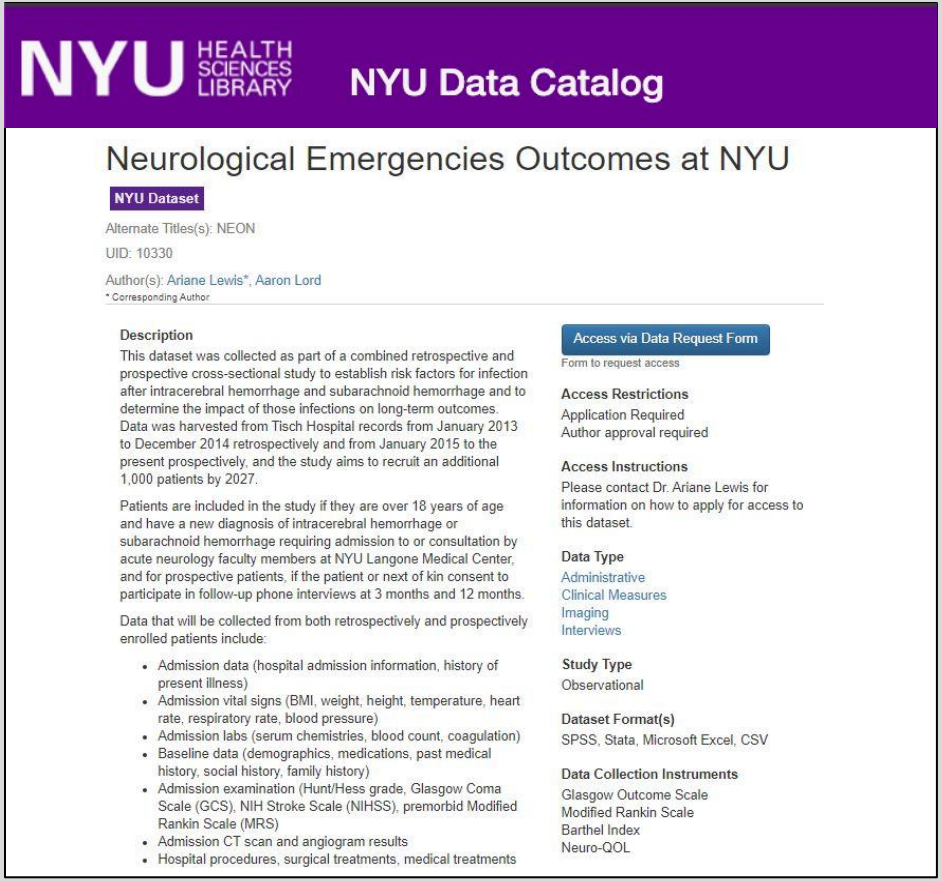
SPSS, Stata, Microsoft Excel, CSV

Data Collection Instruments

Glasgow Outcome Scale
Modified Rankin Scale
Barthel Index
Neuro-QOL

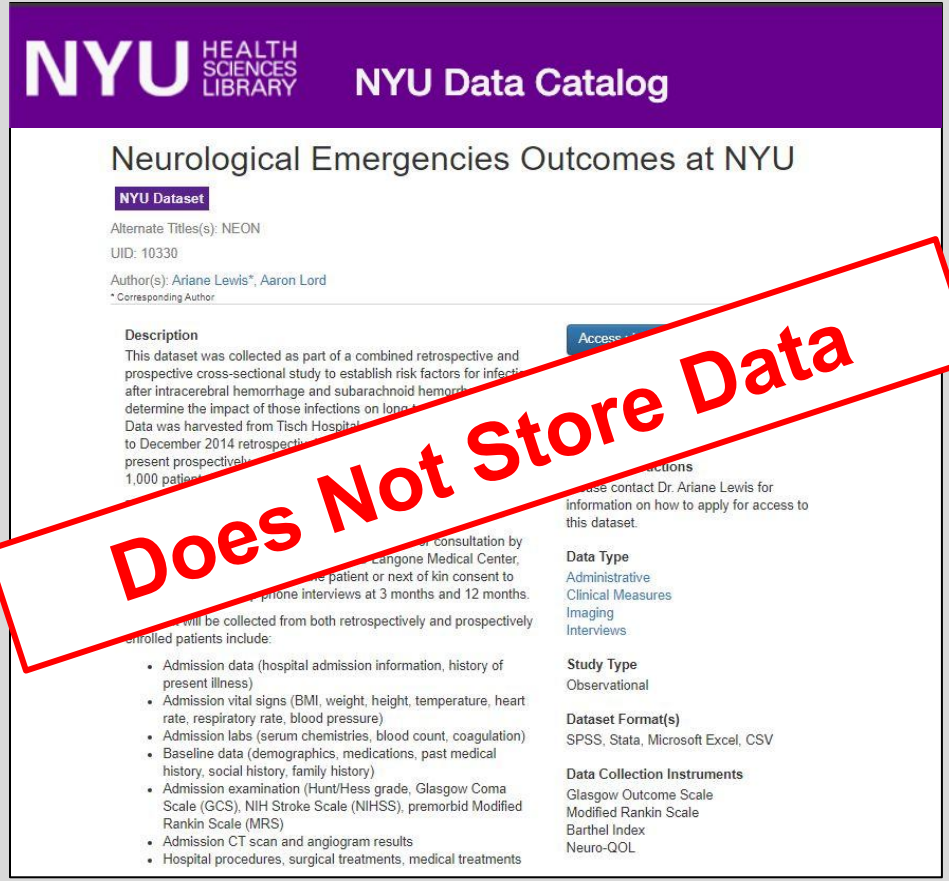
The NYU Data Catalog: An Overview

- Provides a standardized metadata schema to describe data
- Makes research data discoverable regardless of where it is stored



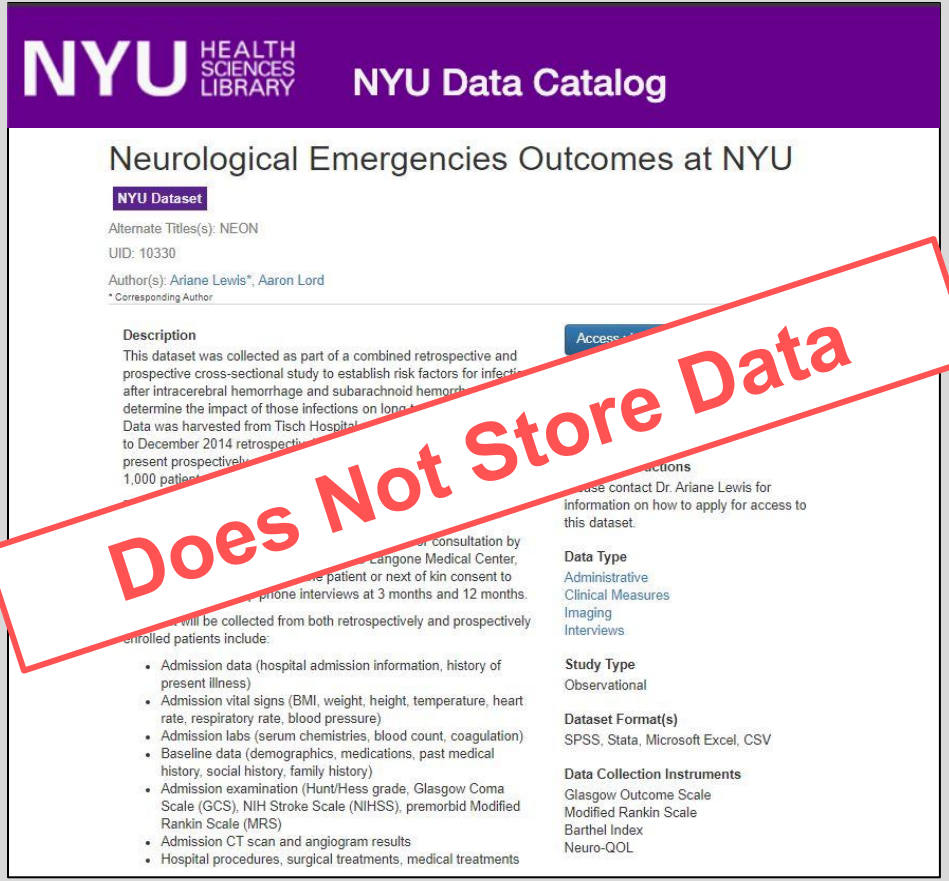
The NYU Data Catalog: An Overview

- Provides a standardized metadata schema to describe data
- Makes research data discoverable regardless of where it is stored



The NYU Data Catalog: An Overview

- Provides a standardized metadata schema to describe data
- Makes research data discoverable regardless of where it is stored
- Open source
  - Code on GitHub
  - Documentation on OSF



An Extremely Abridged History, Part One

The NYU Data Catalog

- Increase visibility of research data for researchers at NYU
- Encourage collaboration

An Extremely Abridged History, Part One

The NYU Data Catalog

- Increase visibility of research data for researchers at NYU
- Encourage collaboration

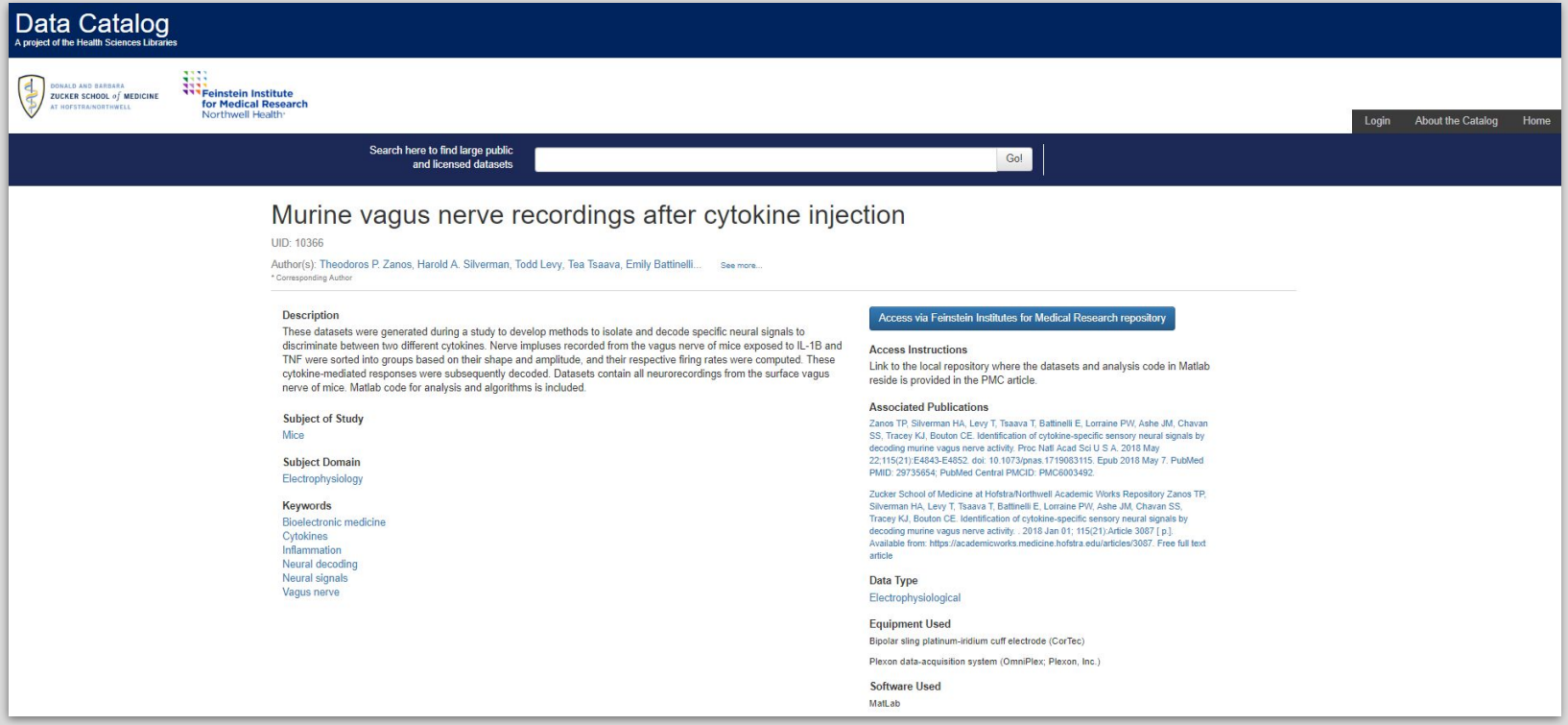
The Data Catalog Collaboration Project

- Improve data discovery through institutional data catalogs
- Work together to develop, implement, and improve the Data Catalog software

Collaboration Members

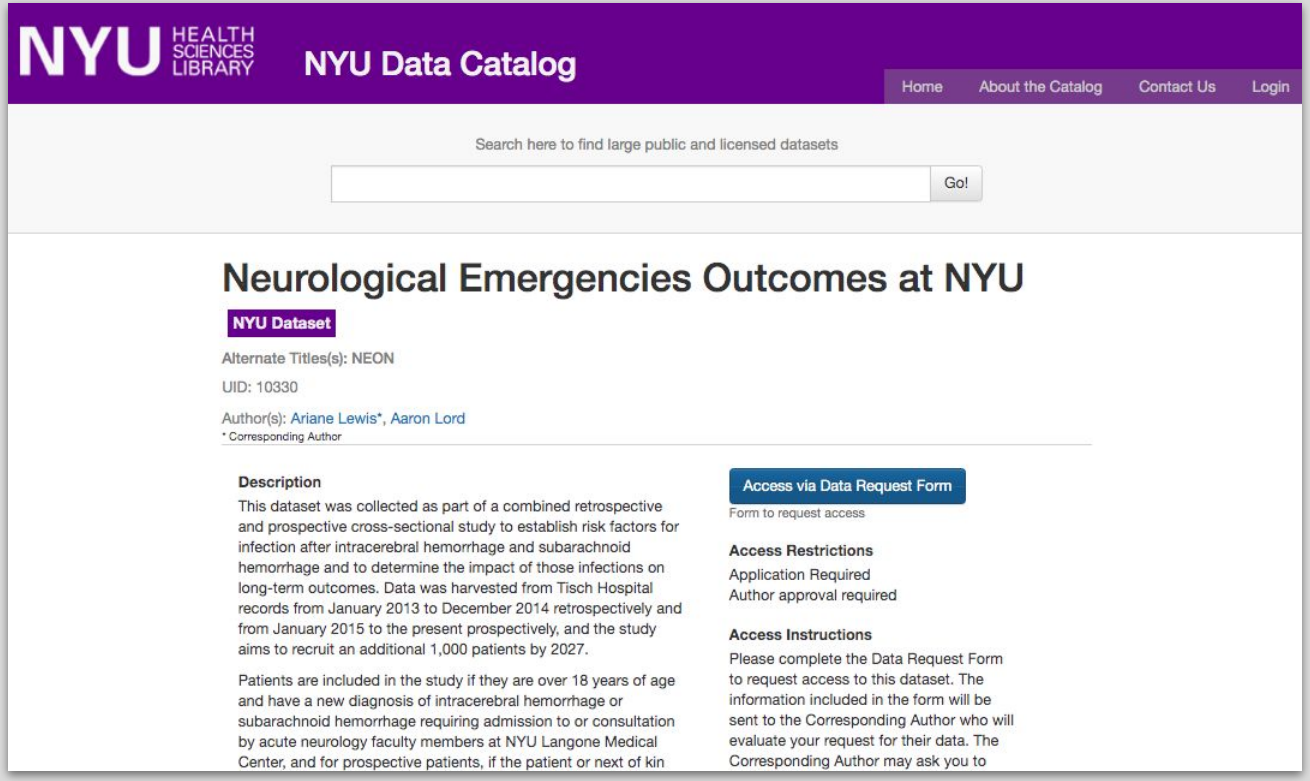
- NYU Langone Health
- University of Pittsburgh
- University of Maryland at Baltimore
- Weill Cornell Medicine
- Wayne State University
- Zucker School of Medicine at Hofstra/Northwell
- Memorial Sloan Kettering Cancer Center
- The University of North Carolina at Chapel Hill
- Duke University
- University of Virginia

Storage Agnostic Data Discovery



Institutional Repository

Storage Agnostic Data Discovery

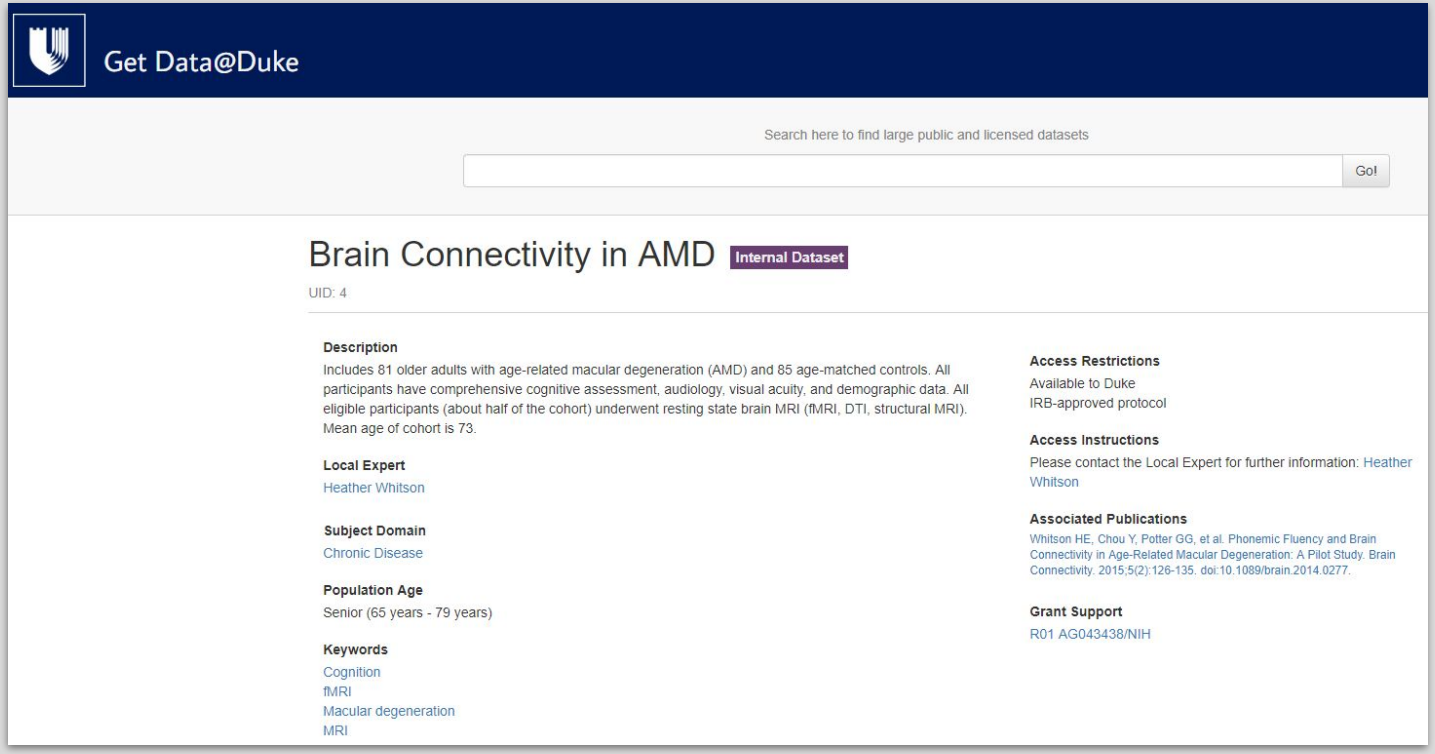


Institutional Center or Department

Storage Agnostic Data Discovery

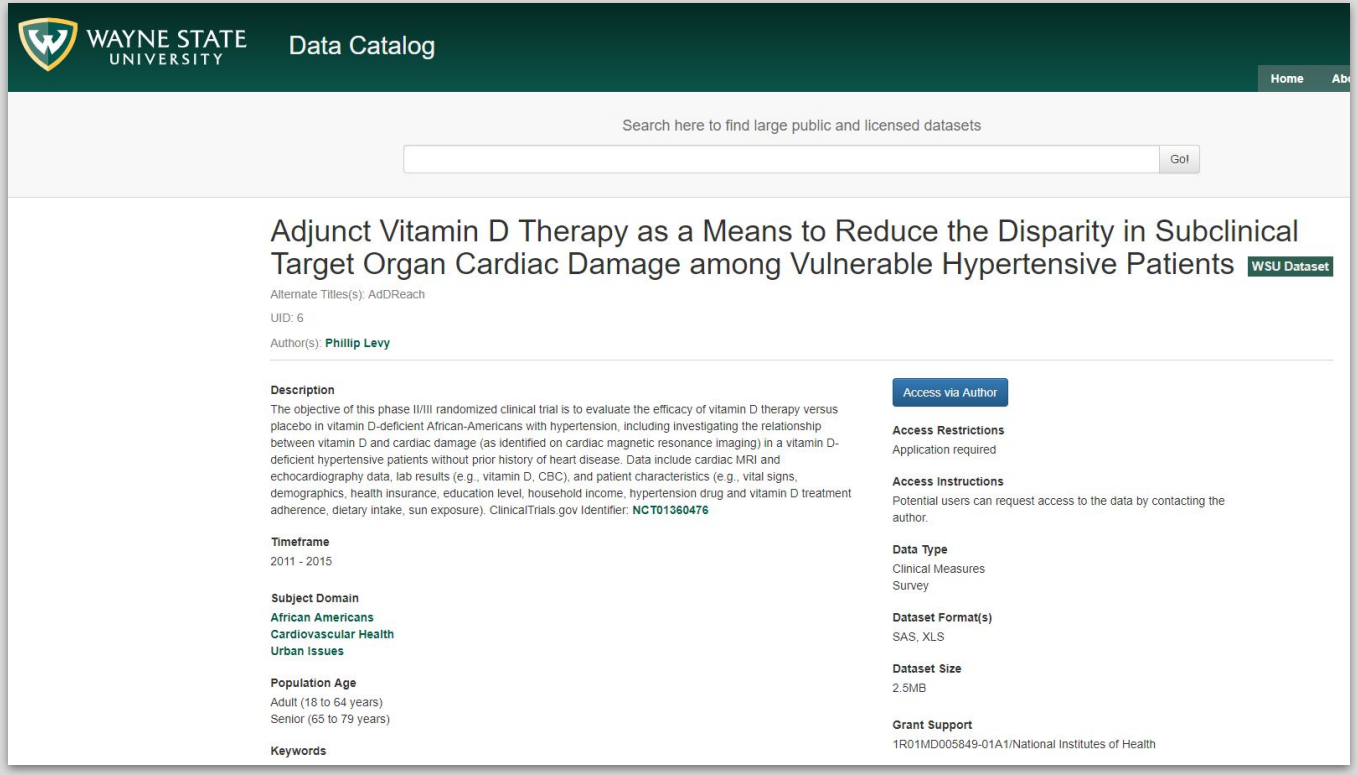
External Repository

Storage Agnostic Data Discovery



Restricted Access Data

Storage Agnostic Data Discovery

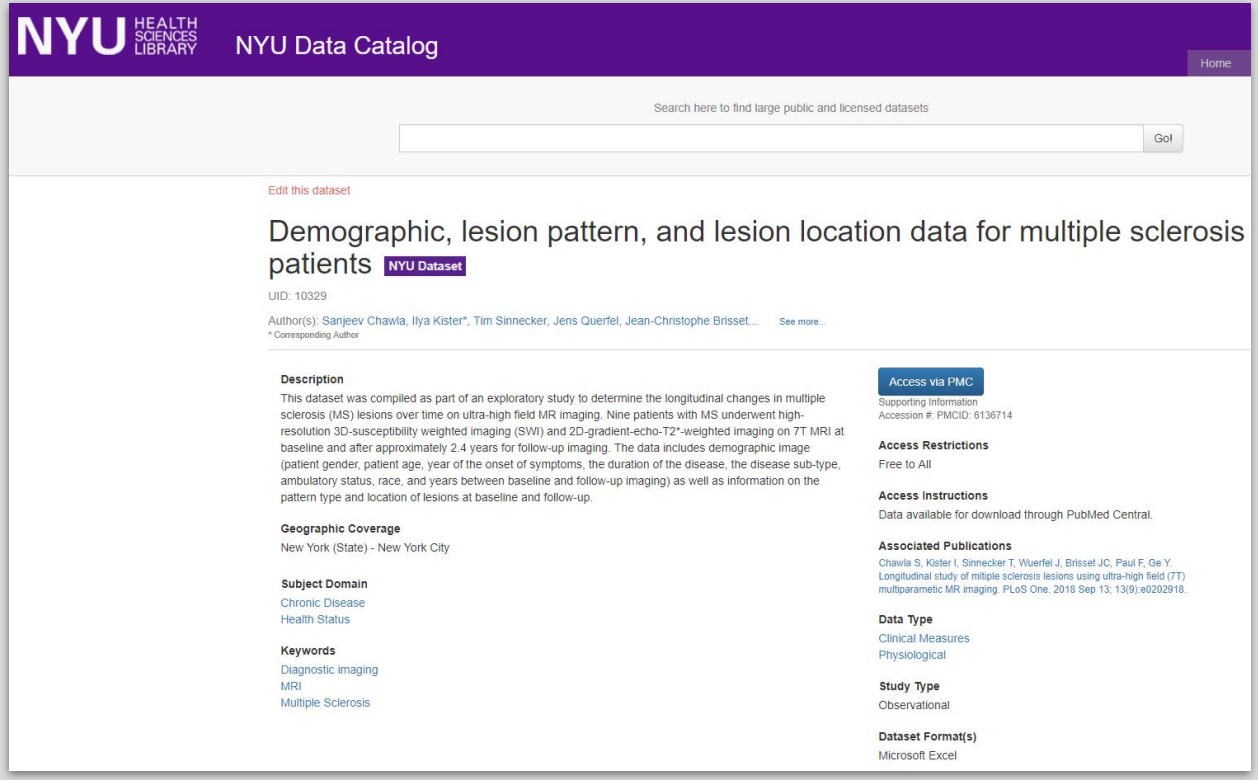


Datasets Via Author Only

Storage Agnostic Data Discovery

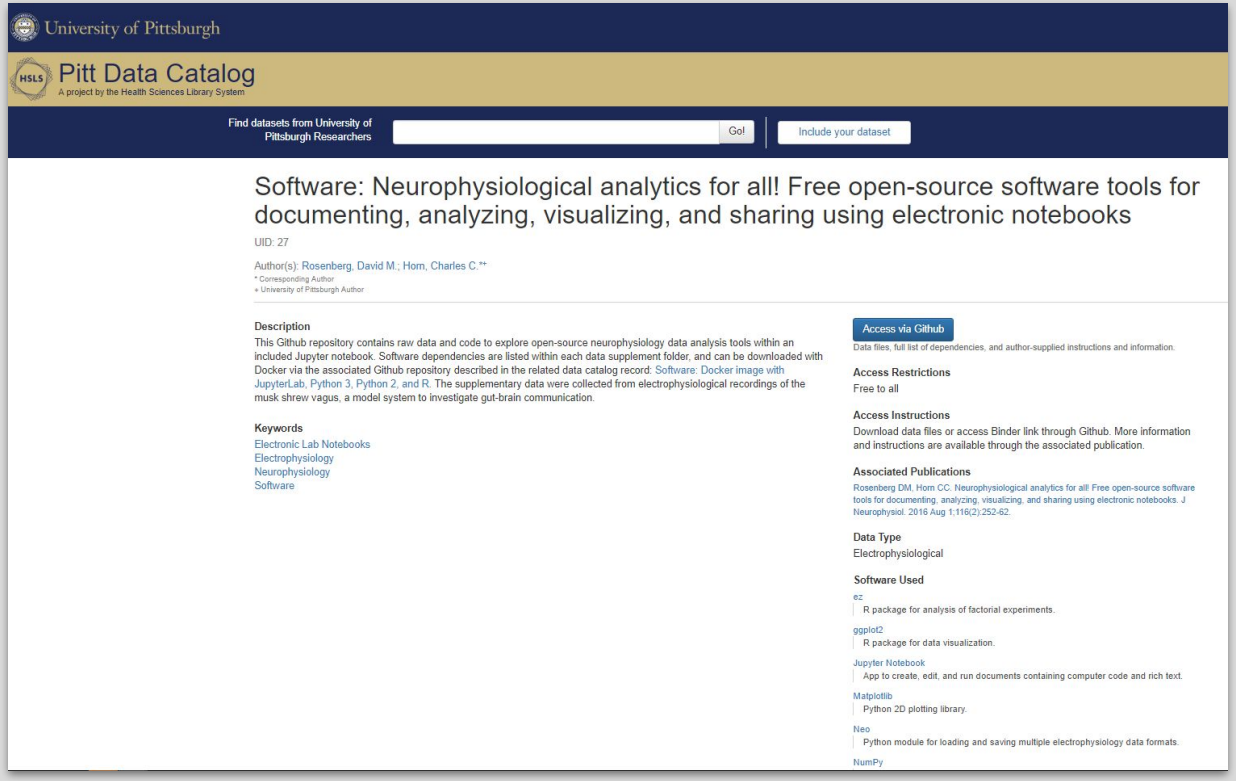
Electronic Health Record Data

Storage Agnostic Data Discovery



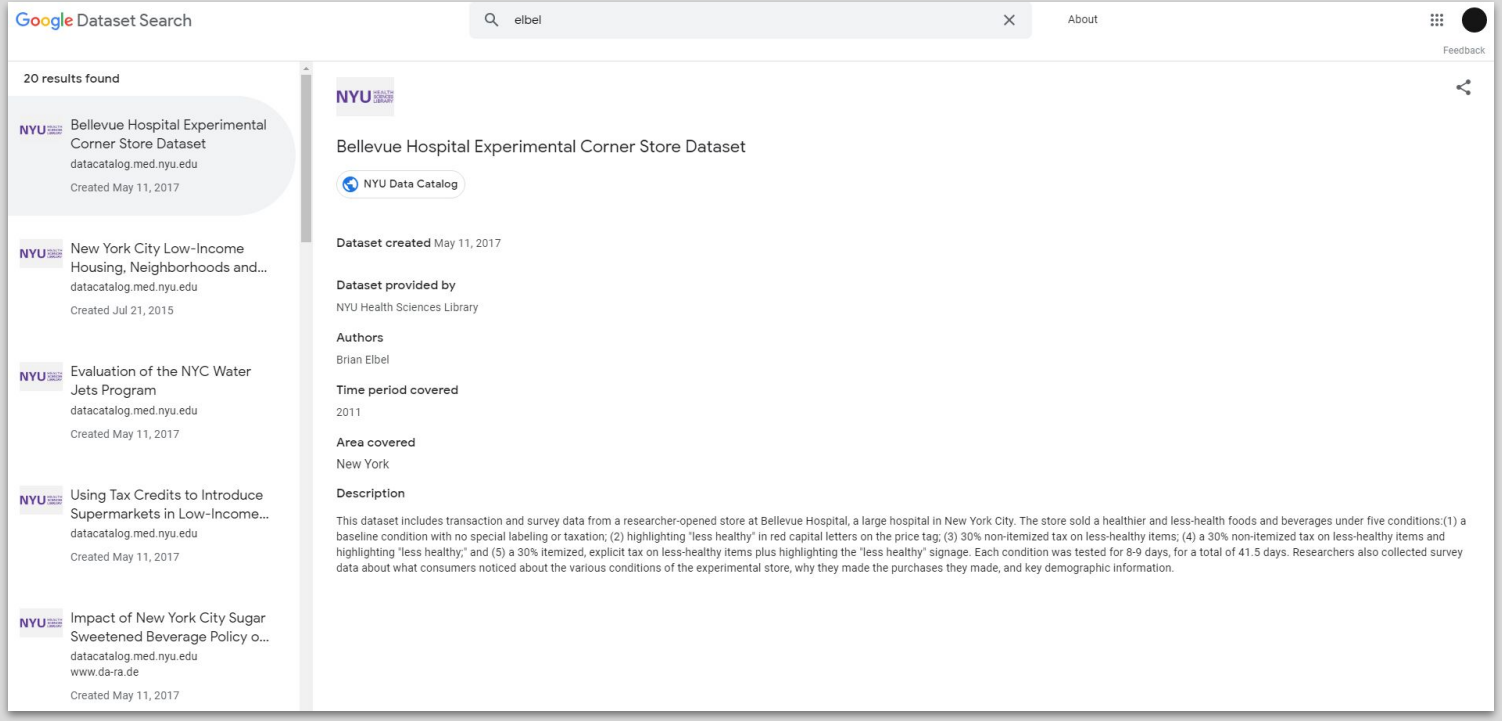
Datasets in Supplemental Files of Published Articles

Storage Agnostic Data Discovery



Datasets with Associated Software

Interoperability with Schema.org



All Datasets Discoverable Via Google Dataset Search

An Extremely Abridged History, Part Two

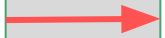
The Data Catalog Collaboration Project

- Improve data discovery through institutional data catalogs
- Work together to develop, implement, and improve the Data Catalog software

An Extremely Abridged History, Part Two

The Data Catalog Collaboration Project

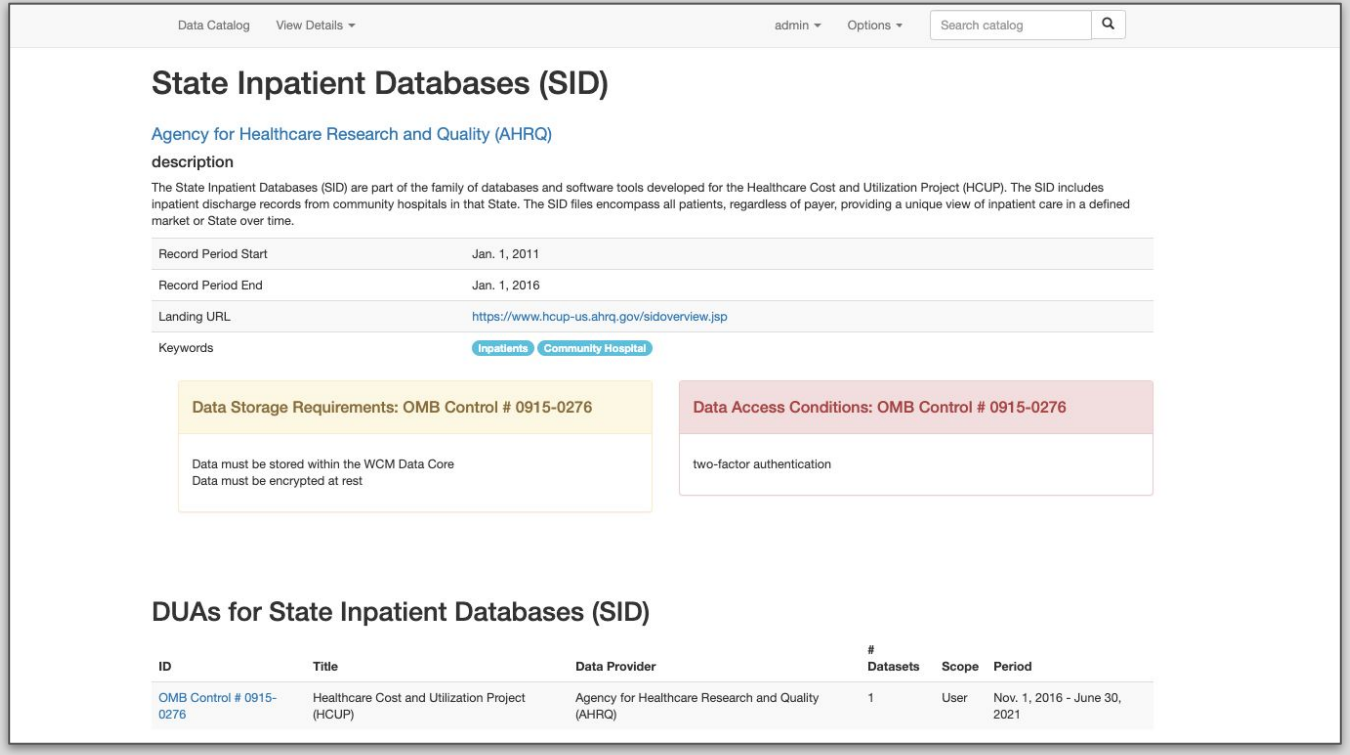
- Improve data discovery through institutional data catalogs
- Work together to develop, implement, and improve the Data Catalog software



The Data Discovery Collaboration Project

- Improve data discovery regardless of tool or platform
- Work towards metadata interoperability between discovery tools
- (Continue developing the Data Catalog software)

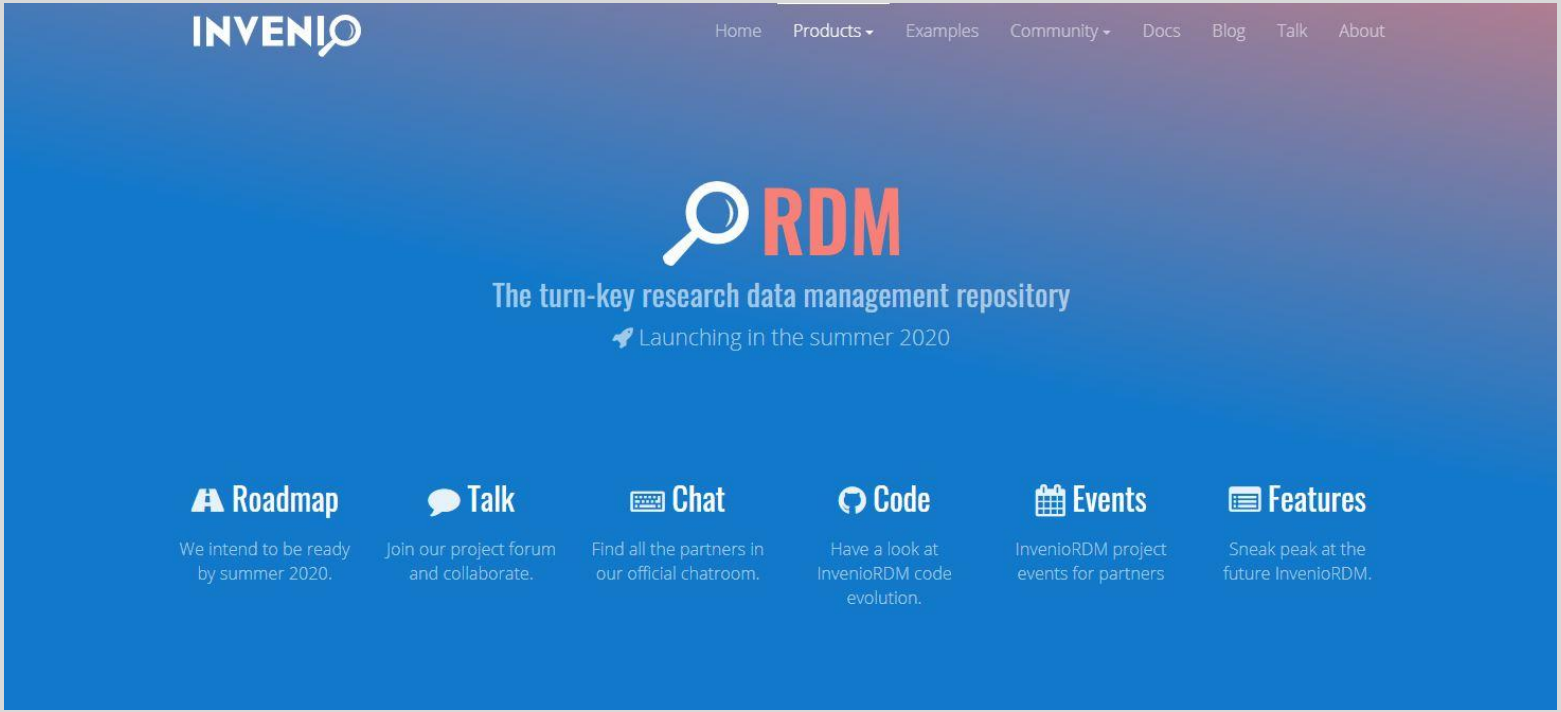
Tool Agnostic Data Discovery



Weill Cornell Medicine

Metadata Relating to Data Governance and Access

Tool Agnostic Data Discovery



Northwestern

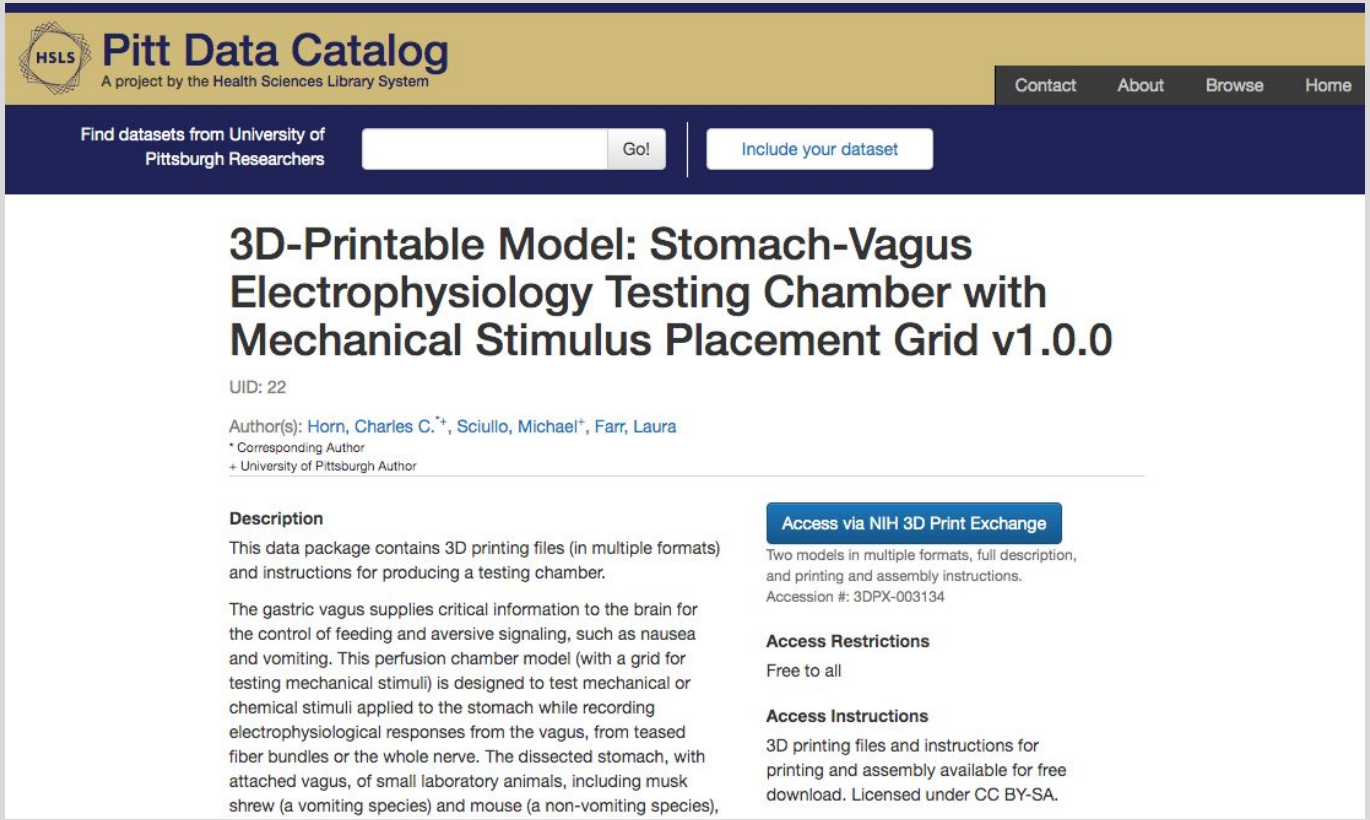
CD2H-Developed RDM Repository

The People

Helenmary Sheridan, MLIS

- Data Services Librarian at the University of Pittsburgh
- In fall 2018, she worked with researcher Charles Horn to catalog software and 3D models created during his research

Describing 3D Models to Improve Discovery



Pitt Data Catalog
A project by the Health Sciences Library System

Contact About Browse Home

Find datasets from University of Pittsburgh Researchers

Go!

Include your dataset

3D-Printable Model: Stomach-Vagus Electrophysiology Testing Chamber with Mechanical Stimulus Placement Grid v1.0.0

UID: 22

Author(s): Horn, Charles C.*+, Sciullo, Michael+, Farr, Laura
* Corresponding Author
+ University of Pittsburgh Author

Description

This data package contains 3D printing files (in multiple formats) and instructions for producing a testing chamber.

The gastric vagus supplies critical information to the brain for the control of feeding and aversive signaling, such as nausea and vomiting. This perfusion chamber model (with a grid for testing mechanical stimuli) is designed to test mechanical or chemical stimuli applied to the stomach while recording electrophysiological responses from the vagus, from teased fiber bundles or the whole nerve. The dissected stomach, with attached vagus, of small laboratory animals, including musk shrew (a vomiting species) and mouse (a non-vomiting species),

Access via NIH 3D Print Exchange

Two models in multiple formats, full description, and printing and assembly instructions.
Accession #: 3DPX-003134

Access Restrictions

Free to all

Access Instructions

3D printing files and instructions for printing and assembly available for free download. Licensed under CC BY-SA.

Describing Models at NYU Langone Health

- Working on describing models created by Christopher Collins, PhD at NYU
- Using Helenmary Sheridan's records as a model and building on her expertise

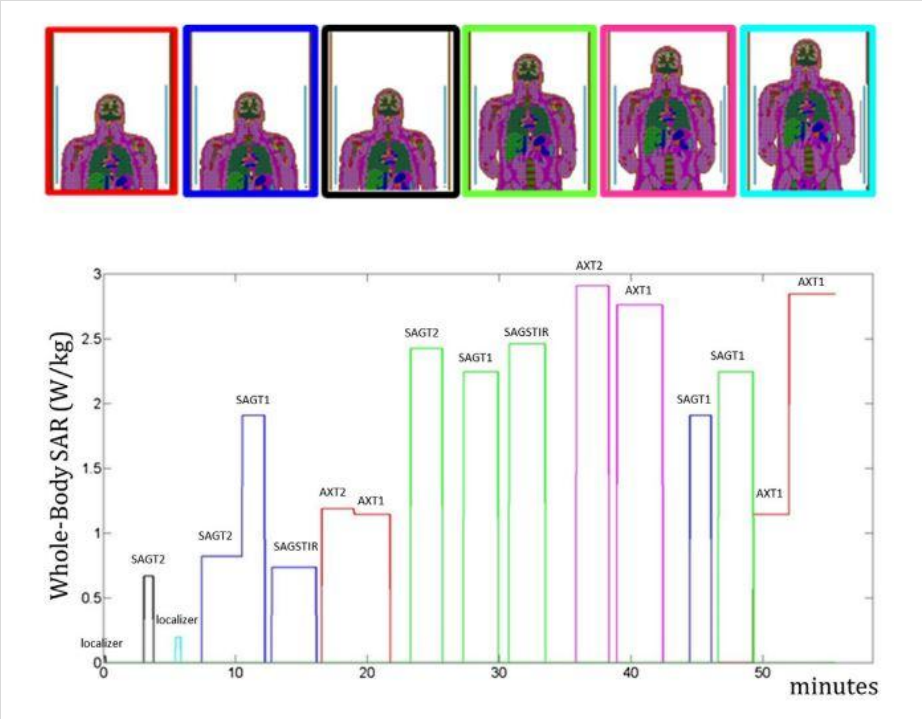


Figure from: Carluccio G, Collins CM. Optimization of the order and spacing of sequences in an MRI exam to reduce the maximum temperature and thermal does. Magn Reson Med. 2019 Mar;81(3):2161-2166. doi: 10.1002/mrm.27503

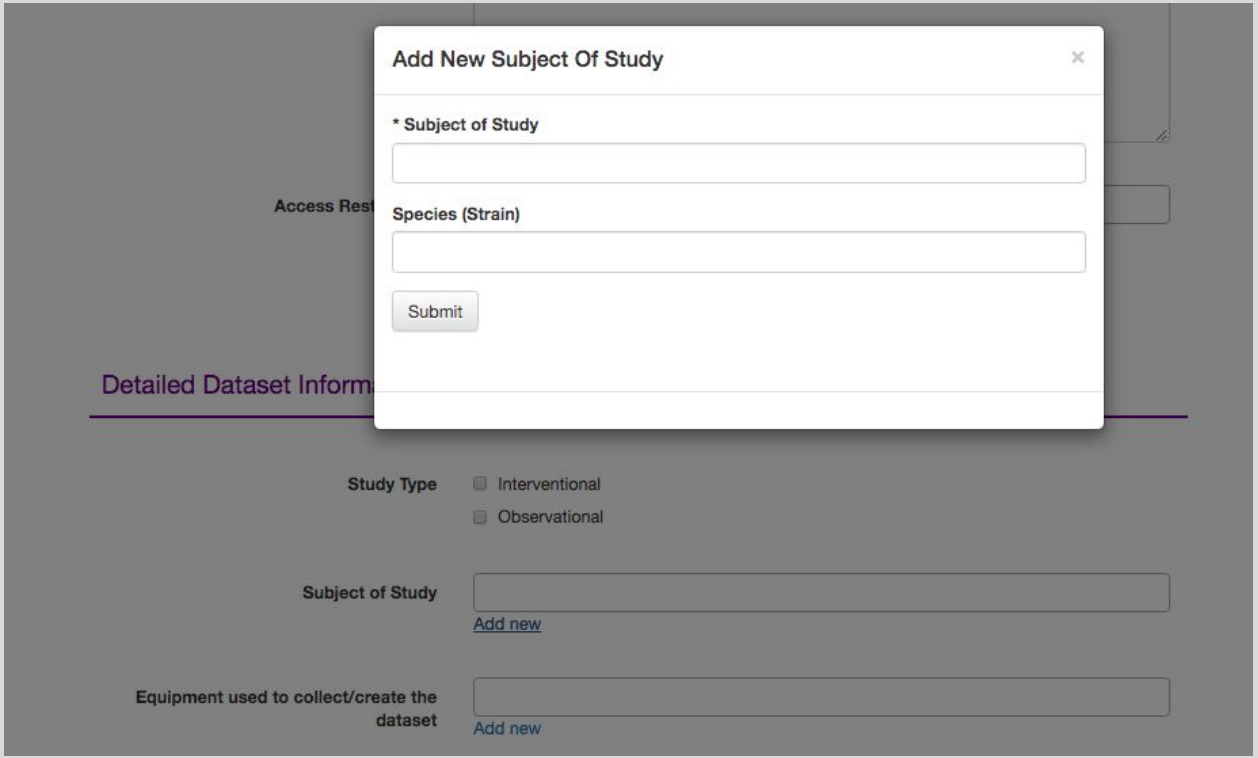
Carrie Iwema, PhD, MLS, AHIP

Formed the Basic Science Working Group in order to:

- Adapt the metadata to better describe basic science data
- Investigate existing metadata elements to ensure that they work for basic science as well

Read more at: http://bit.ly/2r6A1OS

Basic Science Working Group



Suggested changes to:

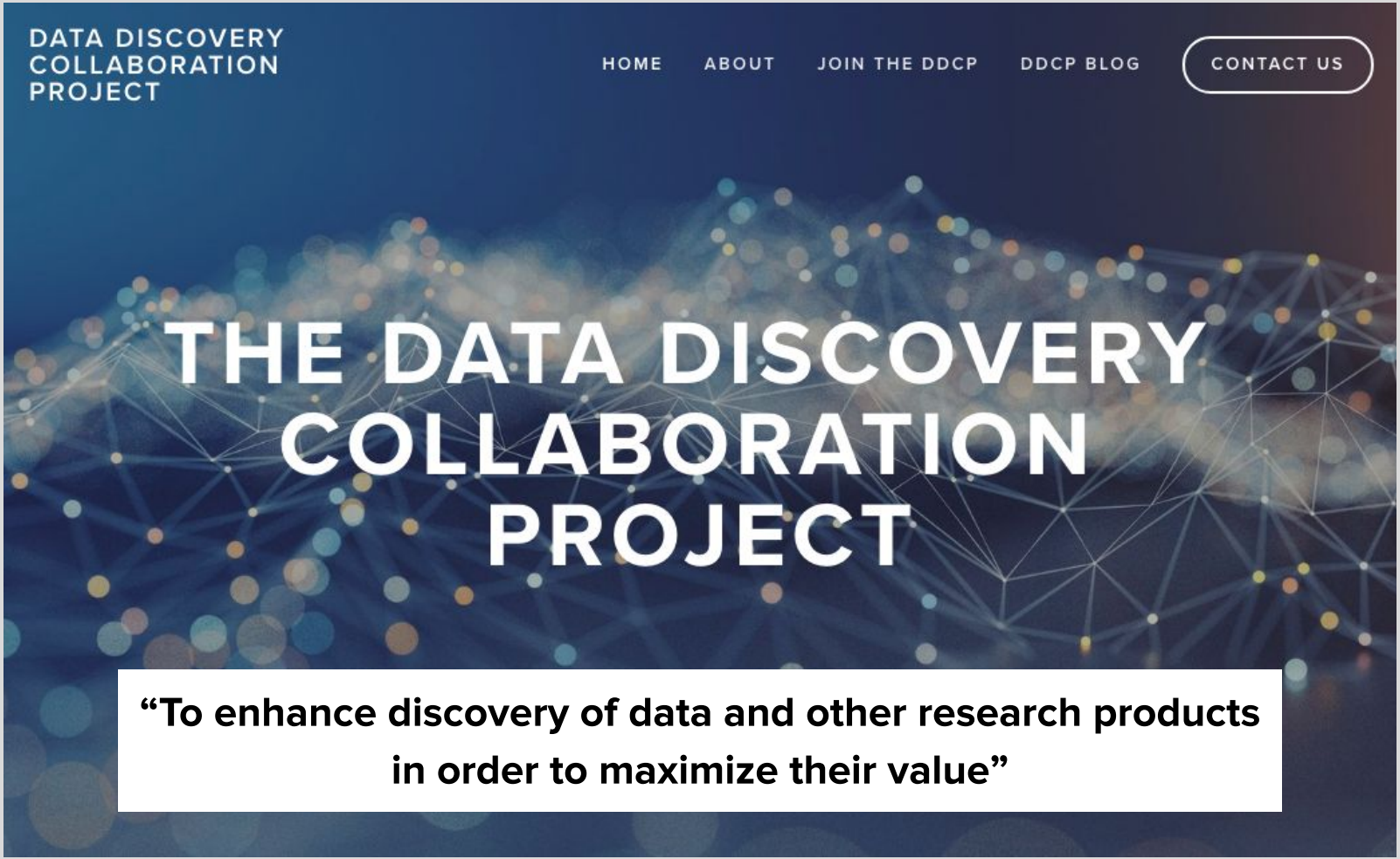
- Data Type
- Subject of Study
- Equipment
- Software
- Study Type

The Future

Future Growth

- Establish governance structure
- Create an advisory board
- Recruit institutional members who use/develop relevant platforms and tools
- Investigate possibilities for integration of existing tools
- Continue working within our communities to improve data discovery

Icon created by Gan Khoon Lay from Noun Project


DATA DISCOVERY COLLABORATION PROJECT
HOME
ABOUT
JOIN THE DDCP
DDCP BLOG
CONTACT US
THE DATA DISCOVERY COLLABORATION PROJECT
“To enhance discovery of data and other research products in order to maximize their value”

What we offer to each other

- Boots-on-the-ground experience with developing metadata to suit researcher needs
- Experience addressing obstacles to data sharing including researcher apathy and institutional barriers
- A platform for addressing issues like metadata interoperability

Questions?

Nicole Contaxis, MLIS
Lead, NYU Data Catalog
nicole.contaxis@nyulangone.org