Poster presented at AcademyHealth annual research meeting

Erika Martin, Rockefeller Institute of Government and University at Albany

June 14, 2015

Minneapolis, MN

Evaluating the Quality, Usability, and Fitness of Open Health Data for Public Health Research

Erika G. Martin PhD MPH[1,2], Jennie Law MPA[2], Weijia Ran MPhil[3], Natalie Helbig PhD MPA[4], Guthrie S. Birkhead MD MPH[4,5]

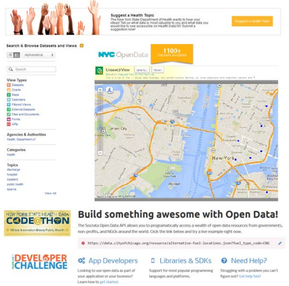
[1] Nelson A. Rockefeller Institute of Government; [2] Rockefeller College of Public Affairs & Policy, University at Albany; [3] College of Computing & Information, University at Albany; [4] New York State Department of Health; [5] School of Public Health, University at Albany-SUNY

Contact: emartin@albany.edu

Supported by a grant from the Robert Wood Johnson Foundation's Public Health Services & Systems Research Program [Grant ID#71597 to E.G.M. and G.S.B]

OPEN DATA FOR PUBLIC HEALTH SERVICES AND SYSTEMS RESEARCH (PHSSR)

- **Open government data is a new source of information for public health research**
  - Motivated by President Obama's Memorandum on Transparency and Open Government (2009)
- **Thousands of government datasets released on open data platforms meeting several "openness" criteria**
  - Publicly accessible
  - Available in non-proprietary format
  - Free of charge
  - Unlimited use and distribution rights
- **Features of open data platforms**
  - Search engines locate data objects
  - Capabilities to interact directly with data in the platform
  - Offer incentives and resources for developers
  - Opportunities to submit ideas for new datasets
  - Provide user feedback



RESEARCH QUESTIONS

- To what extent are open health data **usable** and **fit** for public health research?
- How could government agencies improve the **quality** of the **data** and corresponding **metadata**, to make these data more usable and fit for public health researchers and practitioners?

RESEARCH DESIGN: SYSTEMATIC REVIEW OF DATA OBJECTS

- **Systematic review of open health data offerings on federal, state, and local platforms**
  - Adapted from Institute of Medicine and Patient Centered Outcomes Research Institute guidelines for systematic literature reviews
  - Interdisciplinary research team with collective expertise in epidemiology, health services research, informatics, ontology development, digital government, database management, and the production of public health datasets.
  - Two reviewers ensured consistency in scoring
- **All data offerings examined with a structured coding guide to evaluate**
  - Data quality (intrinsic, contextual)
  - Consistency with five-star deployment
  - Adherence to Dublin Core metadata standards
  - Descriptive information
  - Platform usability

★ make your stuff available on the Web (whatever format) under an open license
★★ make it available as structured data (e.g., Excel instead of image scan of a table)
★★★ use non-proprietary formats (e.g., CSV instead of Excel)
★★★★ use URIs to denote things, so that people can point at your stuff
★★★★★ link your data to other data to provide context

http://5stardata.info/

METADATA INNOVATION Dublin Core®Metadata Initiative

http://dublincore.org/

RESEARCH DESIGN: SAMPLING PROCEDURES

- **Final selection (N=183)**
  - All New York City open Data offerings related to health (N=37) https://nycopendata.socrata.com/
  - 25% random sample of New York State open data offerings (N=71) https://health.data.ny.gov/
  - 5% random sample of federal open data offerings (N=75) https://www.healthdata.gov/
- **Systematic random sampling of data offerings**
  - Metadata from platforms scraped into three Excel spreadsheets
  - Random number generator used to select objects

FINDINGS: DESCRIPTIVE INFORMATION ABOUT DATA OFFERINGS

- **Only 25% of open data offerings are structured datasets**
- **Most offerings do not contain demographic variables commonly used in public health research**

Characteristic	NYC Open Data (city, N=38)[1]	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Primary presentation format in web browser, N (%)			
Table	17 (44.7)	17 (23.9)	12 (16.2)
Chart	--	27 (38.0)	--
Map	9 (23.7)	10 (14.1)	1 (1.4)
External file	1 (2.6)	9 (12.7)	27 (36.5)
Application programming interface	--	2 (2.8)	1 (1.4)
Query tool	4 (10.5)	2 (2.8)	8 (10.8)
Documents about data	3 (7.9)	1 (1.4)	18 (24.3)
Not viewable in a browser[2]	4 (10.5)	3 (4.2)	7 (9.5)
Ability to download data, N (%)			
Available via platform	--	10 (14.1)	--
Available via data access page	--	--	19 (25.7)
Available from both sources	32 (84.2)	56 (78.9)	23 (31.1)
Not available for download	6 (15.8)	5 (7.0)	32 (43.2)

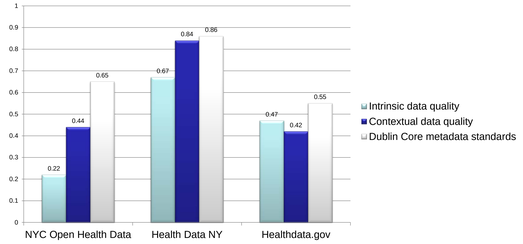
Characteristic	NYC Open Data (city, N=38)[1]	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Inclusion of demographic variables, N (%)			
Age	2 (5.3)	21 (29.6)	18 (24.3)
Gender	2 (5.3)	13 (18.3)	14 (18.9)
Race/ethnicity	2 (5.3)	8 (11.3)	10 (13.5)
Insurance status	2 (5.3)	20 (28.1)	18 (24.3)
Education	2 (5.3)	10 (14.0)	2 (2.7)
Income	7 (18.4)	5 (7.0)	8 (10.8)
Geographic identifier	17 (44.7)	45 (63.4)	28 (37.8)
Provider and/or health facilities	18 (47.4)	36 (50.7)	24 (32.4)
Size of data object,[3] median (IQR)			
Number of rows	11 (69)	161 (3340)	357 (2011)
Number of columns	6 (4)	18 (8)	11 (17)

FINDINGS: INTRINSIC AND CONTEXTUAL DATA QUALITY AND ADHERENCE TO DUBLIN CORE METADATA STANDARDS

Health Data NY scores highest on Intrinsic Data Quality measures

Health Data NY scores highest on Contextual Data Quality measures

Health Data NY scores highest on Adherence to Dublin Core Metadata Standards

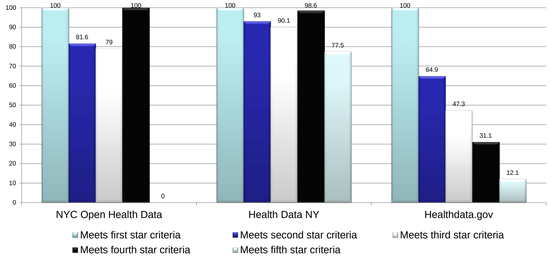


FINDINGS: GAPS IN MEETING FIVE-STAR OPEN DATA CRITERIA

All offerings met "basic web availability" open data standards (first star criteria)

Fewer met higher standards of being hyperlinked to other data (fifth star criteria)

35% of offerings met all five criteria



DISCLOSURES: Guthrie Birkhead and Natalie Helbig are employees of the New York State Department of Health, which maintains the Health Data NY open data platform reviewed in this study.

PLATFORM USABILITY

Common Features	Areas for Improvement
• **Hosting data on platforms, with external links** ***(Health Data NY, NYC Open Data)***	• **Healthdata.gov primarily serves as a search engine**
• **Handbooks to standardize metadata** ***(Health Data NY, NYC Open Data)***	• All offering hosted on external webpages
• **Multiple functions:**	• Limited interaction with data on platform
• Search and download	• Difficult to locate offerings
• Post comments and ideas	• **Technical problems limit functionality**
• Engage developers and the public with pictures, story boards, social media	• Frequent broken links (*Heatlthdata.gov)*
• **Help functions: tutorials, help email**	• Problems loading map visualizations (*NYC Open Data)*
• **Visualizations in external pages** ***(Health Data NY, NYC Open Data)***	• **No response to email queries**
	• **Low visibility on Google searches**

IMPLICATIONS FOR POLICY AND PRACTICE: IMPROVING FITNESS AND USABILITY OF OPEN DATA

- **Government agencies have little guidance on how to release open data for different user communities**
- **All three platforms have areas needing improvement, but Health Data NY scored the highest on all indices**
- **Sustained effort on improving the usability and quality of open data is necessary for improving their value for public health**
- **Future work is needed to develop standard measures of quality and usability**
  - Additional research on the factors that make some open data sites more successful
  - Development of checklists of "best practices" for open data managers

STUDY LIMITATIONS

- **New York platforms are not nationally representative**
- **Coding guide limited to simple fact-based questions**
  - Subjective nature of data quality, which depends on intended use
  - Time constraints
  - Unanticipated finding that most data objects are not tabular datasets
  - Finding that the three platforms present information in inconsistent formats and locations
- **Coding guides do not capture:**
  - Representational consistency (one aspect of platform usability)
  - Metadata consistency (one aspect of metadata quality)
- **Indices need further validation**

For more information visit: www.rockinst.org/ohdoo or http://www.publichealthsystems.org/building-sustainable-open-data-ecosystem-public-health-services-and-systems-research

The Nelson A. Rockefeller Institute of Government
The public policy research arm of the State University of New York

University at Albany
State University of New York

New York State Department of Health

Robert Wood Johnson Foundation