

# Data.gov Wiki: A Semantic Web Approach to Government Data

Li Ding, Dominic DiFranzo, Sarah Magidson,  
Alvaro Graves, James R. Michaelis, Xian Li,  
Deborah L. McGuinness, Jim Hendler

Tetherless World Constellation  
Nov 2, 2009



Rensselaer

# Synergy

- Government: data is out there “as is”
- Loop: gov data and linked data
- Loop: gov data and web developers
- Loop: gov data and end users



# Government Data on the Web

**Data.gov - Mozilla Firefox**  
File Edit View History Bookmarks Tools Help  
http://www.data.gov/

FRIDAY, AUGUST 07, 2009

## DATA.GOV

HOME | CATALOGS | STATE/LOCAL | ABOUT | FAQ | CONTACT US | SUGGEST OTHER DATASETS

### DISCOVER. PARTICIPATE. ENGAGE.

Search the following Data.gov catalogs:

- CSV XSL KML SHP
- 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.

DATA.GOV Data Policy | Accessibility | Contact Info | Privacy Policy

**Recovery.gov - Mozilla Firefox**  
http://www.recovery.gov/

## RECOVERY.GOV

HOME | ABOUT | BOARD | INVESTMENTS | OPPORTUNITIES | IMPACT | NEWS | FAQ | CONTACT US

### CHAIRMAN'S CORNER

Earl E. Devaney is chairman of the Recovery Accountability and Transparency Board. The Recovery Act spending and manages this website. His next report to the American people will focus on development of the new state-of-the-art Recovery.gov.

### FORESTRY PROJECT CREATES JOBS

The Nevada Division of Forestry has begun a \$1.3 million Recovery-Act funded project in Lincoln County, creating 26 jobs and saving three. The project will reduce hazardous forest fuels through the removal of Pinyon and Juniper trees encroaching on Highway 93 and State Route 319.

[LEARN MORE](#)

**Data and Statistics - General Reference Resources: USA.gov - Mozilla Firefox**  
http://www.usa.gov/Topics/ReferenceResources/

## USA.gov

Home | Site Index | FAQs | E-mail | Phone | Chat | Our Blog | Mobile | **Español** | Other Languages

1 (800) FED INFO | 1 (800) 333-4636

For Citizens | For Businesses and Nonprofits | For Government Employees | For Visitors to the U.S.

### Data and Statistics - General Reference Resources

[E-mail me when this page is updated](#)

- 2002 Census of Governments
- Aging Statistics
- American Factfinder
- Budget of the U.S. Government
- Budget of the U.S. Government: State-by-State Tables FY 2009
- Census Data
- Child and Family Statistics

### Featured Sites

[Data.gov](#)

**Federal IT Dashboard - Mozilla Firefox**  
http://it.usaspending.gov/

## IT DASHBOARD

HOME | INVESTMENTS | DATA FEEDS | ANALYSIS | FAQ

### your window into the federal IT portfolio

Bar chart showing IT spending by agency: DHS, HHS, Treasury, DOC, DOT, DOJ, VA, USDA, Energy, Others.

### Department of Defense

Major Investments: 62  
Spending on Major Investments: \$3.6 B (FY 2009)

Overall Rating:

Agency:  Investment Spotlight:

# Objectives

- Investigate the role of semantic web in producing, processing and utilizing government datasets
  - To enrich the value of data via normalizing, linking and information-extraction
  - To realize the value of data via applications, esp. visualization
  - To support web developers via machine friendly data access and web services

# Convert Data



 tagCloud



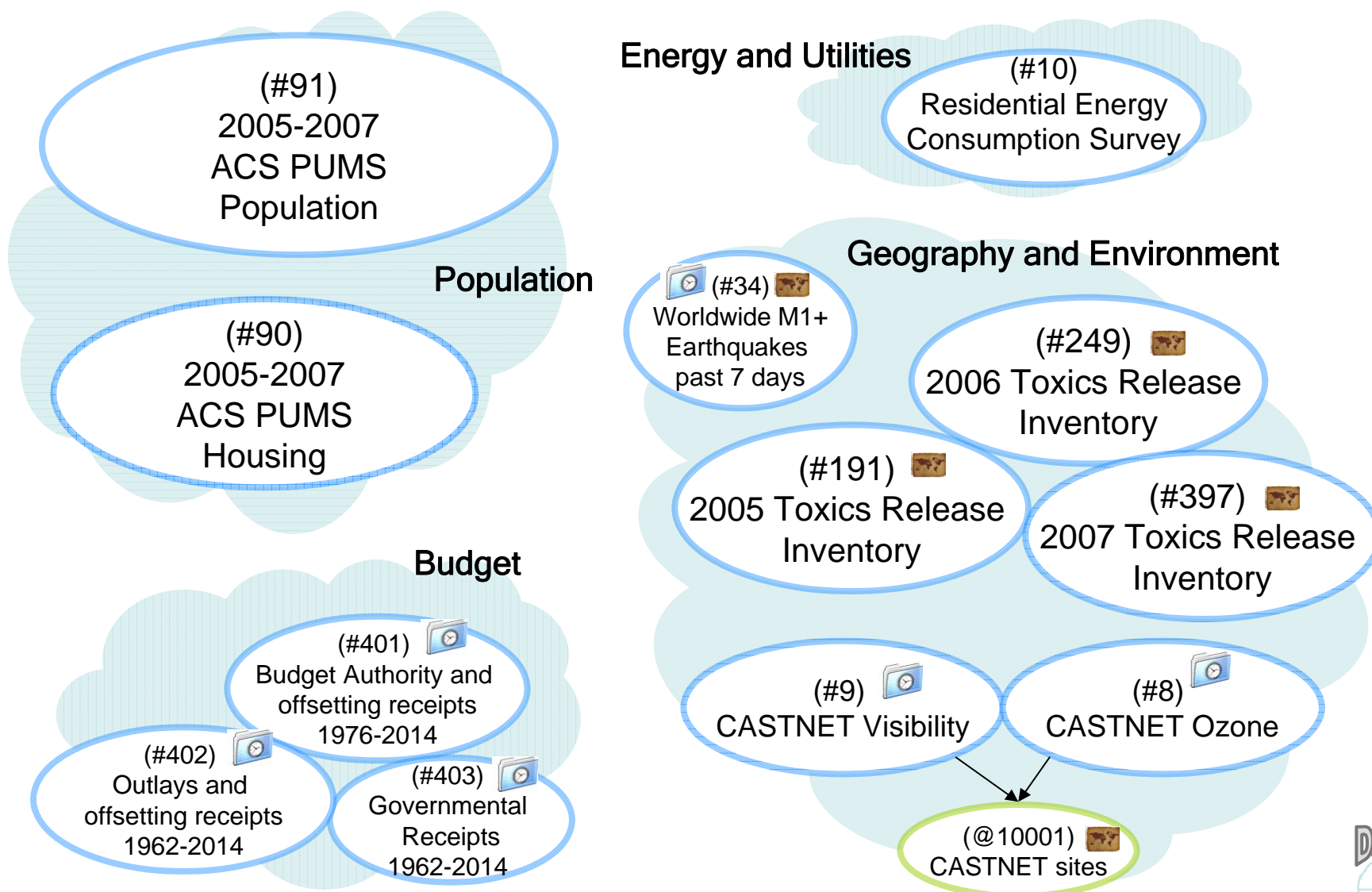
# The Landscape

# The catalog data

▼ <a href="http://data-gov.tw.rpi.edu/raw/92/data-92.rdf">http://data-gov.tw.rpi.edu/raw/92/data-92.rdf</a>		
Patent Application Bibliographic Data (2009)	92/agency	Department of Commerce
	92/agency data series page	<a href="http://www.uspto.gov/web/menu/patdata.html">http://www.uspto.gov/web/menu/patdata.html</a>
	92/agency program page	Patent Application Bibliographic Data <a href="http://www.uspto.gov/web/menu/patdata.html">http://www.uspto.gov/web/menu/patdata.html
	92/applicable agency information quality guideline designation	Department of Commerce/United States Patent and Trademark Office
	92/category	Business Enterprise
	92/citation	USPTO Patent Application Bibliographic Data <a href="http://www.uspto.gov/web/patents/howtopat.htm">http://www.uspto.gov/web/patents/howtopat.htm
	92/collection mode	person/paper and person/computer
	92/data collection instrument	<a href="http://www.uspto.gov/web/patents/howtopat.htm">http://www.uspto.gov/web/patents/howtopat.htm</a>
	92/data dictionary variable list	<a href="http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/2009/patdata.html">http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/2009/patdata.html</a>
	92/data gov data category type	Raw Data Catalog
	92/data quality certification	Yes
	92/date released	15-Mar-2001
	92/date updated	Thursdays
	92/description	Contains the bibliographic text (i.e., front page) of each patent application. This file is a subset of the Patent Application Bibliographic Data (PABD) International Common Element (ICE) Document Type Definition (DTD) file. The file is updated each week [where "yyyymmdd" is a Thursday publication date]. For more information about these files: <a href="http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/2009/patdata.html">http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/2009/patdata.html
	92/frequency	weekly
	92/geographic coverage	National and international
	92/identifier	USPTO Patent Application Bibliographic Data

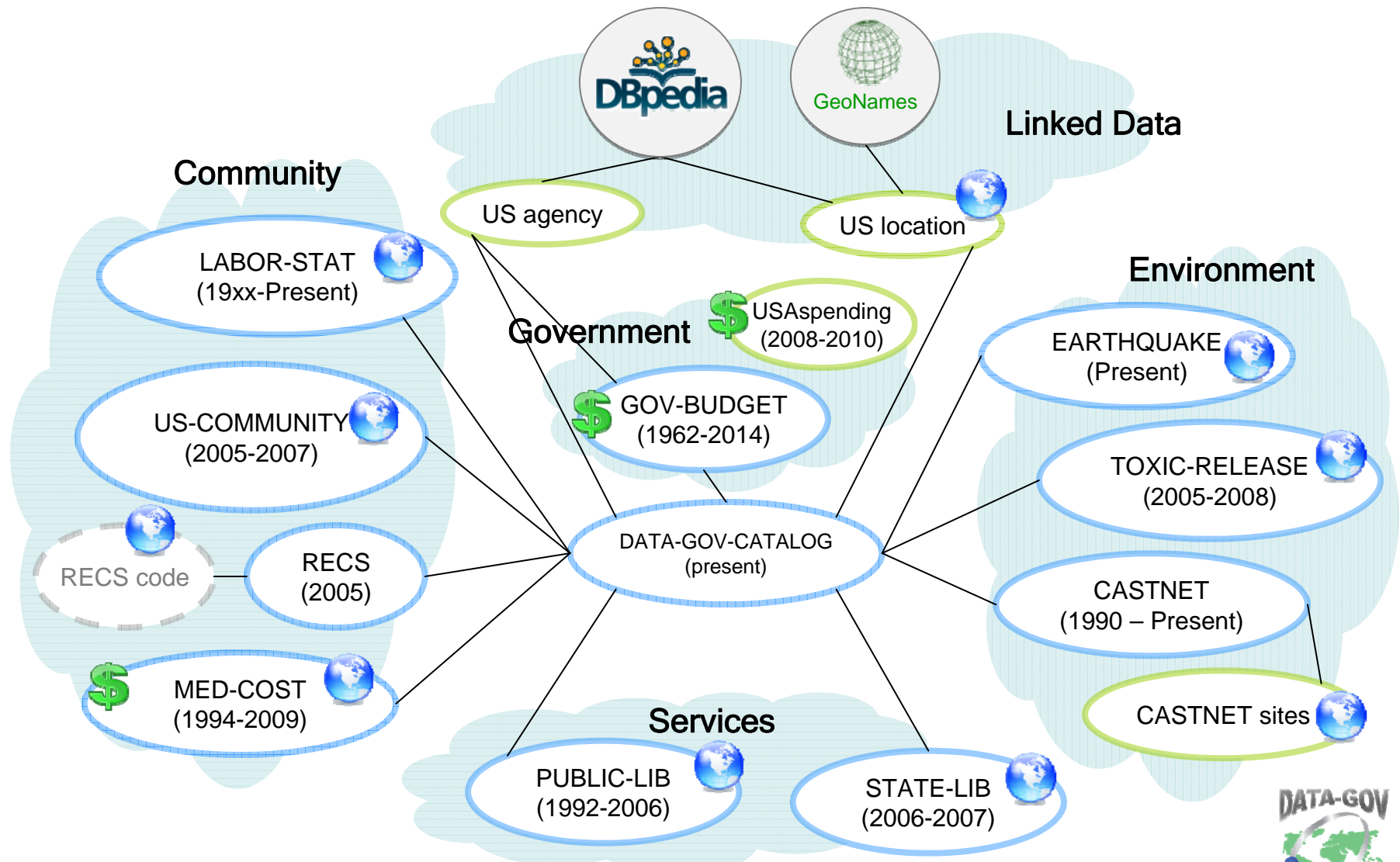


# Data-gov Cloud (Aug 2009)

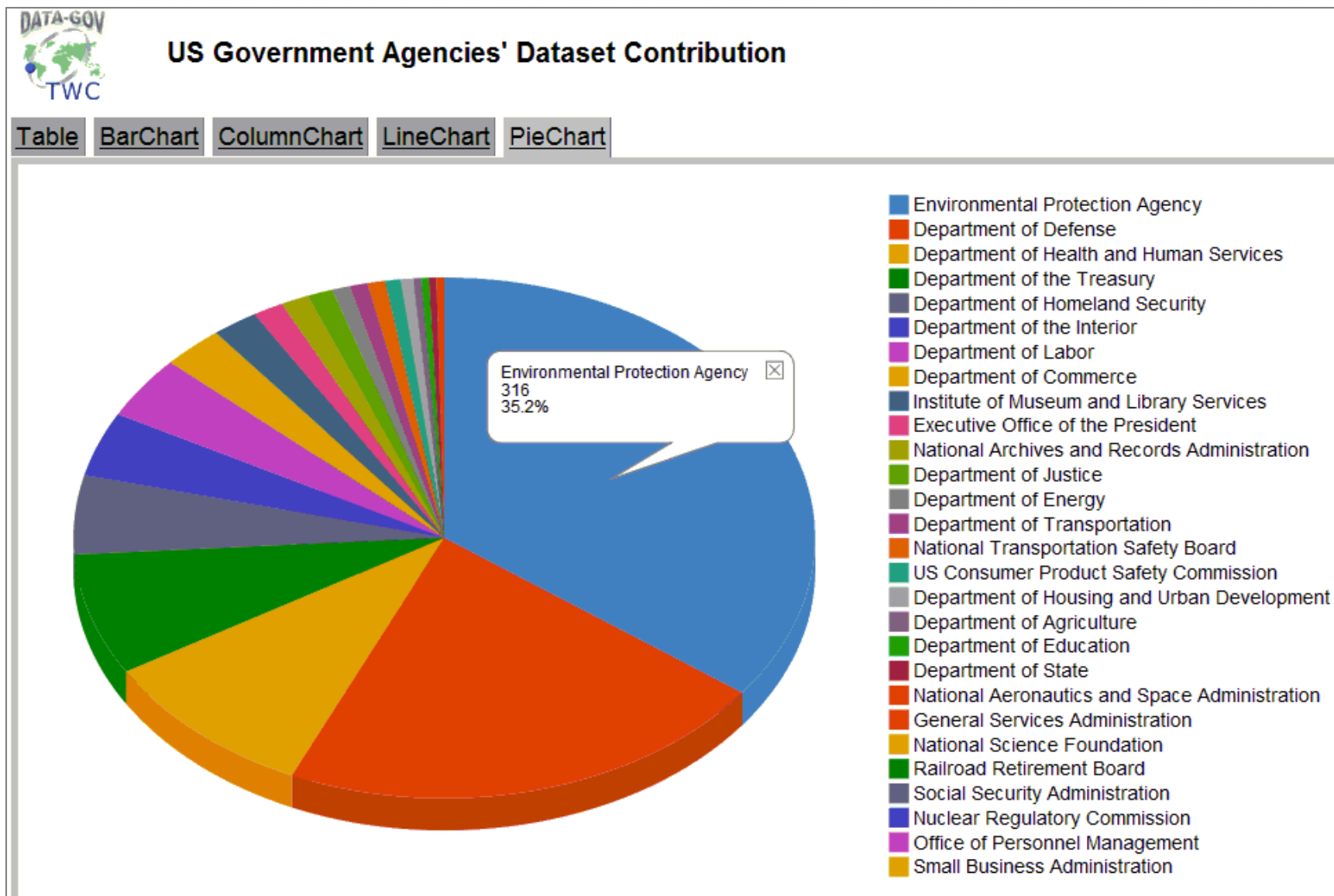




# Data-gov Cloud (Oct 2009)

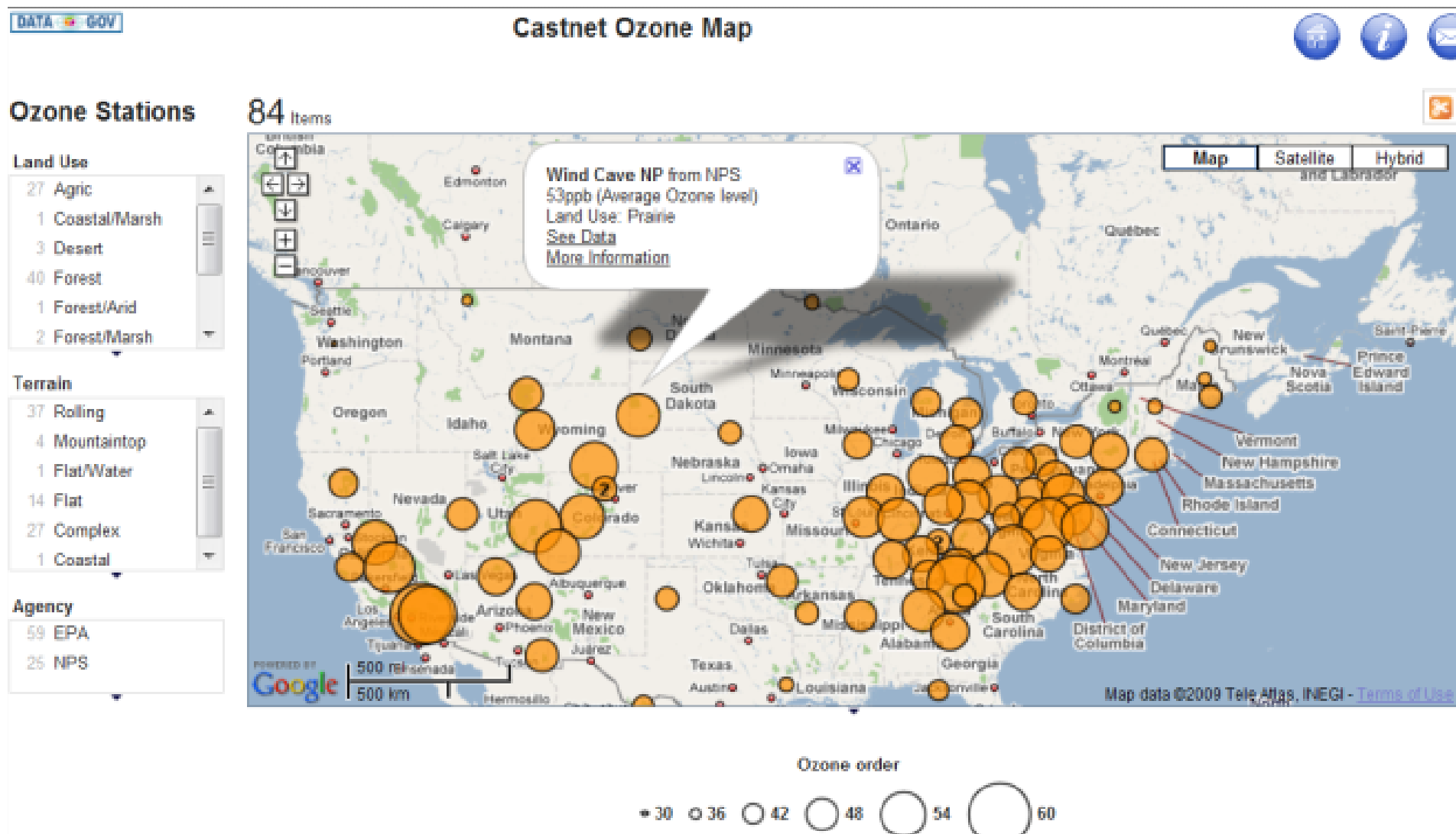


# More statistics



# Demos

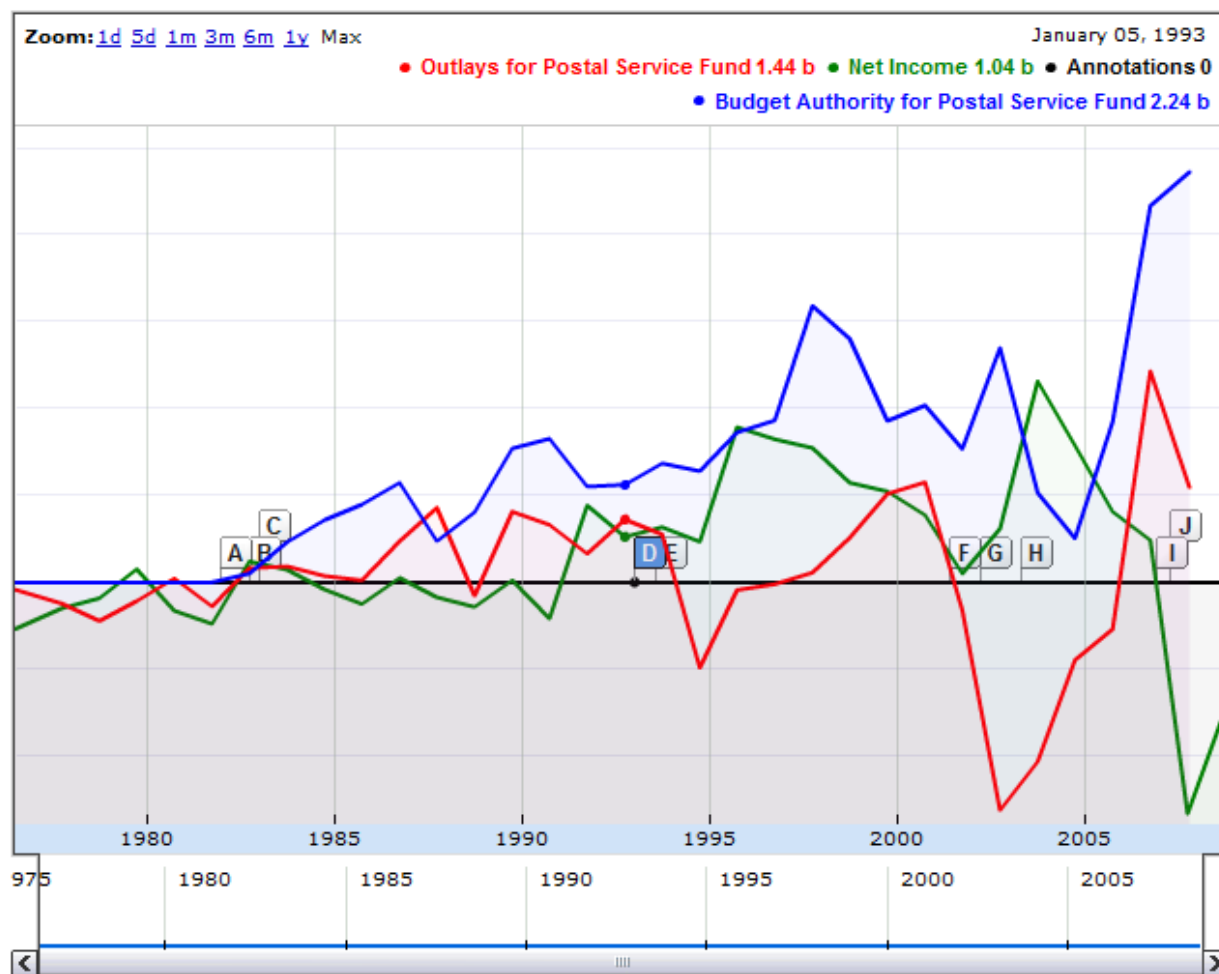
# Data.gov + epa.gov



# Gov Data + Corporate Data + User Data

DATA-GOV

## Annotated Timeline for USPS Money and Historical Events



billion rise in net income.  
2003-4-23

**G. 2002 Transformation Plan released** The USPS releases a report on major structural changes and reforms it can make to meet the challenges of the 21st century, as well as recommendations to Congress on how USPS can be helped.  
2002-4-1

**F. John E. Potter becomes postmaster** He is the 72nd Postmaster General of the US.  
2001-6-1

**E. National Postal Museum opens** New museum showcases US postal history. It is funded by the USPS, the Smithsonian Institute, and private donations.  
1993-7-30

**D. Government Performance and Results Act** This act requires government agencies to periodically release a 5-year plan on goals and how they plan to achieve them.  
1993-1-5

Net Income = Total income - total expenses

# Computing Difference of Revisions

The image shows two overlapping browser windows. The left window is a Mozilla Firefox browser displaying a Twitter search for the hashtag #datagov. The search results show several tweets from users like lidingpku and govwiki, mentioning updates to data.gov datasets. The right window is a web browser displaying the 'Changes from' page for the URL <http://data-gov.tw.rpi.edu/raw/92/2009-10-07/data-92.rdf> to <http://data-gov.tw.rpi.edu/raw/92/2009-10-10/data-92.rdf>. The page indicates a total of 6 changes. The changes are listed as follows:

- [new]Savings Bond Issues, Redemptions, and Maturities by Series**
- [add instance] <http://data-gov.tw.rpi.edu/raw/92/data-92.rdf#entry00908>
- > :xls\_access\_point . [values]{[http://www.treasurydirect.gov/govt/reports/pd/pd\\_sbredemptionsissuesbyseries.xls](http://www.treasurydirect.gov/govt/reports/pd/pd_sbredemptionsissuesbyseries.xls)}
- > :agency . [values]{Department of the Treasury}
- > :statistical\_characteristics . [values]{Not Relevant}
- > :data\_quality\_certification . [values]{Yes}
- > :applicable\_agency\_information\_quality\_guideline\_designation . [values]{Treasury}
- > :statistical\_methodology . [values]{Not Relevant}
- > :questionnaire\_design . [values]{Not Relevant}
- > :privacy\_and\_confidentiality . [values]{Not Relevant}
- > :title . [values]{Savings Bond Issues, Redemptions, and Maturities by Series}





# More demos?

- <http://data-gov.tw.rpi.edu/wiki/demos>

# Technical Issues

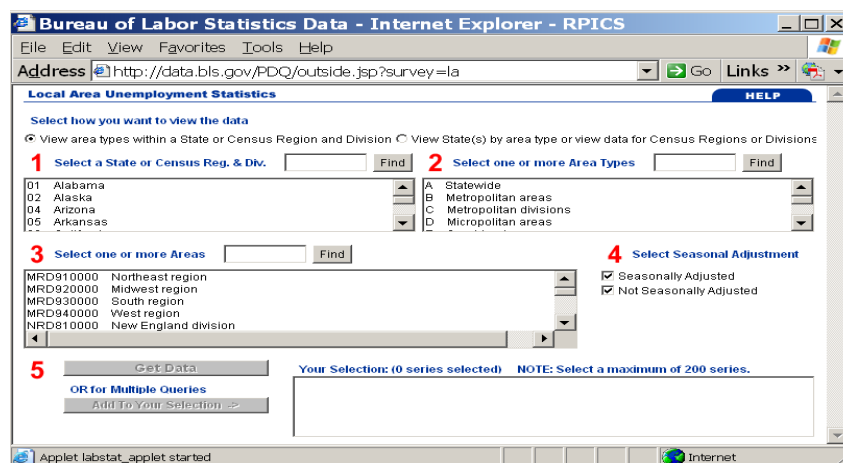
# Issues in Data.gov

- Duplicated Datasets- Some datasets are part of another dataset
  - Dataset 140 (2005 Toxics Release Inventory data for the state of California (EPA)) is a subset of Dataset 191.
- Formatting Issues - The format of some datasets is not friendly to machine processing.
  - Dataset 37 (Lower Colorado River Daily Average Water Elevations and Releases (US Bureau of Reclamation)).
  - Dataset 335 (National Longitudinal Surveys (US Bureau of Labor Statistics)) tells you how to order data from the government.
- Access Point Issues - The access points are interactive webpage which is not friendly for machine access.
  - Dataset 330 (Local Area Unemployment Statistics (US Bureau of Labor Statistics))

15	16900	3642.20	16274	13400	6033.71	3490	1890	6066.54	1460	970	7517.10	606.0
16	16900	3642.30	16276	13400	6033.71	3490	1890	6066.54	1460	970	7516.96	607.5

GLEN CANYON TO HOOVER DAM LOSSES												
JULY 2009												
DATE	GLEN RELEASE CFS	HOOVER RELEASE CFS	LAKE HEAD STORAGE 1000 AF	STORAGE CHANGE 1000 AF	LAKE HEAD PUMPING CFS	LAKE HEAD EVAPORATION CFS	CHANGE BANK STORAGE CFS	LOSS CFS	ACCUMUL. A.F.	INFLOW CFS		
1	12900	12300	11066	-5	677	1134	-164	554	1099	11426		
2	12900	11700	11065	-1	686	1134	-33	-1334	-1547	12983		
3	13300	11100	11066	1	673	1134	33	-499	-2536	13444		
4	12100	9650	11070	4	683	1134	131	-713	-3950	13615		
5	12100	10800	11074	4	619	1134	131	-1432	-6791	14701		
6	13200	13300	11070	-4	697	1134	-131	-847	-8470	12983		
7	13200	11900	11067	-3	713	1134	-98	-59	-8587	12136		
8	13200	10700	11073	6	723	1134	197	-2620	-13784	15779		
9	13200	14100	11066	-7	759	1134	-229	955	-11889	12234		
10	13400	15700	11059	-7	726	1134	-229	-632	-13144	13801		
11	12800	14700	11057	-2	743	1134	-66	-2353	-17811	15503		
12	11700	15600	11049	-8	715	1134	-262	228	-17357	13154		
13	12800	17400	11039	-10	726	1134	-328	-1115	-19568	13891		
14	13200	18800	11022	-17	763	1134	-557	135	-19302	11569		
15	13400	17000	11015	-7	757	1134	-229	-2357	-23976	15132		
16	13400	14800	11011	-4	765	1134	-131	-1393	-26739	14551		
TOTAL	206500	219500			11425	18144			-26740	216903		



The screenshot shows the 'Bureau of Labor Statistics Data - Internet Explorer - RPICS' interface. The address bar shows 'http://data.bls.gov/PDQ/outside.jsp?survey=la'. The page title is 'Local Area Unemployment Statistics'. The main content area has a section 'Select how you want to view the data' with two radio buttons: 'View area types within a State or Census Region and Division' (selected) and 'View State(s) by area type or view data for Census Regions or Divisions'. Below this are four numbered steps: 1. Select a State or Census Reg. & Div. (with a dropdown menu showing Alabama, Alaska, Arizona, Arkansas); 2. Select one or more Area Types (with a dropdown menu showing A. Statewide, B. Metropolitan areas, C. Metropolitan divisions, D. Micropolitan areas); 3. Select one or more Areas (with a dropdown menu showing MRD910000 Northeast region, MRD920000 Midwest region, MRD930000 South region, MRD940000 West region, MRD910000 New England division); 4. Select Seasonal Adjustment (with checkboxes for 'Seasonally Adjusted' and 'Not Seasonally Adjusted'). At the bottom, there is a 'Get Data' button and a 'Your Selection: (0 series selected) NOTE: Select a maximum of 200 series.' message.

# Linking Data

1. link similar datasets by reusing property namespace
2. link to `rdfs:label` (via `rdfs:subPropertyOf`) using semantic wiki
3. link to DBpedia (via `owl:sameAs`) using *wikipedia widget*
4. link instances (via common `<property, literal-value>` pair)
5. link government data with web data (via time and location)
6. link revisions of government data (via knowledge provenance)

# Semantic mapping: AI + CI

Georgia

**Infobox (State of the U.S.)**

- modified: 14 October 2009 18:29:53
- abbreviation: GA
- geoname: <http://sws.geonames.org/4197000/>

Your *continued donations* keep Wikipedia running

[article](#) [discussion](#)

## Georgia

From Wikipedia, the free encyclopedia

**Georgia** has two principal meanings:

- [Georgia \(country\)](#), previously known as:
  - [Georgian Soviet Socialist Republic](#)
  - [Democratic Republic of Georgia](#)
  - [Georgian Kingdom](#), various earlier names
  - [Georgia \(U.S. state\)](#), previously known as:
    - [Province of Georgia](#) (1732–1776)

WIKIPEDIA  
The Free Encyclopedia

navigation

- [Main page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)

need manual  
disambiguation!

Georgia

Map to Wikipedia/DBpedia Name

**Infobox (State of the U.S.)**

- modified: 14 October 2009 18:48:09
- abbreviation: GA
- geoname: <http://sws.geonames.org/4197000/>

Your *continued donations* keep Wikipedia running

[article](#) [discussion](#) [edit this page](#)

## Georgia (U.S. state)

From Wikipedia, the free encyclopedia

**Georgia** (/ˈdʒɔrdʒə/ (help·info)) is a state in the Southeastern United States. It was the fourth state to ratify the United States Constitution on September 15, 1788. With an estimated 9,685,000 residents in 2007 to 2008, 14 of Georgia's counties were known as the *Peach State* and the *Empire State of the South*. Georgia is bordered on the south by Florida.

WIKIPEDIA  
The Free Encyclopedia

navigation

- [Main page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)

# RDF => SPARQL => Web

- We use SPARQL to bridge Web developers and Semantic Web data.
- A triple store is used to support handling multi-million triple RDF datasets



# Conclusion

- ☐ semantic web enabled portal for linked government data
- ☐ 5 billion triples from data.gov
- ☐ hosts apps, demos & services
- ☐ provide education services
- ☐ integrates web users' contributions