



TW/C

Tetherless World Constellation

# Interoperability, Standards and Linked Data: promise and challenge to SDK

Jim Hendler

Tetherless World Professor of Computer and Cognitive Science  
Assistant Dean of Information Technology and Web Science

Rensselaer Polytechnic Institute

<http://www.cs.rpi.edu/~hendler>

@jahendler (twitter)



- What can Science learn from data-sharing on the World Wide Web
- Shattering a couple of myths
  - Science is not the big kid anymore
  - Expressive ontologies are not a scalable solution
    - +1: Lightweight integration trumps ontology
    - See <http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html>
- Posing some new issues/approaches
  - Data search is critical
  - Usable technologies are transitioning
- Challenges remain
  - Which I'll let the next speaker address



Tera, Peta, zeta bytes

Tetherless World Constellation

- Yes, science uses some extremely large databases and many of them are crucial to society
- World Wide Web data is also extremely large
  - Maybe not crucial to society
- And much better “funded”
  - eg. Facebook
    - 25 Terabytes of logged data per day; valuation \$33B (NIH budget ~ \$31B)
  - eg. Google
    - In 2008 it was estimated at 20 petabytes per day (not including youTube); current valuation \$190B (about 1/3 the **entire** DoD budget)



What I'm not talking about

Tetherless World Constellation

- Annotation and semantics in papers/literature search
  - Important, but really not the key *unaddressed* problem
    - Keyword search
      - Available as a service
    - Informatics/language extraction
      - Available as open source products
        - » Application costs for known technologies is a problem, but not the research problem
        - » Most researchers are trying to invent this rather than use the off the shelf solutions
    - IBM Watson shows what Q/A will look like in a few years
    - ...
  - Note there's some interesting new work in embedding the annotations
    - (Ask Hal)



So what can we learn?

Tetherless World Constellation

- Moving away from relational models
  - cf. NoSQL
  - cf. BigData
- Moving towards interoperability and exchange
  - Example: like buttons from Facebook
  - Example: Google
- **Simple metadata and lightweight semantics are becoming a big deal**
  - <http://www.slideshare.net/jahendler/the-semantic-web-2010-update>

# Example: Government Data on the Web



**DATA.GOV**

DISCOVER. PARTICIPATE. ENGAGE.

Search the following Data.gov catalogs:

- RAW DATA CATALOG
- TOOL CATALOG
- GEODATA CATALOG

**FEATURED TOOL:**  
**U.S. GEOLOGICAL SURVEY (USGS)**  
USGS Global Visualization Viewer for Aerial and Satellite Data

Ten million archive images of the Earth's surface are available for immediate selection and free download via the USGS Earth Resources Observation and Science (EROS) Center's Global Visualization Viewer. Users can preview thumbnails, browse images and download full-image selections from 1.5 million aerial photos of U.S. sites and 8.5 million images captured worldwide by U.S. Earth-observing satellites.

[VIEW THIS TOOL](#)

### Welcome to Data.gov

The purpose of Data.gov is to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government. Although the initial launch of Data.gov provides a limited portion of the rich variety of Federal datasets presently available, we invite you to actively participate in shaping the future of Data.gov by suggesting additional datasets and site enhancements to provide seamless access and use of your Federal data. Visit today with us, but come back often. With your help, Data.gov will continue to grow and change in the weeks, months, and years ahead.

### How to use Data.gov

Data.gov includes searchable [data catalogs](#) providing access to data in three ways: through the "raw" data catalog, the tool catalog and the geodata catalog. Please note that by accessing datasets/tools offered on Data.gov, you agree to the [Data Policy](#), which should read before accessing any dataset or tool. If there are additional datasets that you would like to see included on this site, please [click here](#). For more information on how to use Data.gov, [view our tutorial](#).

DATA.GOV | [Data Policy](#) | [Accessibility](#) | [Contact Info](#) | [Privacy Policy](#)

**RECOVERY.GOV**

CHAIRMAN'S CORNER

FORESTRY PROJECT CREATES JOBS

**USA.gov**

Data and Statistics - General Reference Resources

By Organization

- All Agency Index
- Federal Government
- State Government
- Local Government
- Tribal Government

Contact Government

- Online Services
- Frequently Asked Questions
- E-mail
- Web Chat
- Phone

**FEDERAL IT DASHBOARD**

your window into the federal IT portfolio

Department of Defense



# Government Data Sharing

## Tetherless World Constellation



January 1, 2009

“Openness will strengthen our democracy and promote efficiency and effectiveness in Government.”

--- President Obama



May 21, 2009

data.gov online  
57 Data Sets



December 8, 2009

“Open Government Directive” released  
~2000 Data Sets



May 21, 2010

data.gov relaunch with semantic web featured  
>305,000 Data Sets

2009

2010 ...

June 30, 2009

Putting Govt Data online-  
Data.gov.uk beta

January 19, 2010

data.gov.uk online

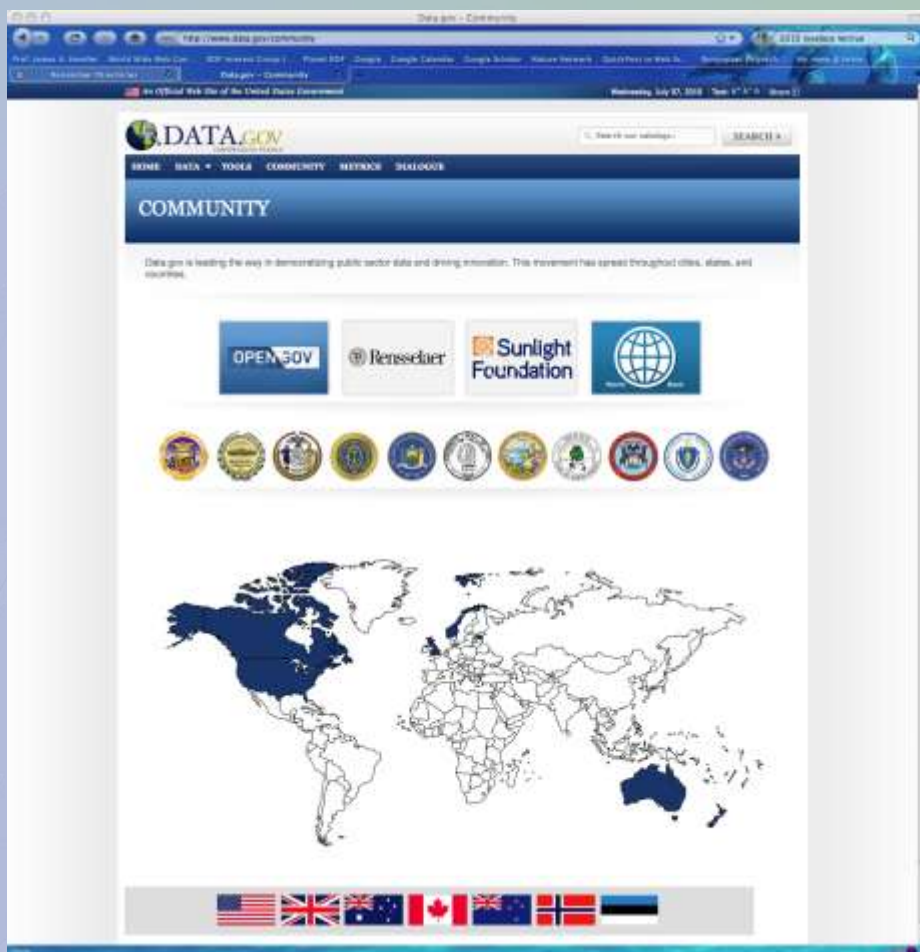
~6000 Data Set





# Data.gov community: International

## Tetherless World Constellation



### Examples:

US	305,000
Japan	30,184
Denmark	17,086
UK	6,000
Korea	833
Australia	700
World Health Org	400
Ireland	263
Catalonia	246





# Creating/Using Data "app" technologies

## Tetherless World Constellation



(a) White House visitor search



(b) US-UK Foreign Aid Comparison



(c) Agency Budget and NYTimes



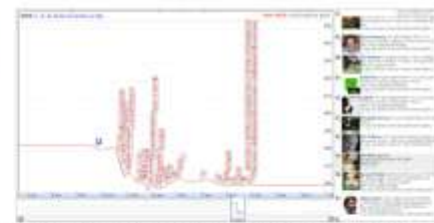
(d) Wildland fire and DBpedia



(e) [Health] Tobacco Prevalence and Correlated Factors



(f) [Policy] About Supreme Court Justices



(g) [Financial] Stock price and Twitter events



(h) [Yahoo! Pipes] World Earthquake Map



(i) [IBM ManyEyes] White House visitor network



(j) [RDFa] semantic search



(k) [RSS] data.gov updates

See more than 50 of these at <http://logd.tw.rpi.edu>

Thought: Easy to credit app downloads/usage



# Linking GDP of the US and China

## Tetherless World Constellation

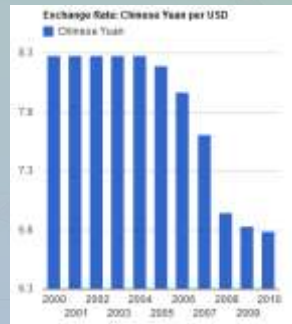
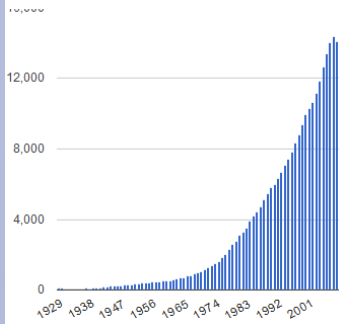


Federal Reserve Statistical Release

H.10  
Foreign Exchange Rates (Weekly)

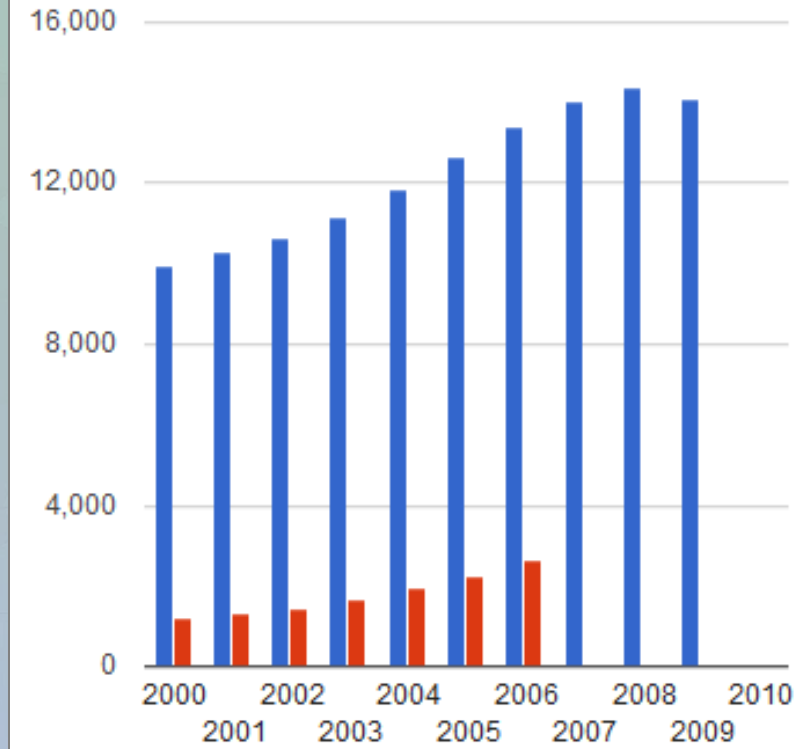


GDP of the US (Billion Dollar)

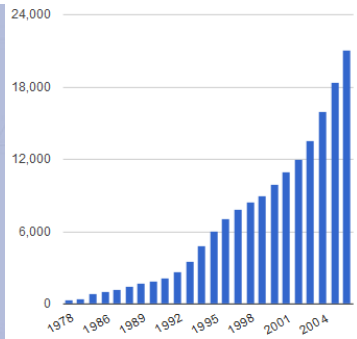


GDP (Billion USD) adjusted by Exchange Rate

■ US ■ China



GDP of China (Billion Chinese Yuan)



[Temporal Mashup] [bea.gov](http://bea.gov) + [federalreserve.gov](http://federalreserve.gov) + [stats.gov.cn](http://stats.gov.cn)



# Linking GDP of the US and China

## Tetherless World Constellation

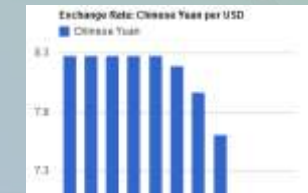
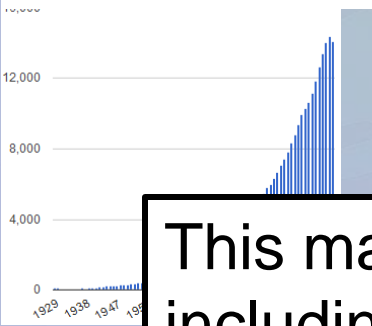


Federal Reserve Statistical Release

H.10  
Foreign Exchange Rates (Weekly)



GDP of the US (Billion Dollar)

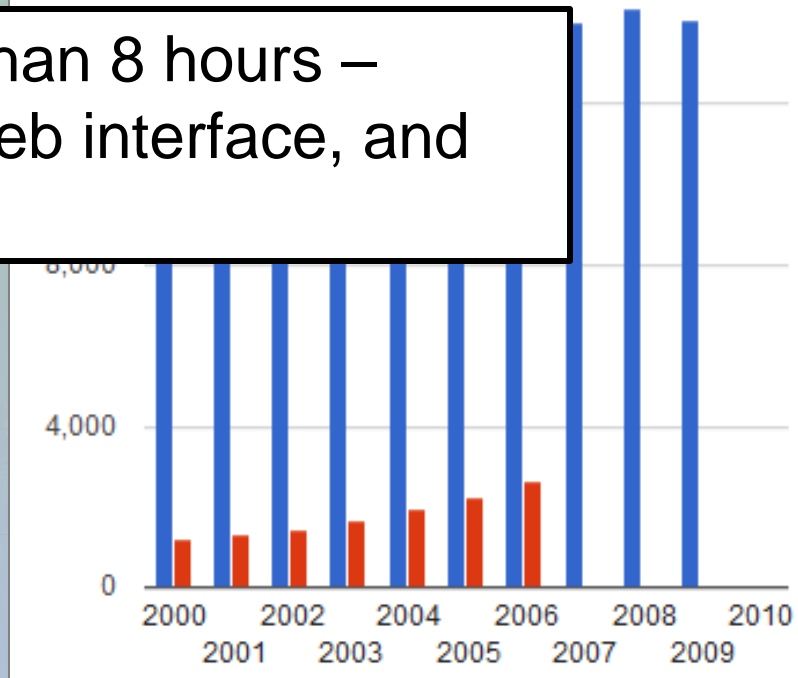


GDP (Billion USD) adjusted by Exchange Rate

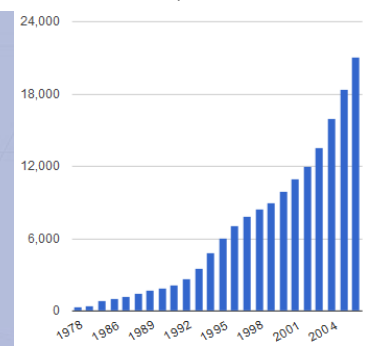
■ US ■ China

16,000

This mashup was built in less than 8 hours – including conversion of data, web interface, and visualization!



GDP of China (Billion Chinese Yuan)

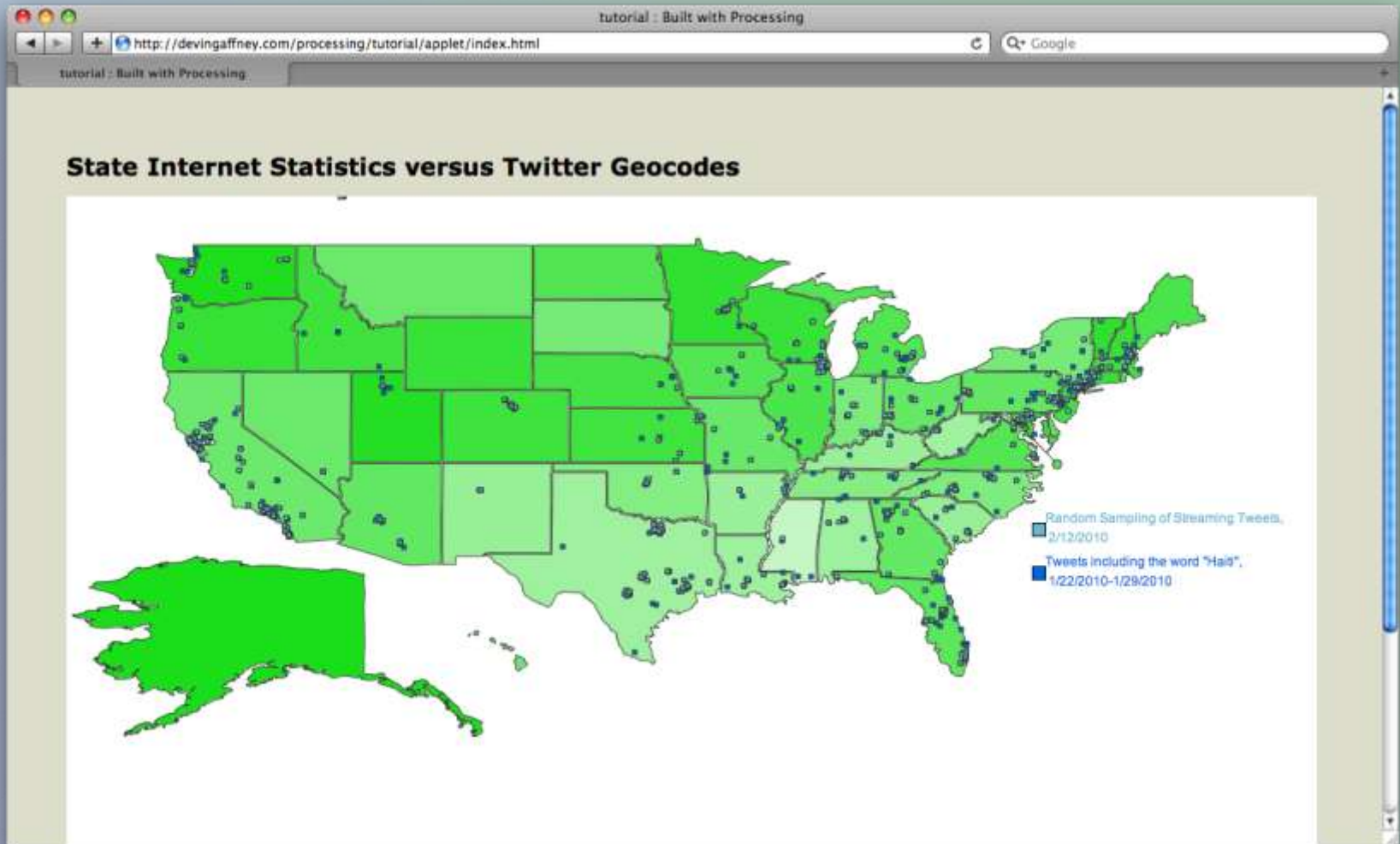


[Temporal Mashup] [bea.gov](http://bea.gov) + [federalreserve.gov](http://federalreserve.gov) + [stats.gov.cn](http://stats.gov.cn)



# Govt data linked to Social Media Metadata

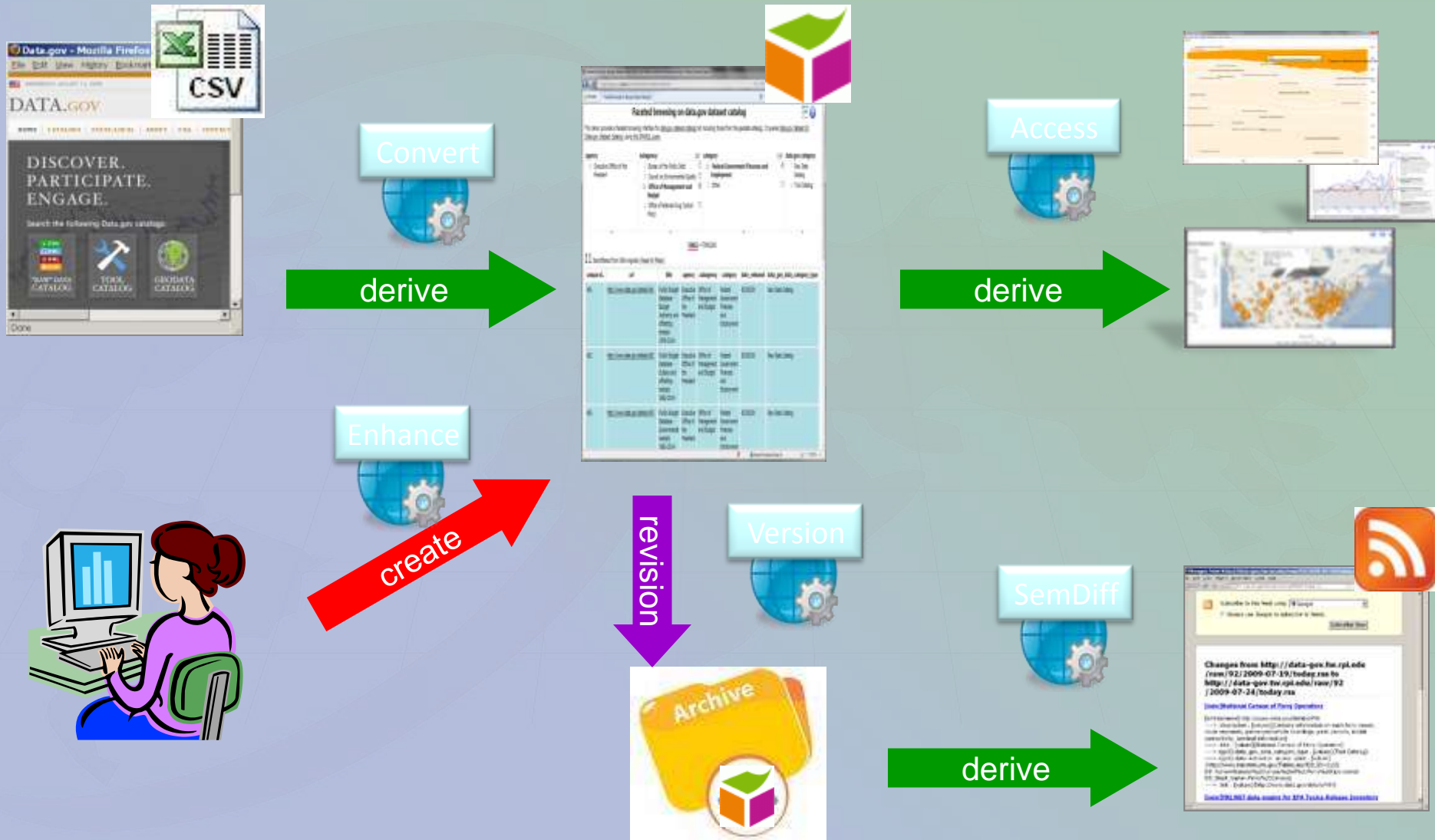
## Tetherless World Constellation





There is a lot of workflow information in the mix

## Tetherless World Constellation





**State-related Dataset Faceted Search**

**Datasets**

- 249
- 307

**States**

- 17 Alabama
- 12 Alaska
- 3 American Samoa
- 6 Arizona
- 10 Arkansas
- 9 California

**Agency**

- Environmental Protection Agency

**Category**

- Geography and Environment

2 Dataset filtered from 21 originally (Reset All Filters)

**249 (link)**

label: 249

type: Dataset

URL: <http://data.gov.tw.epi.edu/data-gov/item/249>

url: [http://data.gov.tw.epi.edu/vocab/Dataset\\_249](http://data.gov.tw.epi.edu/vocab/Dataset_249)

title: 2006 Toxics Release Inventory National data file of all US States and Territories

agency: Environmental Protection Agency

category: Geography and Environment

state: New Jersey, Connecticut, Wisconsin, Illinois, American Samoa, Mississippi, Ohio, Virginia, Louisiana, Nebraska, Idaho, Oklahoma, Delaware, Texas, Indiana, Missouri, Kansas, Guam, Massachusetts, Maryland, Kentucky, Florida, California, Hawaii, Utah, Michigan, Vermont, Iowa

2 Dataset filtered from 21 originally (Reset All Filters)

Dataset ID	URL	Title	Agency	Category	State-related Predicates
249	<a href="http://data.gov.tw.epi.edu/vocab/Dataset_249">http://data.gov.tw.epi.edu/vocab/Dataset_249</a>	2006 Toxics Release Inventory National data file of all US States and Territories	Environmental Protection Agency	Geography and Environment	<ul style="list-style-type: none"> <li>• <a href="#">(191)stack_air_emissions_release_pounds</a></li> <li>• <a href="#">(191)mailing_state</a></li> <li>• <a href="#">(191)facility_state</a></li> <li>• <a href="#">(191)total_transferred_off_site_for_further_waste_management</a></li> <li>• <a href="#">(191)other_on_site_waste_management</a></li> <li>• <a href="#">(191)total_stack_air_emissions</a></li> </ul>
307	<a href="http://data.gov.tw.epi.edu/vocab/Dataset_307">http://data.gov.tw.epi.edu/vocab/Dataset_307</a>	2007 Toxics Release Inventory National data file of all US States and Territories	Environmental Protection Agency	Geography and Environment	<ul style="list-style-type: none"> <li>• <a href="#">(191)mailing_state</a></li> <li>• <a href="#">(191)stack_air_emissions_release_pounds</a></li> <li>• <a href="#">(191)facility_state</a></li> <li>• <a href="#">(191)total_stack_air_emissions</a></li> </ul>

How can we search for data?



Metadata is crucial

## Tetherless World Constellation

keyword:

search results (note: This demo only lists the first 200 search results)

**agency**

- 1 Department of Justice
- 1 Department of Labor
- 4 Department of the Interior
- 11 DOC/NOAA/NEROIS/WDC > National Geophysical Data Center, NEROSIS, NOAA, U.S. Department of Commerce
- 21 Environmental Protection Agency
- 3 N/A
- 1 National Oceanic and Atmospheric Administration (NOAA), National Ocean Service (NOS), Office of Coast Survey (OCS)
- 11 NOAA National Oceanographic Data Center
- 1 U.S. Geological Survey
- 1 U.S. Geological Survey, Central Energy Resources Team
- 1 U.S. Geological Survey, Central Energy Resources Team, Data Management Project
- 111 US National Oceanographic Data Center

**category**

- 21 Data
- 1 boundaries
- 11 climatology/Meteorology/Atmosphere
- 2 elevation
- 12 Geography and Environment
- 11 geoscientificInformation
- 1 Health and Nutrition
- 1 Labor Force, Employment, and Earnings
- 3 Law Enforcement, Courts, and Prisons
- 2 N/A
- 2 Natural Resources

**catalog**

- 111 geodata catalog
- 111 Raw Data Catalog
- 1 Tool Catalog

200 items

dataset	agency	category	catalog	geographic coverage
Wisdom Information Office	Department of Labor	Labor Force, Employment, and Earnings	Raw Data Catalog	<a href="http://logd.fw.rpi.edu/source/data-gov/dataset/92?value-of/geographic_coverage/States_Metric_Areas">http://logd.fw.rpi.edu/source/data-gov/dataset/92?value-of/geographic_coverage/States_Metric_Areas</a>
U. S. Department of Energy (DOE) Category: Exclusion (CE) Determinations Under the National Environmental Policy Act (NEPA)	Department of Energy	Geography and Environment	Tool Catalog	<a href="http://logd.fw.rpi.edu/source/data-gov/dataset/92?value-of/geographic_coverage/U_S_and_U_S_Territories">http://logd.fw.rpi.edu/source/data-gov/dataset/92?value-of/geographic_coverage/U_S_and_U_S_Territories</a>
Total percentage of restricted use carried on commercial waterways by traffic box	Department of Defense	Transportation	Raw Data Catalog	<a href="http://logd.fw.rpi.edu/source/data-gov/dataset/92?value-of/geographic_coverage/United_States">http://logd.fw.rpi.edu/source/data-gov/dataset/92?value-of/geographic_coverage/United_States</a>
Temperature and salinity profile	NOAA National Oceanographic	oceans	Geodata Catalog	lower corner: 47.66,-158.7 upper corner: 61.07,-122.26

What kinds of metadata are: simple to create, powerful enough for search and internationalizable (esp. beyond English)



Example, integrating data and info search

Tetherless World Constellation

# Search in SciVerse Hub on Climate Change

Home | ScienceDirect | Scopus | Applications TWC.HK is logged in | User Control Panel

Home | My Settings

Search  Search ? Search Tips

Found 688464 results for ALL (Climate change)

Previous | 104 to page 10 | 4227458 | Next page >

### Search within results

### Refine Results

   
**Content Sources** A

- Scopus (198842)
- ScienceDirect (397345)
- Digital Archives (87384)
- HDL7D (48254)
- RePEc (3114)

View more >>

**Year** A

- 2011 (1467)
- 2010 (430557)
- 2009 (281717)

### My Applications

[Add | Manage Applications](#)

- 1. [Enlighten Your Vision...](#)
- 2. [Workforce Productivity](#)
- 3. [Workforce Productivity](#)
- 4. [Workforce Productivity](#)
- 5. [Workforce Productivity](#)

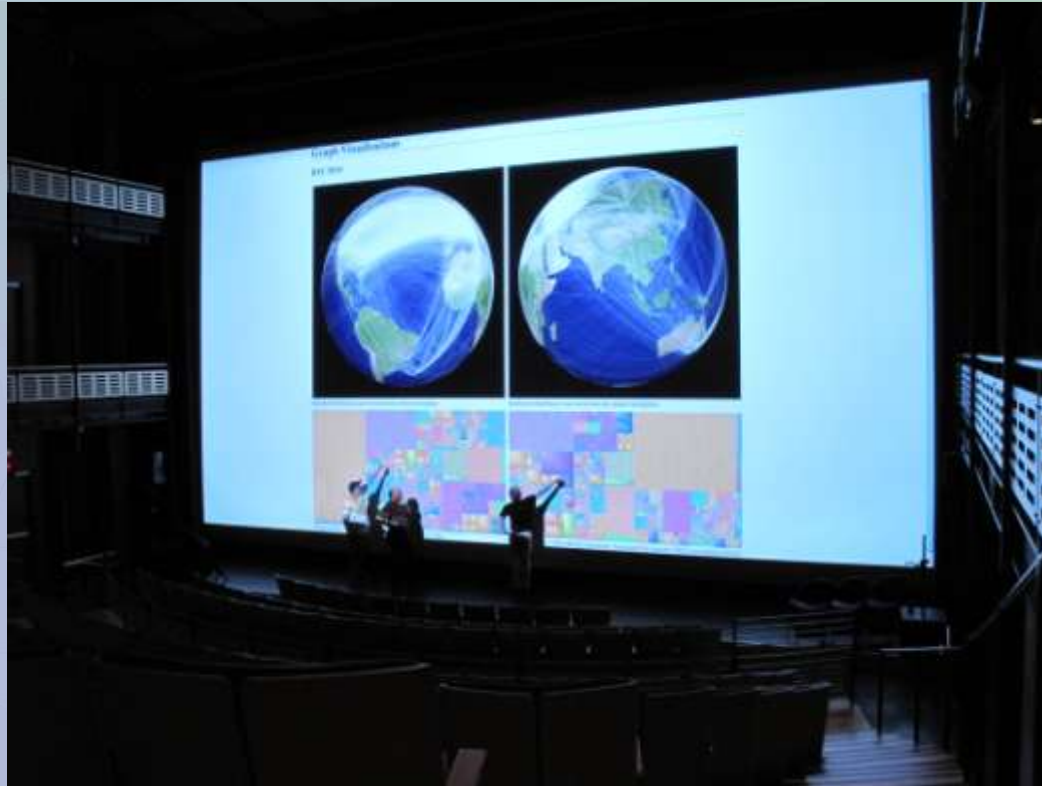
- Climate Change**  
RbUwww.greenpeace.org/usa/en/energy/... December 2009  
... Revolution Step Four Step Steps The Rainbow Home **Climate Change** What does dangerous **climate change** mean to Asia? Philippines **Climate** Impacts Solutions Welcome to Greenpeace Philippines! **Climate Change** First Step **Climate Change** is real and happening.
- IPCC Special Reports on Climate Change**  
RbUwww.grid.ac.uk/information/... October 2009  
IPCC Special Reports on **Climate Change** Ordering these reports (About the web-site Emissions Scenarios Analysis and the Global Atmosphere Land Use, Land-Use **Change** and Forests Biotechnological and Technological issues in Technology Transfer The Regional Impacts of **Climate Change**
- IPCC - Intergovernmental Panel on Climate Change**  
RbUwww.ipcc.ch/ November 2009





Challenge: Visualization can be more than art

Tetherless World Constellation



Fox & Hendler, Changing the Equation in Scientific Visualization Science, 2/11/1

How can we make data visualization a 1<sup>st</sup> class citizen in the scientific methods?



What's promising for SDK?

Tetherless World Constellation

- Linked open data (cf. [data.gov](http://data.gov), [data.gov.uk](http://data.gov.uk))
  - Being explored in Genomics, Astronomy
  - Needs a lot more attention in other scientific domains
- Markup languages and semantics and tools can enable “transparency”
  - Interdisciplinary science requires data communication
- Lower barriers to internet visualization, e.g. Google vis, MIT simile, many more...
- Web 2.0 to put people in the loop and use and contribute to annotations



What's challenging?

Tetherless World Constellation

- Finding Stuff!

- How do I find a dataset in the many out there that might be of use to me?
  - Cannot keyword search in data
  - Cannot
- How do I know what is in a large data store? In a virtual observatory? In the cloud?
  - What is the coverage?
  - What is the access?
  - Who do I need to ask for what
- What are the rules about using it?
  - What can I combine it with?
  - How do downstream users know I've combined it



# TWIC

What else

## Tetherless World Constellation

- Trust
  - Government data is controversial, and potentially biased
    - How do we confirm or dispute?
- Combination
  - When we combine data we need to keep the provenance of information (see trust)
    - How can we show and use?
- Scaling
  - Data-gov Wiki has already converted 5,448,693,510 triples
- Versioning and updating
- Archiving
- Searching
- ...



BTW

Tetherless World Constellation

# ASK ME ABOUT MY New PhD Program

RPI is considering the creation of a  
**SCIENCE INFORMATICS**  
PhD program