The National Committee on Vital and Health Statistics

Data Access and Use Group

Data Access & Use Matrix

Joshua Rosenthal, PhD
06.06.16

Definition & Goals

What is the Data Access & Use Matrix?

A working draft cataloguing the major components of and describing the information necessary to support, increase and extend the ability of different users to find, access and use data.

Information necessary to answer the following questions:

1 – How can a data consumer find data (speed, ease, accuracy, etc.)?

2 – How can a data consumer quickly tell if the data will be meaningful for their abilities and goals?

3 – How can a data producer/distributer ensure the data is meaningful to different users (usability, utility, etc.)?

4 – How can a data producer/distributer increase the reach of, and ability for a user to find, the data? (Pull vs. Push Models)

Distribution Models

Pull vs. Push Models

Pull Model - An early model of web and information structuring based around a destination site that required users to find the site and then go to it (pulling users to a specific site). Think, "Build it and they will come."

Push Model - A later model where information is pushed to users where they already are through various distribution channels. This is/was known as Web 2.0. Think, "throwing seeds to the wind to let a 1,000 flowers bloom."

Currently, thousands of users visit HHS destinations sites in a pull model, while hundreds of millions of users interact with data though distribution channels in a push model (meta-sites, data-browsers and distribution mechanism), although these are not often known or tracked (see appendix for examples and secondary and tertiary distribution mechanisms).

Ensuring that data gains a wide reach requires many of the same solutions that also make the data meaningful for users regardless of model of delivery (taxonomy, metadata, machine readability, etc.). The data access & use matrix attempts to address issues that drive both.

DATA ACCESS & USE MATRIX

Complementary Efforts

The Data Access & Use Matrix focuses on various categories of information, and information about the information including:

Data Publisher
Data Description
Data Currency/Frequency
Data Attributes
Data Usage/Adoption

It is not designed to replace standards (e.g. NCHVS standards work), or be a technical system or architecture (e.g. RWJ's recent work).

It is designed to address the specific elements around the access and use of health data beyond general data practices and universal distribution principles such as multidiscplinany guides and machine-readable clauses in legislation (e.g. US CTO / OSTP work).

DATA ACCESS & USE MATRIX

Outstanding Work & Questions

The current Data Access and Use Matrix needs to be extended and expanded and reviewed by a wide variety of user groups and experts. (It's current form is intentionally at a level to high enough to communicate the concepts without getting lost in the details but granular enough to support specific use cases.)

Key Questions:

1 – Is this work useful for data producers and distributers?

As a mechanism to inform production and distribution in order in facilitate use and access? As a guide to evaluate current data assets and distribution mechanisms and the degree to which they facilitate access & use?

2 – What is the best form/format/artifact for this work (letter, workbook, independent or conjoined with other work)?

3 – Does this work over the necessary elements to raise issues around and inform strategy and policy about the access and use of HHS data?

4 – What is the best name for this work (matrix/framework/scorecard)?

Appendix

Systems & Practices
of Data Access and Use

SYSTEMS & PRACTICES

Structures

Ecosystem Interaction

How does someone find a site / delivery mechanism or information within a site / delivery mechanism via another channel?

UI / UX

[User Interaction / User Experience]
How does someone use and experience a site / delivery mechanism?

Information Architecture

How does someone find information within a site / delivery mechanism?

Data

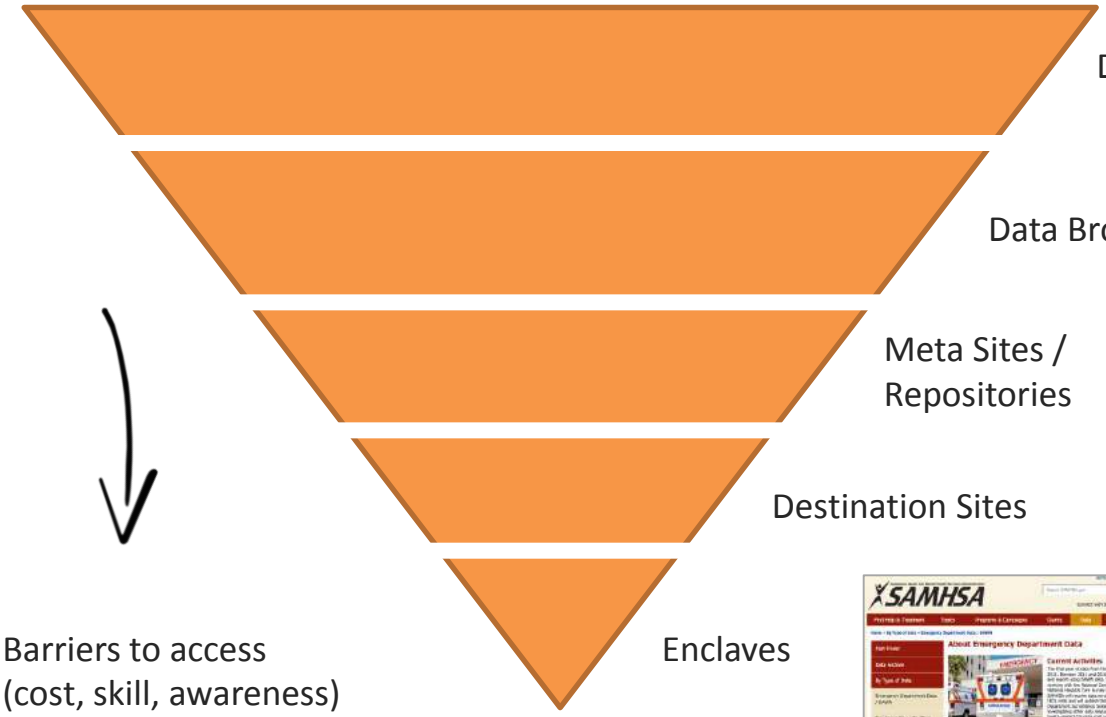
How useful and usable is the data within a site / delivery system?

Ecosystems

Ecosystem Interaction

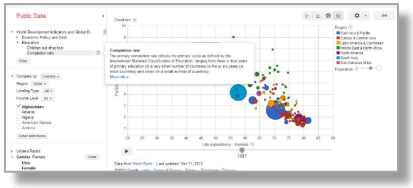
How does someone find a site / delivery mechanism or information within a site / delivery mechanism via another channel?

Example Ecosystem



Delivery Mechanisms

Data Browsers



Meta Sites / Repositories

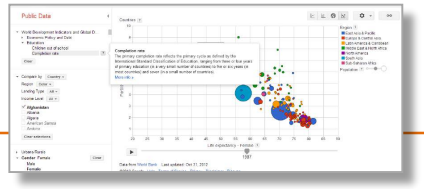
Destination Sites

Barriers to access
(cost, skill, awareness)

Enclaves

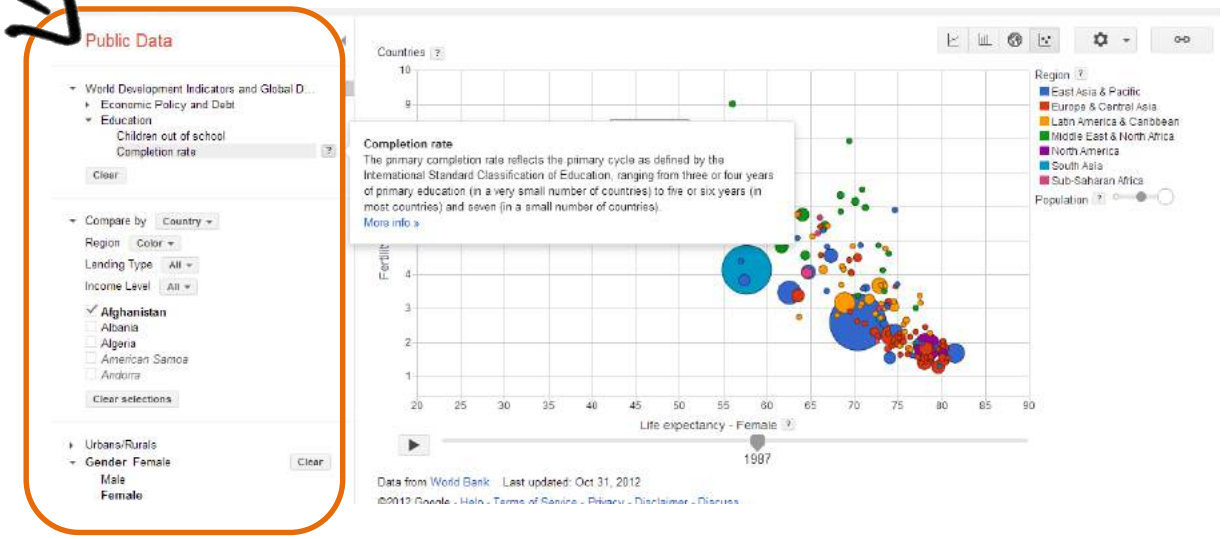
NCVHS

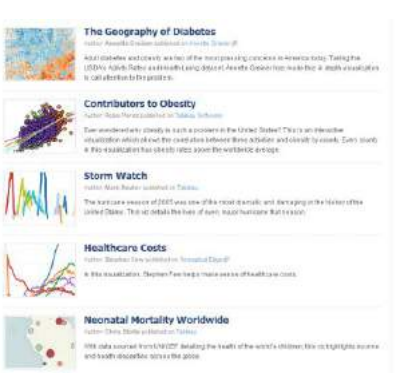
Data Browsers as Distribution

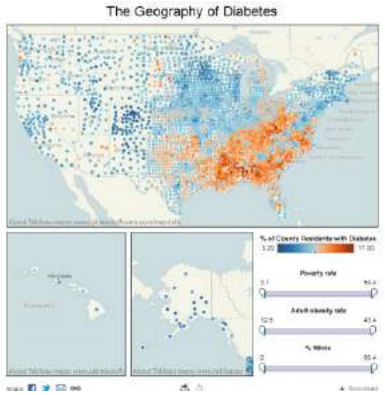


Sample: Data Browsers use HHS data
See Google Public Data Explorer and Tableau Public

Data Browsers navigate via a taxonomy (meta data), not what a top down standard has stated, but what's implicit / latent / already in the data (via scraping)







Delivery Mechanisms

HHS Data Taking on Secondary & Tertiary Distribution
Many, Many, Many Consumers/Users Using HHS Here

ProPublica repacking CMS data, distributing embedded in Yelp

yelp Official Blog
"We know just the place."

« Your Next Yelp Small Business Advisory Council Unveiled | Main | Review Federal Agencies on Yelp...and Maybe Get a Response »

August 05, 2015

Yelp's Consumer Protection Initiative: ProPublica Partnership Brings Medical Info to Yelp

Posted by Jeremy, Yelp CEO

Yelp exists to empower and protect consumers, and we're continually focused on how we can enhance our service while enhancing the ability for consumers to make smart transactional decisions along the way.

A few years ago, we partnered with local governments to launch the LIVES open data standard. Now, millions of consumers find restaurant inspection scores when that information is most relevant: while they're in the middle of making a dining decision (instead of when they're signing the check). Studies have shown that displaying this information more prominently has a positive impact.

Today we're excited to announce we've joined forces with ProPublica to incorporate health care statistics and consumer opinion survey data onto the Yelp business pages of more than 25,000 medical treatment facilities. Read more in today's Washington Post story.

ProPublica repackaging CMS data, making more useful, selling

The ProPublica Data Store

ProPublica is making available the datasets that power our data journalism. The raw data we received as the result of a FOIA request is available for free, and datasets that reflect substantial cleaning and processing by our staff are available for a one-time fee. Journalists and academic researchers can purchase premium datasets, and interested commercial users can contact us for pricing, by clicking the "Purchase" button on any dataset. We also provide a pass-through link when a data download is available on another site. Related Story »

Premium Datasets (Purchase)	APIs and Raw Data (Free)	External Data
Cleaned up, categorized and often created from multiple sources, these Premium datasets are unique to ProPublica and sold for a nominal fee.	ProPublica's APIs and other free-to-download raw datasets that we've compiled from our own research or received via FOIA request.	ProPublica frequently uses datasets that are free and available online. So instead of downloading copies from us, we send you straight to the source.

Health Datasets	Source	JOURN ($)	ACAD ($)	
Premium: Surgeon Scorecard Dataset ProPublica's Surgeon Scorecard data. *Size: 23,370 rows, Date Released: September 2015*	Centers for Medicare & Medicaid Services	$200	$2,000	PURCHASE Try a Sample!

Delivery Mechanisms

HHS Data Taking on Secondary & Tertiary Distribution
Many, Many, Many Consumers/Users Using HHS Here

RowdMap HHS data, partnering with USNEWS to help consumers

RowdMap (start up) repackaging HHS & Dartmouth data, selling to US Market (100MM patients, 48 States cf. CMMI presentation) and partnering with USNEWS

RowdMap

U.S. & WORLD REPORT News

NCVHS

This workbook highlights key attributes of data that either is or can be made available to support community health
This attributes captured in this matrix/template should be sufficient to allow different user types to make an informed decision about the usefulness of the data and insight into the degree of difficult in "consuming" it.
Focus: More on the characteristics of the data and its attributes

Version 1.2

Sources:	Rosenthal presentation to the NCVHS		
	Rippen additions, data.gov		
	NCVHS Data group comments and recommendations		
	Characteristics		
Attributes	Usable		
	Useful		
	Timely		
	Sustainability		
Dissemination	Source website		
	Data channels	Google	
		Repositories	
		Metasites	data.gov

		Title	Type	Description
Data Owner				
	Publisher	Data Owner	Organization name	Define "owner" of the data
		Contact		
		Data Publisher	Organization name	Defines who is responsible for posting the data
		Contact		Contact for external users of the data
		Data Producer/Collector	Organization name	Defines who is responsible for collecting the data
		Contact		SME to contact for more information
Theme				
	Data description	Topic	Example Categories: This links to the population health matrix categories to ensure consistency	The tags that help define the high level themes that the data describes and faciliates users find the data (e.g., clinical data, health-financial data)
			Summary focus (text)	
		Purpose	Example Categories: Research, Evaluation, Planning, Payment, Resource Allocation, other	What was the purpose that the data was collected for.
			Summary of purpose (text)	
		Audience	Example Categories: Researchers, Policy makers, Consumers, Communities, Clinicians, Government, Data distributers, Clinicians, Public Health, Educators, Insurers, Healthcare providers, Other	Who (what audience) was the data intended for?
			Details	

	Geographic Scope	Example Categories: Zip3, Local Community, City, County, State, Region, National, International, Other	What is the geographic reach of the data?
		Details	
	Funding	Mandated, Unmandated	If funding mandated by a government?
		Example Categories: Federal, State, Local, Non-Profit, Foundation, Business, Other	Who paid to collect the data, post the data?
		Details	
	Priority status	Example categories: Critical, Urgent, Important, Normal	What is the priority status associated with this data for the data owner
	Data Source	Example categories: Survey (direct), Survey (interview), Device (e.g., sensor), Activity Output (e.g., billing, arrest, EHR), Other	How was the data captured?
		Details	
Data Currency	Start date	Date - Day, month, year	First date associated with the data (start date)
	Latest Date	Date - Day, month, year	Date of latest data available
	Publication-date (of data)	Date - Day, month, year	Latest date data was published to the site
	Frequency of refresh/collection	Example categories: instantaneous, hourly, daily, weekly, monthly, quarterly, semi-annually, annually, every x years, other	How often is the data collected?

	Publication lag time	Days, months, year(s)	What is the expected lag time between data collection and when it is made available on the site.
	Expected End date	Example categories: contiuous, one time, until end date	How long will this type of data be collected?
		Date - Day, month, year	
		Details	
Data Attributes			
	Data Availability	Example categories: Open source, Request (free), Request (fee), Deidentified, Matched deidentified, Restricted, Internal Use Only, Other	Are there any barriers, restriction to the use of the data
		Details	
	How can data be obtained?	Details	How data can be obtained
	Limitations of the data	Details	What are important limitations in the data that a user should know.
	File Format	Example categories: JSON, XML, RDG, xlsx	what file format is the data in? Is it machine readable?
	Data Dictionary	Yes, No	Is there a data dictionary available?
		Status (complete/parial, N/A)	
		Machine readable?	
		Details	include a link
	ERD (entity relationship diagram)		
		Machine readable?	
		Number of entities	
		Number of attributes	
	File attributes		

		Size (mb, gb)	
		Number of records	
		Number of colums (if applicable)	
	Geocoded		
		Country, State, County, PUMA, HRR, HAS, Zip, Neighborhood Block…)	What granularity?
			Can more detail be requested?
	API		
	Verification		
About the use of the data (not core to the description)			
Usage from System based attributes (meta information captured about usage form site/location data is described or hosted)			
	number of times page desciption of data viewed		
	number of downloads		
	number of "likes"		
	average star ratings (1-5 stars)		