



Improving the Discovery of Health Data in a Domain Repository

Margaret C. Levenstein



ICPSR



- ➤ Founded in 1962 by 22 universities, now consortium of 800 institutions world-wide
- Focus on social and behavioral science data, broadly defined
- Current holdings
 - ➤ 10,000 studies, quarter million files
 - > 1500 are restricted studies, almost always to protect confidentiality
 - ➤ Bibliography of Data-related Literature with 75,000 citations
- ➤ Approximately 60,000 active MyData ("shopping cart") accounts
- ➤ NIA, NICHD, and NIDA sponsored repositories
 - > Others like HMCA (RWJF) also have significant health data

What is Data Curation?

- ➤ Curation, from the Latin "to care," is the process used to add value to data, maximize access, and ensure long-term preservation
- > Data curation is akin to work performed by an art or museum curator.
 - ➤ Data are organized, described, cleaned, enhanced, and preserved for public use, much like the work done on paintings or rare books to make the works accessible to the public now and in the future



Data Documentation Initiative

Metadata standard developed and led by ICPSR

- Preservation
- Codebook creation
- Data discovery

6,600+ studies have DDI at variable level
2900 studies have question text

```
- <codeBook version="1.2.2" ID="ICPSR22626">
  -<docDscr>
    -<citation>
      -<titlStmt>
        -<titl>
            Metadata record for India Human Development Survey (IHDS), 2005
        </titlStmt>
      --cprodStmt>
        ---cproducer abbr="ICPSR">
            <ExtLink URI="http://www.icpsr.umich.edu/images/icpsr-logo.gif" title="ICPSR Logo" role="image"/>
            Inter-university Consortium for Political and Social Research
            <ExtLink URI="http://www.icpsr.umich.edu/ICPSR/" title="URL of ICPSR Web Site"/>
        -<copyright>
            ICPSR metadata records are licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License
            <ExtLink URI="http://creativecommons.org/licenses/by-nc/3.0/us/" title="Link to full text of license"/>
          </copyright>
        -<verStmt>
          <version date="2012-12-30">2012-12-30</version>
        </re>
        <holdings URI="http://www.icpsr.umich.edu/icpsrweb/ICPSR/ddi2/studies/22626"/>
      </citation>
    </docDscr>
  -<stdvDscr>
    -<citation>
      -<titlStmt>
          <titl>India Human Development Survey (IHDS), 2005</titl>
          <IDNo agency="ICPSR">22626</IDNo>
          <IDNo agency="CrossRef">10.3886/ICPSR22626.v8</IDNo>
        </titlStmt>
      -<rspStmt>
          <a href="AuthEnty affiliation="University of Maryland">Desai, Sonalde</a>/AuthEnty>
          <a href="AuthEnty affiliation="University of Maryland">Vanneman, Reeve</a>/AuthEnty>
        - < AuthEnty affiliation="National Council of Applied Economic Research, New Delhi">
            National Council of Applied Economic Research, New Delhi
         </AuthEnty>
```

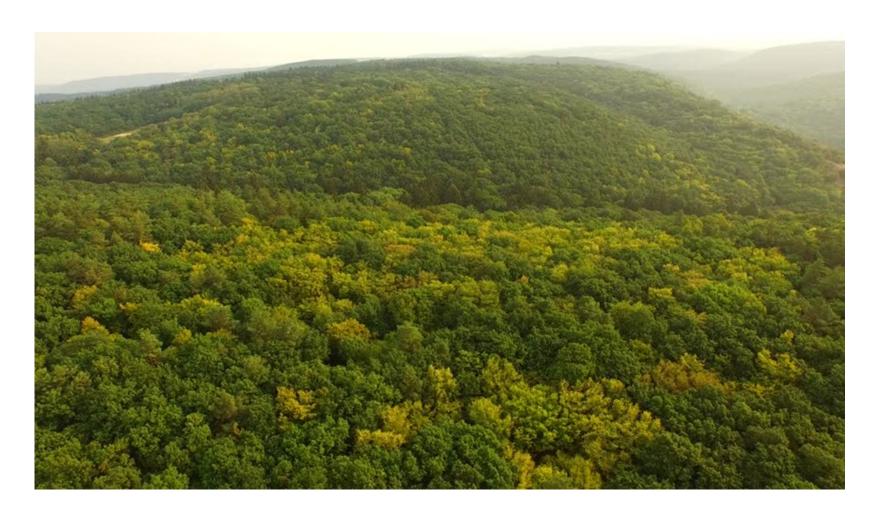
Data jumble

Whilethemetadataapplicationismanifoldcoveringalargevarietyoffieldsthereares pecializedandwellacceptedmodelstospecifytypesofmetadataBretherton&Singley(1994)distinguishbetweentwodistinctclassesstructural/controlmetadataandguidemetadataStructuralmetadatadescribesthestructureofdatabaseobjectssuchastablescolumnskeysandindexesGuidemetadatahelpshumansfindspecificitemsandareusuallyexpressedasasetofkeywordsinanaturallanguageAccordingtoRalphKimballmetadatacanbedividedinto2similarcategoriestechnicalmetadataandbusinessmetadataTechnicalmetadatacorrespondstointernalmetadataandbusinessmetadatacorrespondstoexternalmetadataKimballaddsathirdcategoryprocessmetadataOntheotherhandNISOdistinguishesamongthreetypesofmetadatadescriptivestructuralandadministrative

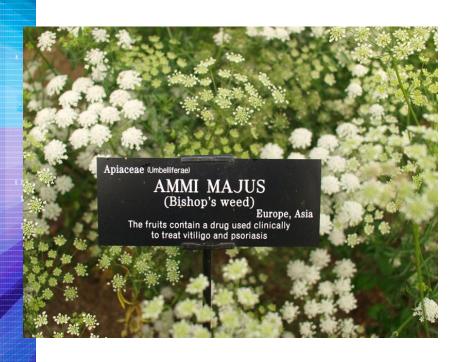
Metadata is like punctuation

While the metadata application is manifold, covering a large variety of fields, there are specialized and well-accepted models to specify types of metadata. Bretherton & Singley (1994) distinguish between two distinct classes: structural/control metadata and guide metadata. Structural metadata describes the structure of database objects such as tables, columns, keys and indexes. Guide metadata helps humans find specific items and are usually expressed as a set of keywords in a natural language. According to Ralph Kimball metadata can be divided into 2 similar categories: technical metadata and business metadata. Technical metadata corresponds to internal metadata, and business metadata corresponds to external metadata. Kimball adds a third category, process metadata. On the other hand, NISO distinguishes among three types of metadata: descriptive, structural, and administrative.

Aerial view



Tree placards





What metadata can do

Like arboretums, greenhouses, and conservatories can have trees and plants organized by types, metadata can be the way to organize, describe, identify and define data for discovery and identification.

Bibliographical grouping

- Study Title
- Alternate Title
- Principal Investigators (Individual and Organizational)
- Distributors
- Publication Date
- Funding Agencies
- Version

Scope of Study grouping

- Summary
- Subject Terms
- Geographic Coverage
- Study Time Period
- Collection Date
- Universe
- Data Type
- Collection Note

Access/analysis grouping

- Purpose of the Study
- Study Design
- Description of Variables
- Sampling
- ➤ Time Method (Cross-sectional, longitudinal/panel, repeated cross-sectional (trend))
- Data Source
- Collection Mode (self-administered, interviewer-assisted, mixed-mode)

- Weights
- Response Rates
- Scales
- Unit of Observation
- Geographic Unit

Variable level metadata

- ➤ Variable name
- ➤ Variable description
- ➤ Question text
- ➤ Possible values and definitions

➤In this project, we enhanced this variable level metadata with descriptors (tags) from CDEs and ontologies relevant to health outcomes

Problem

- ➤ Researchers looking for data run into two problems
 - ➤ Can't find data that measures what they are interested in
 - They come to you asking for \$\$ for new data collection (which can never tell you what happened in the past)
 - Find so much data they are frustrated with search and just go back to the same old dataset they already know

Example

- Researcher looking for data to study social networks and teen drug use among Native youth, specifically opioids, and writing a grant proposal to fund the research.
- ➤ Searches: opioid, friends, age, ethnicity
- ➤ Searching on ICPSR, 193 studies
- ➤ Searching NAHDAP, 124 studies
- ➤ But the best study, Drug Use Among Young American Indians, isn't there
 - ➤ It asks about heroin, not opioids

What about the cool new Google dataset search?

Google Dataset Search Q opioid friends age ethnicity X

Your search - opioid friends age ethnicity - did not match any datasets. Suggestions:

- · Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Learn how you can add new datasets to our index.

Goals of the Project

- ➤ Enhance the variable-level metadata of studies and improve variables' discoverability
- ➤ Evaluate the usefulness of alternative systems for classifying data to improve discoverability
- Increase the size of "gold standard" hand-curated data available to estimate machine learning models for automatically tagging data in the future

What is a Common Data Element?

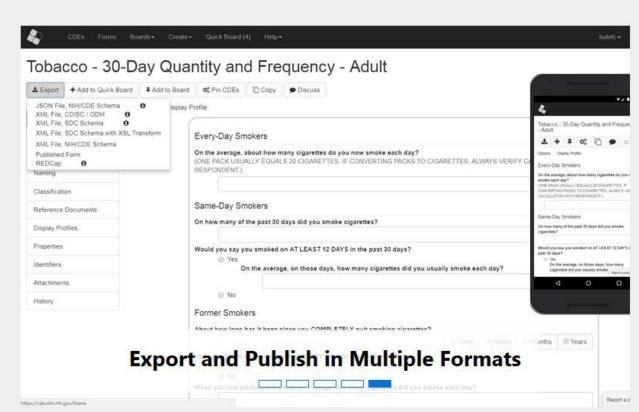
A data element common to multiple data sets across different studies. NIH encourages the use of CDEs in research and patient registries to improve data quality and promote data sharing. The National Library of Medicine hosts the NIH CDE Repository.



The NIH Common Data Elements (CDE) Repository has been designed to provide access to structured human and machine-readable definitions of data elements that have been recommended or required by NIH Institutes and Centers and other organizations for use in research and for other purposes.

Visit the NIH CDE Resource Portal for contextual information about the repository.

http://cde.nlm.nih.gov/



What is an ontology?

Defines a common vocabulary for researchers to share information in a domain, including machine-readable definitions of basic concepts in the domain and relationships, e.g., Global Mental -> Mental Health -> Substance Use – Appeal.



PROMIS® (Patient-Reported Outcomes Measurement Information System) is a set of person-centered measures that evaluates and monitors physical, mental, and social health in adults and children. It can be used with the general population and with individuals living with chronic conditions.

Returned 10 matching items for Domain = Substance Use, System = PROMIS

× Reset

Name LAZ	Domain	Measurement System	Measure Type
PROMIS Bank v1.0 - Appeal of Substance Use (Past 3 months)	Substance Use	PROMIS	Computer Adaptive Test/Item Bank
PROMIS Bank v1.0 - Appeal of Substance Use (Past 30 days)	Substance Use	PROMIS	Computer Adaptive Test/Item Bank
PROMIS Bank v1.0 - Prescription Pain Medication Misuse	Substance Use	PROMIS	Computer Adaptive Test/Item Bank

http://www.healthmeasures.net/explore-measurement-systems/promis

Overall Strategy

- ➤ Select datasets and use cases to conduct pre-test of studies and variables using associated search terms
- ➤ Identify related NLM CDE and ontology terms for variables in datasets
- Add CDEs and ontology terms to variable metadata using new tagging tool
- ➤ Analyze inter-rater reliability
- ➤ Conduct post-test to evaluate improvement of search results

Use Cases

- A researcher studying social networks and teen drug use among Native youth, specifically opioids, and writing a grant proposal to fund the research. Search terms: opioid, friends, age, ethnicity.
- ➤ Student looking for facts for a paper on drug use and school performance. Search terms: drug use, grades, school, achievement.
- ➤ Media looking for facts for a story on trends in HIV rates among drug users. Search terms: HIV, drugs.
- ➤ Government or policy worker looking for factual guidance (e.g., is maternal drug use related to infant outcomes). Search terms: drugs, infant health.

Selected Studies

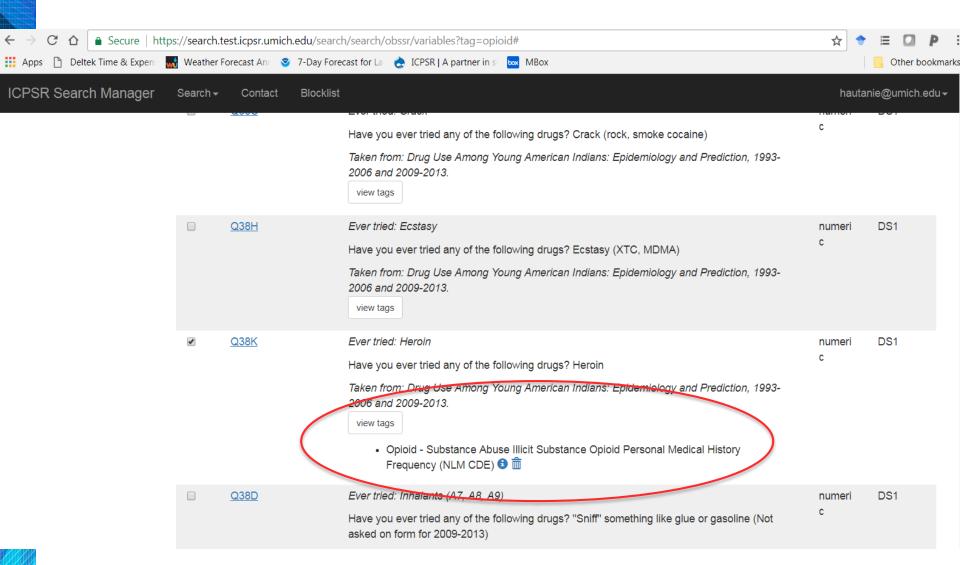
- ➤ Drug Use Among Young American Indians: Epidemiology and Prediction, 1993-2006 and 2009-2013 (ICPSR 35062) YAI: https://doi.org/10.3886/ICPSR35062
- Schools and Families Educating (**SAFE**) Children Study [Chicago, IL]: 1997-2008 (ICPSR 34368) https://doi.org/10.3886/ICPSR34368
- ➤ California Families Project [Sacramento and Woodland, California] [Restricted-Use Files] (ICPSR 35476) (**CFP**) https://doi.org/10.3886/ICPSR35476
- ➤ Maternal Lifestyle Study in Four Sites in the United States, 1993-2011 (ICPSR 34312) (MLS) https://doi.org/10.3886/ICPSR34312.v9

Tagging Tool

Time Period

ICPSR Search Manager Contact Blocklist Feedback/Problem Synonyms The search engine re-indexes at 1pm and 7pm; tags will not appear in the search facets immediately after you apply them. Tags will, however, appear if you use the "view tags" button, which queries the database directly. Filters Search Results > Scope 1 results. public = true q38k ARCHIVE = search tips OBSSR Studies (1) Publications (0) Variables (1) **Series** > Select.... Actions...▼ Study Var. Name Label/Question Text Compare Drug Use Among V Ever tried: Heroin Tag Young Have you ever tried any of the following drugs? Heroin, American Indians: Taken from: Drug Use Among Young American Indians: Epidemiology and Prediction, 1993-2006 and 2009-2013. Epidemiology view tags and Prediction, 1993-2006 and 2009-2013

Tagging Tool Verifies Tag Added



Inter-rater Reliability Results

	Total Vars	Vars In-Scope	Curator 1 agrees with Curator 2	Curator 1 agrees with metadata expert	Curator 2 agrees with metadata expert	All three agree		
CDE: All In-Scope								
CFP	918	379	86.3	88.1	88.1	88.1		
SAFE	377	142	96.5	97.2	98.6	96.5		
YAI	533	184	89.1	13.0	10.3	8.2		
Ontology: All In-Scope								
CFP	918	379	31.9	65.7	56.7	29.6		
SAFE	377	142	33.1	47.2	69.7	28.9		
YAI	533	184	32.6	69.0	51.1	27.7		

Pre- and Post-Test Results

Drug Abuse Use Case	Search terms	Pre-test	Post-test	% increase
1	opioid, friends, age, ethnicity			
	opioid	3	14	366.67
	friends	1712	1738	1.52
	age	1150	1228	6.78
	ethnicity	120	223	85.83
2	drug use, grades, school, achievement			
	drug use	445	446	0.22
	grades	618	651	5.34
	school	2367	2991	26.36
	achievement	332	332	0
3	HIV, drugs			
	HIV	69	101	46.38
	drugs	1858	2301	23.84

What did we learn?

- ➤ Tagging Common Data Elements
 - ≻helps, but ...
 - >CDEs may not exist for relevant domains
 - ➤ CDEs may not map into measures used before they were introduced
- ➤ Will be very helpful for harmonizing measurement moving forward
- ➤ Need to map to more aggregate concepts to improve discoverability of existing data

What is to be done?

- ➤ Tag with CDE domain
 - ➤ Picks up more variables
 - > Reduces problems with inter-rate reliability
 - ➤ Improves discoverability
 - ➤ Unless the researcher is really looking for prior examples of a particular CDE
 - > Too many results still returned
- ➤ More tagging with other nomenclatures
 - >PROMIS
 - ➤ Patient Reported Outcomes Measurement Information System

Leveraging tagging tool

- Can be customized for use with different topical datasets
 - >Select ontologies that build ground truth
 - ➤ Select ontologies that translate across decades
 - ➤ Select ontologies that translate across disciplines
- ➤ Can be adapted for use by experts or nonexperts
 - ➤ New NSF-funded experiments to solicit metadata enhancements from domain experts and non-experts

Conclusions

- ➤NIH invests a *lot* in data collection
- ➤NIH requires data sharing and preservation
- ➤ Effective data re-use requires
 - >Putting data somewhere people can find it
 - >Preserving so that it's accessible in the future
 - ➤ Curating it so that it's discoverable
 - >FAIR principles
 - > Findable, Accessible, Interoperable, Reusable

Study conclusions

- ➤ Discoverability and cost-effective tagging
- ➤ Harmonize, integrated hierarchy of CDEs
 - ➤ Engage social scientists in CDE creation
 - ➤Use experts to tag
- Tag a lot of data imperfectly
 - >Estimate recommender model