Building a Data Catalog
Promoting Data Reuse and Collaboration at an Academic Medical Center

Kevin Read, MLIS, MAS
Alisa Surkis, PhD, MLIS
EXTERNAL DATASETS

- U.S. Census Data
- National Health Surveys
EXTERNAL DATASETS

- U.S. Census data
- National Health Surveys

INTERNAL DATASETS

- Institutional research data
• Population Health
• IT
• Clinical and Translational Science Institute (CTSI)
• DataCore
• Population Health
• IT
• Clinical and Translational Science Institute (CTSI)
• DataCore

NEW YORK UNIVERSITY

• Global Institute of Public Health
• Neuroscience
• Bioengineering
• Dentistry
• Nursing
NYU Data Catalog

REASONS FOR IMPLEMENTATION
Heavy users of large, external datasets (e.g. census, national health surveys, Medicare)

Department of Population Health
Heavy users of large, external datasets (e.g. census, national health surveys, Medicare)

Lack of knowledge about institutional licenses

Department of Population Health
Department of Population Health

- Heavy users of large, external datasets (e.g. census, national health surveys, Medicare)
- Lack of knowledge about institutional licenses
- Difficulty accessing datasets
- Difficulty working with datasets
NYU Data Catalog

THE DATA CATALOG
The Data Catalog

http://datacatalog.med.nyu.edu/
Current population health focus

• Maintaining functionality as catalog expands in scope

• Increase in subject domains
Centers for Medicare and Medicaid Services

Description
Centers for Medicare and Medicaid Services (CMS) offers researchers and other health care professionals a broad range of quantitative information on programs, from estimates of future Medicare and Medicaid spending to enrollment, spending, and claims data, and a broad range of consumer research to help its partners and staff. CMS also conducts demonstration projects to explore alternative policies of health care coverage and delivery. National health expenditure data not specific to the Medicare or Medicaid programs are also included.

Publisher
United States. Centers for Medicare and Medicaid Services

Timeframe
2002 - present

Geographic Coverage
United States

Local Expert
Caroline Baum
Saul Blecker
Heather Gold
Keith Goldfield
Danil Makarov
Todd Mjnovich
James Slover

Subject Matter
Health Care Costs
Delivery of Health Care

Keywords
Medicare
Medicaid
Health expenditures

Access URL
Link

Access Restrictions
Free to all
Fee/Application required

Access Instructions
The CMS DataNavigator provides users with the ability to search CMS data by program, setting/type of care, topic, geography and document type. Some of the CMS data available through the DataNavigator is freely available to download from the website. Each freely available dataset has an associated policy statement, support contact information, and downloadable .ZIP files. Multiple dataset descriptions can also be batch downloaded into .xlsx files. For restricted use datasets, users must fill out a CMS Access Request Form in order to gain access to the data and understand the privacy and security restrictions when using these restricted CMS data files. Restricted or limited use datasets are also a payed service and requests will be submitted through pay.gov.

Dataset Type
Administrative Data

Dataset Format(s)
PDF, RTF, Microsoft Word, Microsoft Excel, CSV, XML

Dataset Size
15-36 MB

PubMed Search
Articles using this dataset

Related Datasets
Surveillance, Epidemiology, and End Results Program

Related Dataset - Note
SEER-Medicare linked database available here: http://applitedresearch.cancer.gov/seermedicare/
Centers for Medicare and Medicaid Services
Alternate Title(s): CMS
UID: 10005

Description
Centers for Medicare and Medicaid Services (CMS) offers researchers and other health care professionals a broad range of quantitative information on programs, from estimates of future Medicare and Medicaid spending to enrollment, spending, and claims data, and a broad range of consumer research to help its partners and staff. CMS also conducts demonstration projects to explore alternative policies of health care coverage and delivery. National health expenditure data not specific to the Medicare or Medicaid programs are also included.

Publisher
United States. Centers for Medicare and Medicaid Services

Timeframe
2002 - present

Geographic Coverage
United States

Local Expert
Caroline Baum
Saul Blecker
Heather Gold
Keith Goldfield
Danil Makarov
Todd Mjanovitch
James Slover

Subject Matter
Health Care Costs
Delivery of Health Care

Keywords
Medicare
Medicaid
Health expenditures

Access URL
Link

Access Restrictions
Free to all
Fee/Application required

Access Instructions
The CMS DataNavigator provides users with the ability to search CMS data by program, setting/type of care, topic, geography and document type. Some of the CMS data available through the DataNavigator is freely available to download from the website. Each freely available dataset has an associated policy statement, support contact information, and downloadable .ZIP files. Multiple dataset descriptions can also be batch downloaded into excel files. For restricted use datasets, users must fill out a CMS Access Request Form in order to gain access to the data and understand the privacy and security restrictions when using these restricted CMS data files. Restricted or limited use datasets are also a payed service and requests will be submitted through pay.gov.

Dataset Type
Administrative Data

Dataset Format(s)
PDF, RTF, Microsoft Word, Microsoft Excel, CSV, XML

Dataset Size
15-36 MB

PubMed Search
Articles using this dataset

Related Datasets
Surveillance, Epidemiology, and End Results Program

Related Dataset - Note
SEER-Medicare linked database available here: http://appliedresearch.cancer.gov/seernedicare/
Centers for Medicare and Medicaid Services

Description
Centers for Medicare and Medicaid Services (CMS) offers researchers and other health care professionals a broad range of quantitative information on programs, from estimates of future Medicare and Medicaid spending to enrollment, spending, and claims data, and a broad range of consumer research to help its partners and staff. CMS also conducts demonstration projects to explore alternative policies of health care coverage and delivery. National health expenditure data not specific to the Medicare or Medicaid programs are also included.

Publisher
United States. Centers for Medicare and Medicaid Services

Timeframe
2002 - present

Geographic Coverage
United States

Local Expert
Caroline Baum
Saul Biecker
Heather Gold
Keith Goldfield
Danil Makarov
Tod Mjano
James Slover

Subject Matter
Health Care Costs
Delivery of Health Care

Keywords
Medicare
Medicaid
Health expenditures

Access URL
Link

Access Restrictions
Free to all
Fee/Application required

Access Instructions
The CMS Data Navigator provides users with the ability to search CMS data by program, setting/type of care, topic, geography and document type. Some of the CMS data available through the Data Navigator is freely available to download from the website. Each freely available dataset has an associated policy statement, support contact information, and downloadable ZIP files. Multiple dataset descriptions can also be batch downloaded into Excel files. For restricted use datasets, users must fill out a CMS Access Request Form in order to gain access to the data and understand the privacy and security restrictions when using these restricted CMS data files. Restricted or limited use datasets are also a payed service and requests will be submitted through pay.gov.

Dataset Type
Administrative Data

Dataset Format(s)
PDF, RTF, Microsoft Word, Microsoft Excel, CSV, XML

Dataset Size
15-36 MB

PubMed Search
Articles using this dataset

Related Datasets
Surveillance, Epidemiology, and End Results Program

Related Dataset - Note
SEER-Medicare linked database available here; http://appliedresearch.cancer.gov/seermedicare/
# NIH Data Sharing Repositories

This table lists NIH-supported data repositories that accept submissions of appropriate data from NIH-funded investigators (and others). Also included are resources that aggregate information about biomedical data and information sharing systems. The table can be sorted according to name and by NIH Institute or Center and may be searched using keywords so that you can find repositories more relevant to your data. Links are provided to information about submitting data to and accessing data from the listed repositories. Additional information about the repositories and points-of-contact for further information or inquiries can be found on the websites of the individual repositories.

<table>
<thead>
<tr>
<th>IC</th>
<th>Repository Name</th>
<th>Repository Description</th>
<th>Data Submission Policy</th>
<th>Access to Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI</td>
<td>The Cancer Imaging Archive (TCIA)</td>
<td>The Cancer Imaging Archive (TCIA) is a large archive of medical images of cancer accessible for public download. All images are stored in DICOM file format. The images are organized as &quot;Collections&quot;, typically patients related to a common disease (e.g. lung cancer), image modality (MRI, CT, etc) or research focus.</td>
<td>How to Submit Data to TCIA</td>
<td>How to Access TCIA Data</td>
</tr>
<tr>
<td>NCI (NHGRI, NIGMS)</td>
<td>PeptideAtlas</td>
<td>PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences.</td>
<td>How to Submit Data to PeptideAtlas</td>
<td>How to Access PeptideAtlas Data</td>
</tr>
<tr>
<td>NHGRI</td>
<td>FlyBase: A Drosophila Genomic and Genetic Database</td>
<td>Drosophila Genomic and Genetic database that includes proteomics data, microarrays and Tiling BAC's.</td>
<td>How to Submit Data to FlyBase</td>
<td>How to Access FlyBase Data</td>
</tr>
<tr>
<td>NHGRI</td>
<td>The Zebrafish Model Organism Database (ZFIN)</td>
<td>ZFIN serves as the zebrafish model organism database. It aims to: a) be the community database resource for the laboratory use of zebrafish, b) develop and support integrated zebrafish genetic, genomic and developmental information, c) maintain the definitive reference data sets of zebrafish research information, d) to link this information extensively to corresponding data in other model organism and human databases, e) facilitate the use of zebrafish as a model for human biology, and f) serve the needs of the research community.</td>
<td>How to Submit Data to ZFIN</td>
<td>How to Access ZFIN Data</td>
</tr>
<tr>
<td>NHGRI</td>
<td>WormBase</td>
<td>WormBase is an international consortium of biologists and computer scientists dedicated to providing the research community with accurate, current, accessible information concerning the genetics, genomics and biology of C. elegans and related nematodes.</td>
<td>How to Submit Data to WormBase</td>
<td>How to Access WormBase Data</td>
</tr>
</tbody>
</table>
Existing metadata schemas

• W3C Data Catalog Vocabulary
• Dryad Metadata Schema
• DataCite Metadata Schema
• NPG’s Scientific Data
Strategy over purity

• Need to describe both external and internal datasets

• PubMed: search link vs article link

• Local Experts = external data only

• Authorship
Give the people what they want

• Focus on external datasets first
• Deal with internal datasets in the future
• Add functionality wherever possible
Building a community of users

- External datasets form a strong core
- Initial internal datasets will be within similar domains to external datasets
  - Department of Population Health
- Aiming for scope creep
  - Clinical research data
  - Social science data
  - Basic science data down the road…
FUTURE DIRECTIONS
Promotion and Scalability

PROMOTION
• Integrated with other systems  
  (e.g. Main Library Catalog, Profiles RNS)
• Linked to from multiple websites  
  (e.g. Library, CTSI, DataCore)

SCALABILITY
• Researcher upload of datasets
• More curators?
Maintenance

• Add researcher information feeds
  • Faculty Data Mart
  • HR feeds from main campus

• Add archival record type

• ORCID for all researchers by Spring 2015
Exciting/New Opportunities

• Collaborate with NYU main campus librarians
  • Ingest datasets from their researchers
  • Display catalog entries in main library catalog

• Integrate catalog entries into Profiles Research Network software

• Align data catalog efforts with BioCADDIE national Data Discovery Index development

• Collaborate with DataCore to describe clinical research datasets and EHR analysis datasets
DataCore

NYULMC DataCore

Clinical Research Data Management
- Single Center Studies
- Multicenter Studies

Accessing Clinical Data for Research; Data Warehouse
- Clinical Data
- Claims Data
- Omics Data

Data Set Curation
- External Data Sets
- Internal Data Sets
Clinical Research Data Flow
Thank you!

kevin.read@nyumc.org
alisa.surkis@nyumc.org