Peta vs. Meta: Rethinking Data Interoperability on the World Wide Web

Jim Hendler

Director, Rensselaer Institute for Data Exploration and Applications

Rensselaer Polytechnic Institute, USA

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

@jahendler (twitter)

Roots: Data Exploration

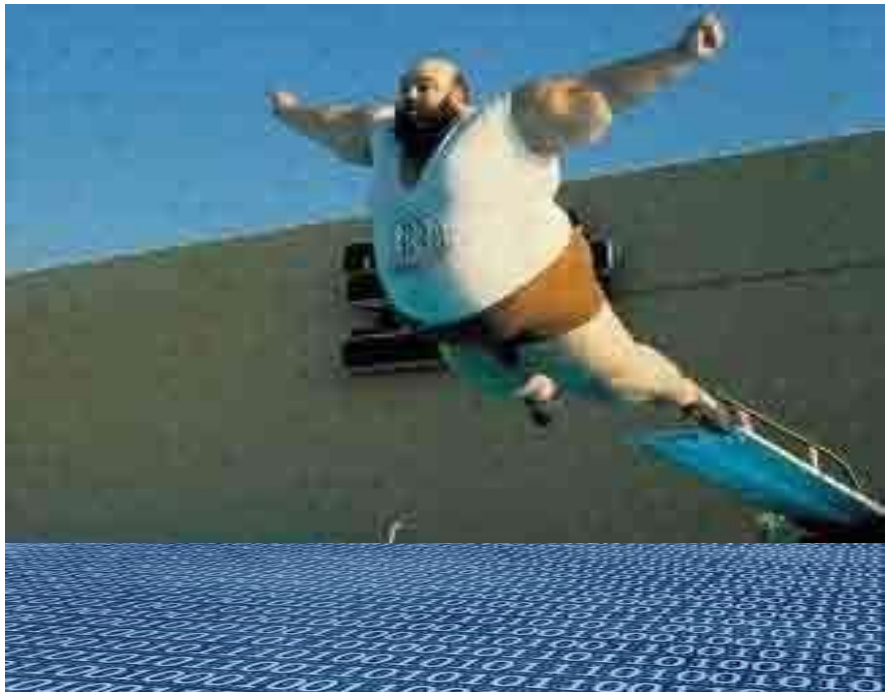
Geekopedia: Data exploration helps a data consumer focus an information search on the pertinent aspect of relevant data before true analysis can be achieved. In large data sets, data is not gathered or controlled in a focused manner. Even in smaller data sets, it is also true that data gathered are not in a very rigid and specific technique can result in a disorganized manner and a myriad of subsets each…

Discover

Integrate

Validate

Explain



Government Data Sharing

Tetherless World Constellation

International OGD Catalog Search, searching **440,386** datasets from **116 catalogs** in **16 languages** representing **38 countries** and international organizations

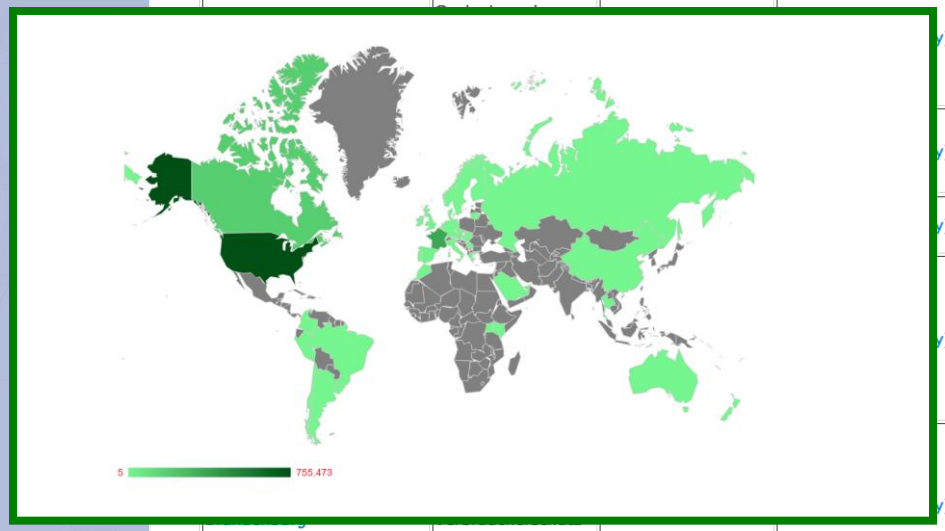
Results 1 to 200 of 440386 next

Dataset Title	Agency	Categories	Catalog	Country
Elektronischer Wasserstraßen-Informationsservice - [Karte Nordwest]	Nieders chsischer Landesbetrieb f r Wasserwirtschaft K sten- und	Wasser	Portalu.de (Germany)	Germany

Catalogs: Deselect All

(395459) Data.gov --- Geodata Catalog
(7702) Data.gov.uk (United Kingdom)
(5978) Data.gov.sg (Singapore)
(3958) Opendatacordoba.com(Cordoba,Spain)
(3707) Data.gov (United States) --- Raw Data
(2489) Data.gov.bc.ca (British Columbia, Can

International OGD Catalog Search, searching **1,022,787** datasets from **192 catalogs** in **24 languages** representing **43 countries** and international organizations.



	Naturschutz			
	Landesamt f r			Germany
				Germany
				Germany
				Germany
	des Landes Brandenburg			Germany
		chemikalien, gesundheit		

(406280) United States
(8940) United Kingdom
(5978) Singapore
(5907) Spain
(4777) Canada
(1687) New Zealand
(1293) Italy
(776) Australia

Agencies: Deselect All

(1666) Environmental Protection Agency
(1468) Health
(1347) Economic Development Board
(1123) Department of Health
(1117) U.S. Geological Survey Earth Resource
(883) Land Information New Zealand
(781) Department for Communities and Local

Categories: Deselect All

(140) Datos Demográficos
(139) Empleo
(139) Infraestructuras
(136) Labour
(134) Finance
(129) Sciences
(126) Transportation and Storage
(126) State sector performance

Semantic Web and Linked Data (UK)

	A	B	C
1	Potholes in our county council		
2			
3	**Date reported**	**Street**	
4	15/01/11	High Street	
5	16/01/11	Tree Lane	
6	16/01/11	High Street	
7	18/01/11	B3068	
8	19/01/11	B3068	
9	22/01/11	B2013	
10	23/01/11	High Street	
11	24/01/11	High Street	
12	24/01/11	London Road	
13	26/01/11	Birmingham Road	
14	27/01/11	High Street	
15	28/01/11	High Street	
16			

County Council

Street	**Post code**
High Street	BM1 3
Tree Lane	BM1 2
B3068	BM1 1
London Road	BM1 2
Birmingham Road	BM 1 3

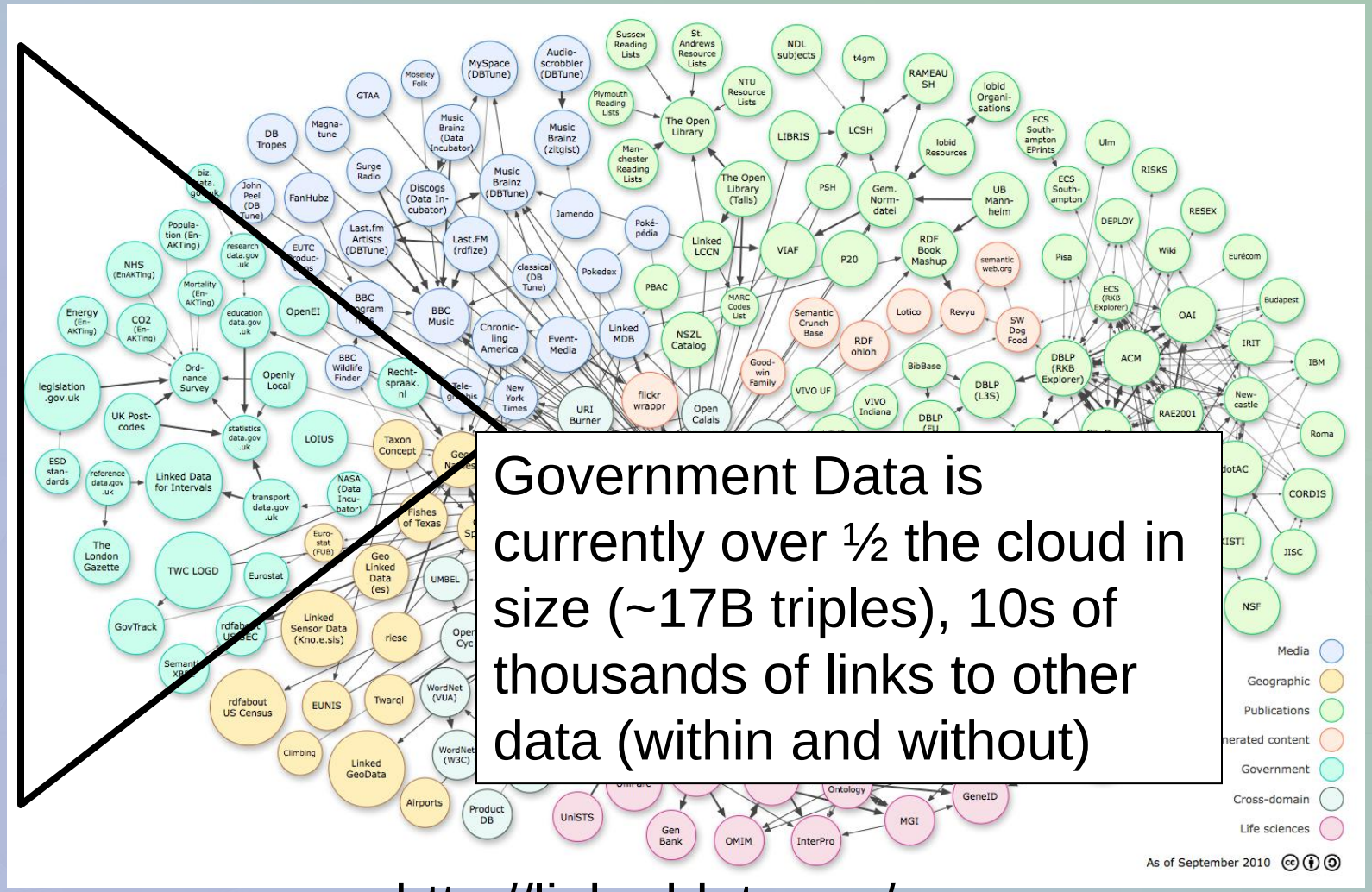
Royal Mail

	A	B	C	D	E	F	
1	BM1 1AE	10	291960	92581	E92000001	E19000002	E18
2	BM1 1AT	10	291778	92355	E92000001	E19000002	E18
3	BM1 1BA	10	291725	92265	E92000001	E19000002	E18
4	BM1 1BB	10	291786	92251	E92000001	E19000002	E18
5	BM1 1BD	10	291763	92290	E92000001	E19000002	E18

Ordnance Survey

Government Data in the linked open data cloud

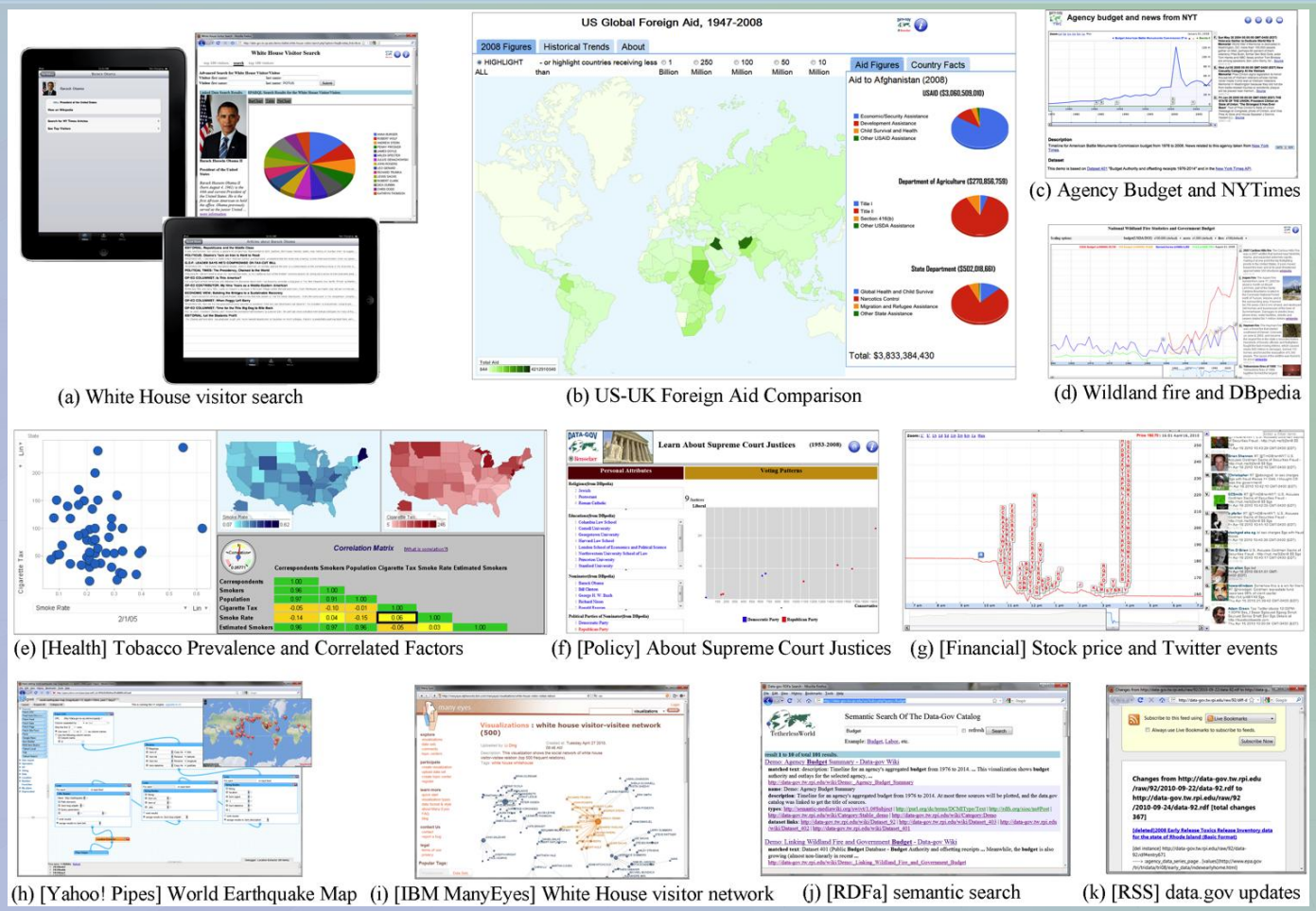
Tetherless World Constellation



http://linkeddata.org/

Creating Data Mashups Requires Semantics

Tetherless World Constellation

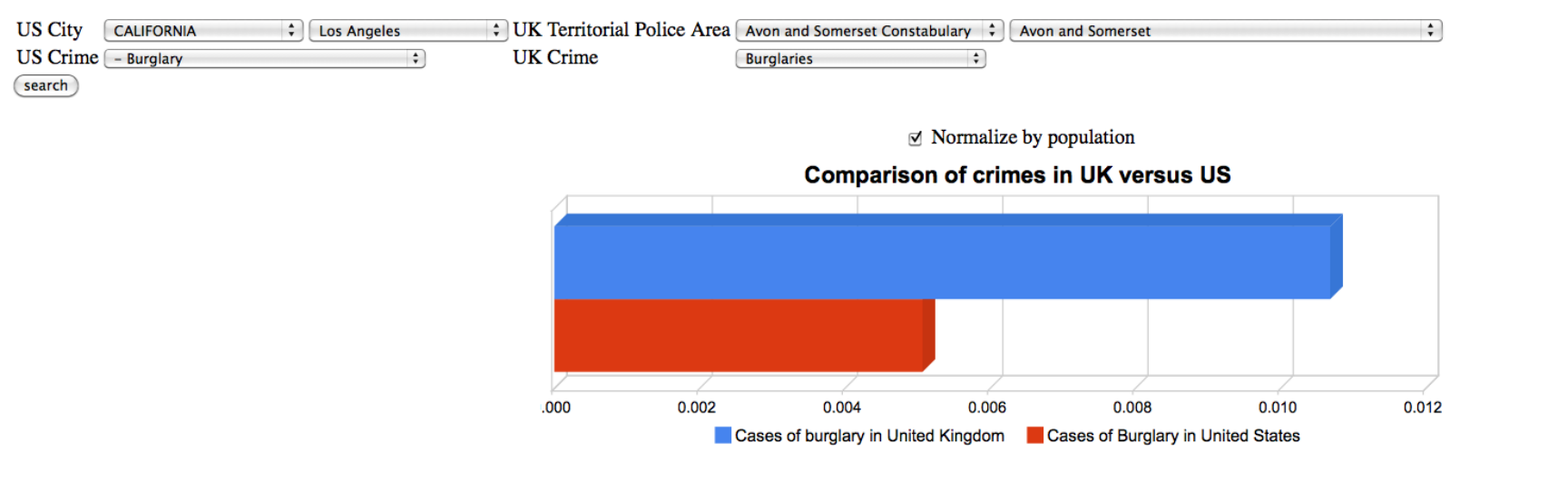


(a) White House visitor search
(b) US-UK Foreign Aid Comparison
(c) Agency Budget and NYTimes
(d) Wildland fire and DBpedia
(e) [Health] Tobacco Prevalence and Correlated Factors
(f) [Policy] About Supreme Court Justices
(g) [Financial] Stock price and Twitter events
(h) [Yahoo! Pipes] World Earthquake Map
(i) [IBM ManyEyes] White House visitor network
(j) [RDFa] semantic search
(k) [RSS] data.gov updates

More than 50 of these at http://logd.tw.rpi.edu

BUT: it’s not that easy

Tetherless World Constellation



Head to head comparions shows that burglaries in Avon and Somerset (UK) far exceed those in Los Angeles, California (one of the highest crime areas in the US)

The problem is (likely) semantics

Same or different?

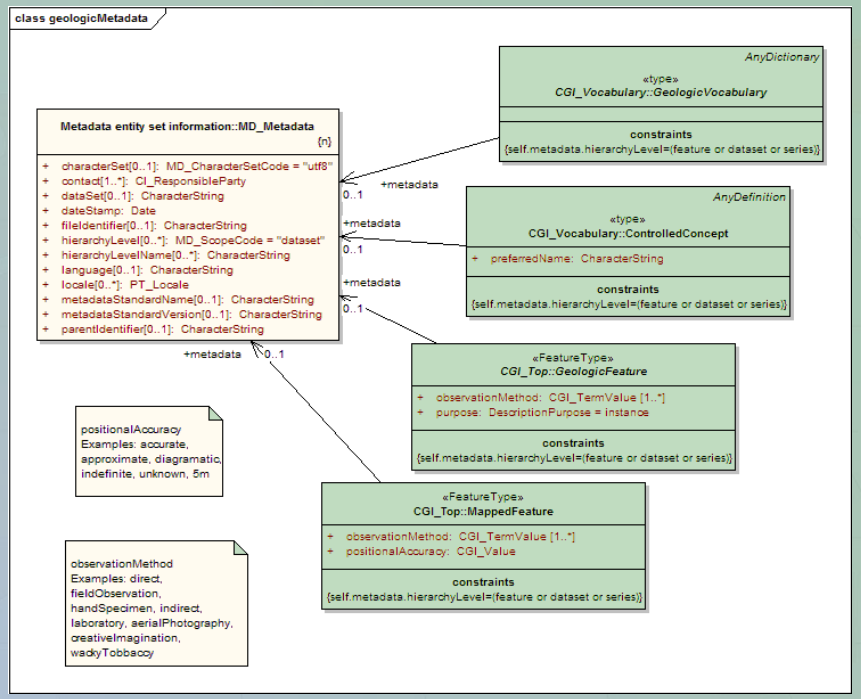
Do the terms mean the same? Are they collected in the same way? Are they processed differently? …

Metadata

Tetherless World Constellation

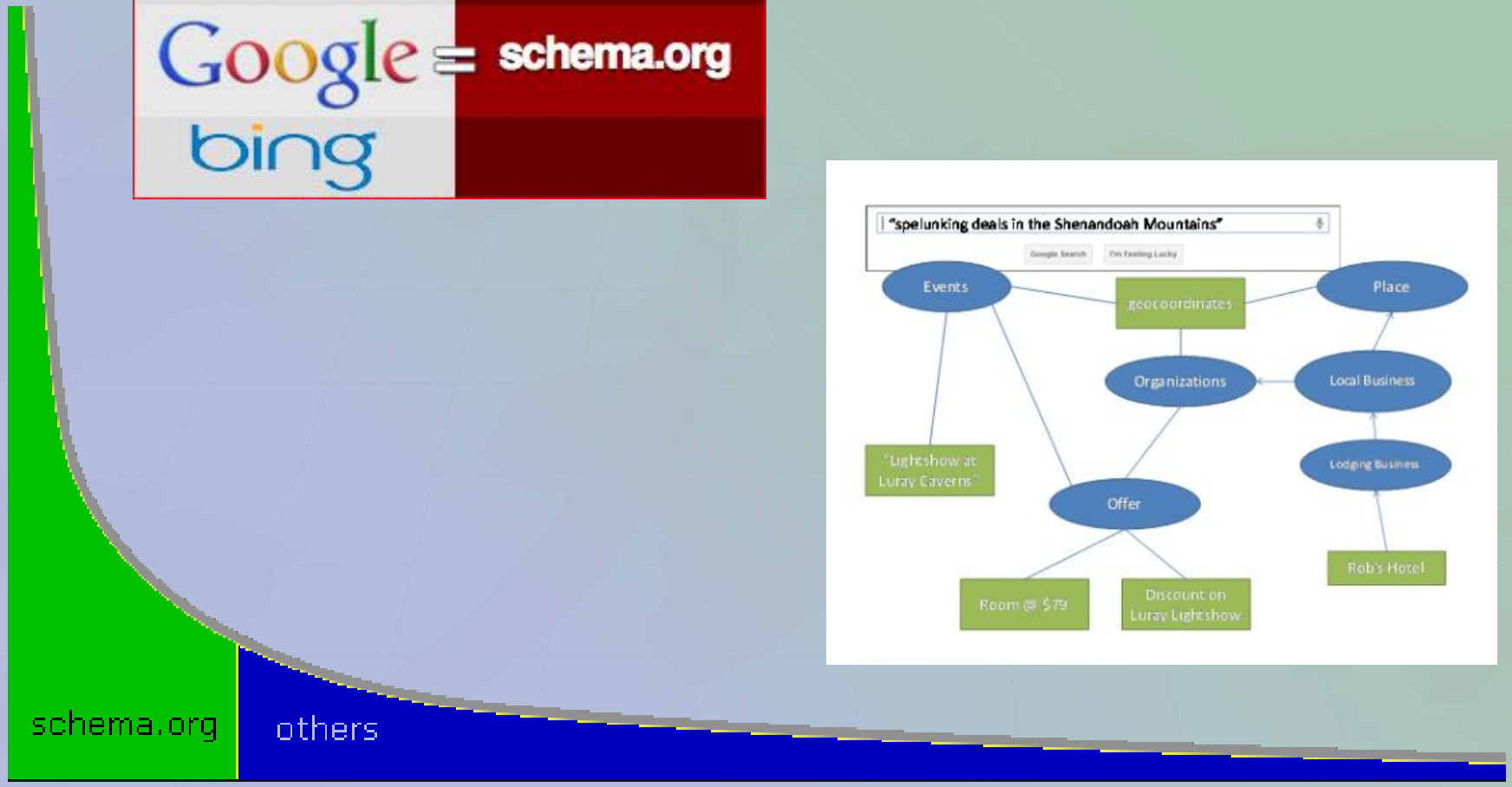
- International data sharing
  - W3C Govt Linked Data Working Group
  - Need for vocabularies within govt sectors
    - Esp for cross-langauge use
      - How can we compare health (or legal, or social, or ….) data between countries like US, UK, India, Kenya (English) with Norway, China, France, etc.
      - How can we link local govts (in traditional languages, local dialects, etc) w/national data
- Modern Metadata design is crucial to govt data sharing
  - Needed for search and federation in large data sharing efforts

- Traditionally metadata tries to be comprehensive
  - Example:ISO 19115 (GIS standard)
    - >400 elements
    - 14 “packages”
    - Dozens of UML models (not all consistent w/ each other)



- Big Data on the web
  - is moving away from traditional relational models (*cf*. NoSQL)
  - Moving towards third party application and extension (cf. Json)
  - Focus on interoperability and exchange with “lightweight” semantics
    - Using ideas from the Semantic Web
      - Search: Schema.org
      - Social Networking: OGP

YAHOO!
Google = schema.org
bing



Dataset extension to schema.org - April, 2013

Tetherless World Constellation

Datasets extension	DCAT	ADMS	VoID
ds:DataCatalog	dcat:Catalog	adms:SemanticAssetRepository	
ds:DataDownload	dcat:Download	adms:SemanticAssetDistribution	
ds:Dataset	dcat:Dataset	adms:SemanticAsset	void:Dataset
ds:catalog		dcterms:isPartOf	
ds:dataset	dcat:dataset	dcterms:hasPart	
ds:distribution	dcat:distribution	radion:distribution	void:dataDump
ds:keyword	dcat:keyword	radion:keyword	
ds:license	dcterms:license	dcterms:license	
ds:spatial	dcterms:spatial	dcterms:spatial	
sdo:about	dcat:theme	dcterms:subject	
sdo:contentSize	dcat:size		
sdo:contentURL	dcat:accessURL	adms:accessURL	
sdo:copyrightHolder			
sdo:Country			
sdo:dateModified	dcterms:modified	dcterms:modified	
sdo:datePublished	dcterms:issued	dcterms:created	
sdo:description	dcterms:description	dcterms:description	
sdo:encodingFormat	dcterms:format	dcterms:format	
sdo:inLanguage	dcterms:language	dcterms:language	
sdo:name	dcterms:title	rdfs:label	
sdo:Organization	foaf:Organization		
sdo:Person	foaf:Person		
sdo:publisher	dcterms:publisher	dcterms:publisher	
sdo:Thing	skos:Concept	(recommends but does not require skos:Concept)	
sdo:url	foaf:homepage		
sdo:version		radion:version	
	dcat:CatalogRecord		
	dcat:dataDictionary		

Schema.org/Dataset – add this to your pages!

SF Shoreline

San Francisco mainland shoreline and in the south, the county line.

Country: United States

Publisher: Department of Public Works

Human-readable database description (HTML)

type: http://schema.org/dataset
property:

name: SF Shoreline
url: http://www.datasf.org/story.php?title=sf-shoreline-
description: San Francisco mainland shoreline and in the south, the county line.
spatial: Item 9
publisher: Item 10

Item 9
type: http://schema.org/country
property:

name: United States

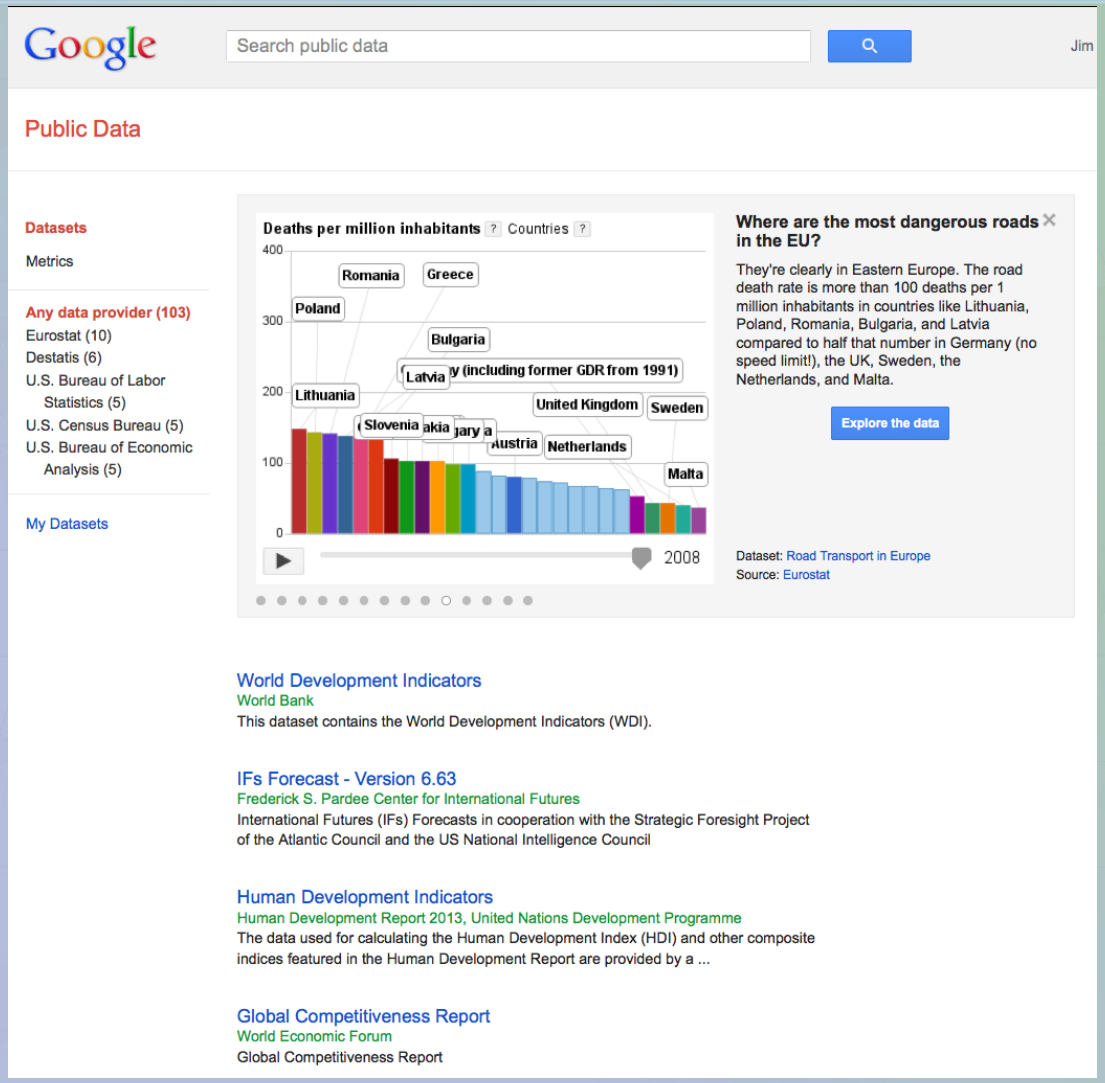
Item 10
type: http://schema.org/organization
property:

Department of Public Works

Embedded meta-data (RDFa)

Schema.org/Dataset
google.com/publicdata (early days)

Tetherless World Constellation



US moving in right direction

Tetherless World Constellation

The White House

Office of the Press Secretary

E-Mail Tweet Share +

For Immediate Release May 09, 2013

Executive Order -- Making Open and Machine Readable the New Default for Government Information

EXECUTIVE ORDER

- - - - - - -

MAKING OPEN AND MACHINE READABLE THE NEW DEFAULT
FOR GOVERNMENT INFORMATION

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

USA “Project Data” – metadata JSON

Tetherless World Constellation


```
[
    {
        "title": "Data Catalog",
        "description": "Version 1.0",
        "keyword": "catalog",
        "modified": "2013-05-09 06:00:00",
        "publisher": "US Department of X",
        "person": "Contact Person",
        "mbox": "contact.person@agency.gov",
        "identifier": "1",
        "accessLevel": "public",
        "distribution": [
            {
                "accessURL": "http://agency.gov/data.json",
                "format": "json"
            }
        ]
    },
    {
        "title": "Public Elementary/Secondary Listing",
        "description": "The purpose of the CCD nonfiscal surveys is to provide a listing of all schools and agencies providing
        "keyword": "education, schools",
        "modified": "2011-11-19 00:00:00",
        "publisher": "US Department of Education",
        "person": "Open Data Initiative",
        "mbox": "opendata@ed.gov",
        "identifier": "ykv5-fn9t",
        "accessLevel": "public",
        "dataDictionary": "http://nces.ed.gov/ccd/pdf/INsc09101a.pdf",
        "distribution": [
            {
                "accessURL": "https://explore.data.gov/views/ykv5-fn9t/rows.csv?accessType=DOWNLOAD",
                "format": "csv",
                "size": "200mb"
            },
            {
                "accessURL": "https://explore.data.gov/views/ykv5-fn9t/rows.json?accessType=DOWNLOAD",
                "format": "json"
            },
            {
                "accessURL": "https://explore.data.gov/views/ykv5-fn9t/rows.xml?accessType=DOWNLOAD",
                "format": "xml"
            }
        ],
        "webService": "http://explore.data.gov/api/views/ykv5-fn9t/rows.json",
        "license": "Public Domain",
        "spatial": "US",
        "temporal": "2009-09-01 00:00:00,2010-05-31 00:00:00",
        "issued": "",
        "frequency": "one-time",
        "language": "English",
        "granularity": "",
        "dataQuality": "true",
        "theme": "education",
        "references": "http://nces.ed.gov/ccd/data/txt/psu091alay.txt",
        "landingPage": "http://ed.gov/developer",
        "feed": "",
        "systemOfRecords": "http://nces.ed.gov/ccd/"
    }]
```


Aimed at developers

Based on DCAT

USA "Project Data" – metadata RDFa

Tetherless World Constellation

Office Locations

- **Title:** Office Locations
- **Description:** A list of the agency's office locations and contact information.
- **Documentation URL:** http://www.agency.gov/data/information/locations
- **Download URL:** http://www.agency.gov/data/raw/locations.zip
- **Format:** csv
- **Tags:** keyword1
- **Last Update:** 1/1/2013
- **Publisher:** Agency
- **Contact Name:** John, Smith
- **Contact Email:** john.smith@agency.gov
- **Unique Identifier:** 1
- **Public:** true
- **Endpoint:** http://www.agency.gov/data/raw/locations.json
- **License:** public domain
- **Spatial:** United States
- **Temporal:** today
- **Release Date:** 7/9/2012
- **Frequency:** 6 months
- **Language:** English
- **Granularity:** Address
- **Category:** Energy
- **Related Documents:** http://www.agency.gov/data/information/locations/document.doc
- **Distribution:**
- **Size:** 44KB
- **Homepage URL:**
- **RSS Feed:**
- **Data Quality:** True

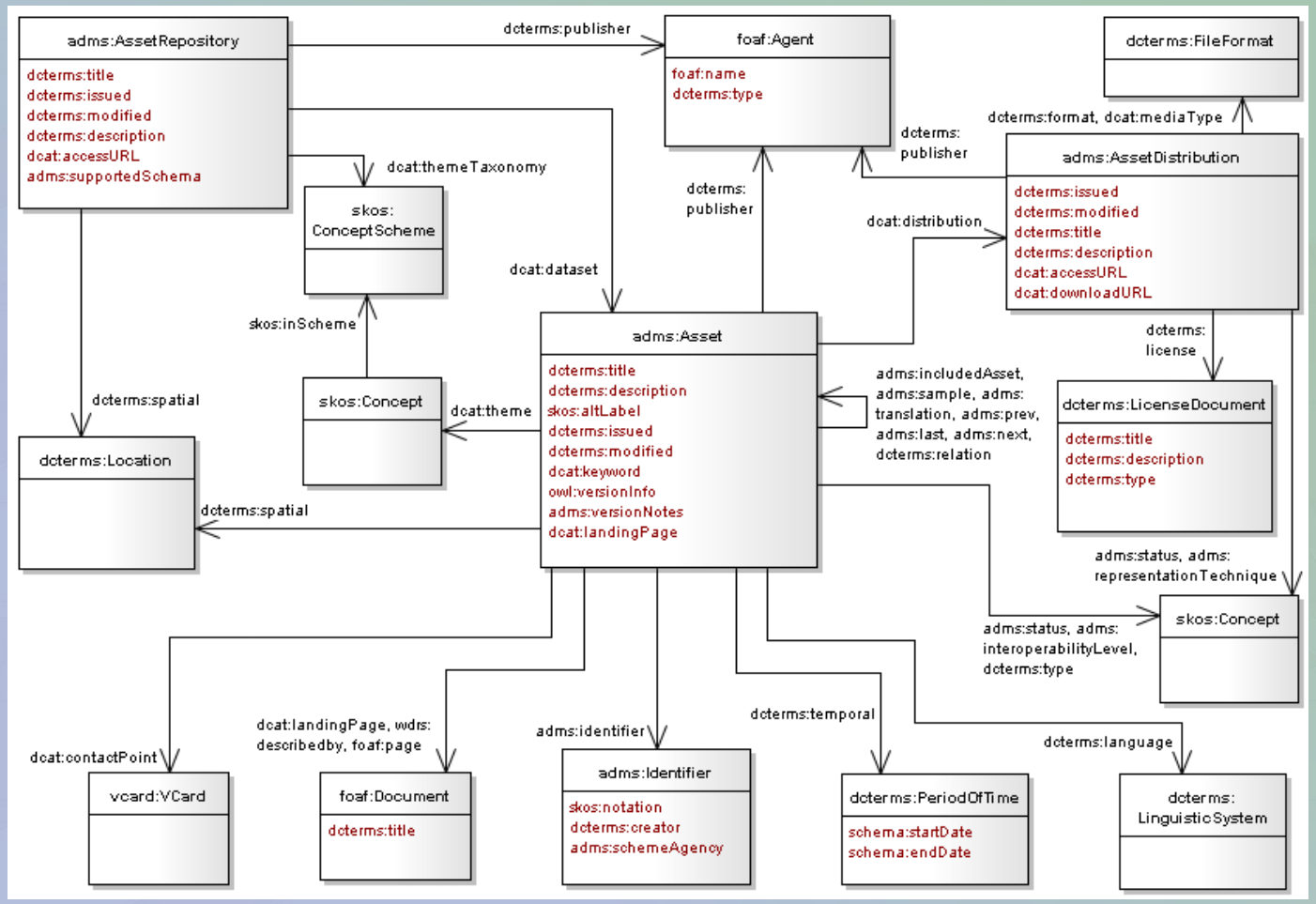
Embedded metadata for Search, Web Apps

Based on Schema.org/Dataset

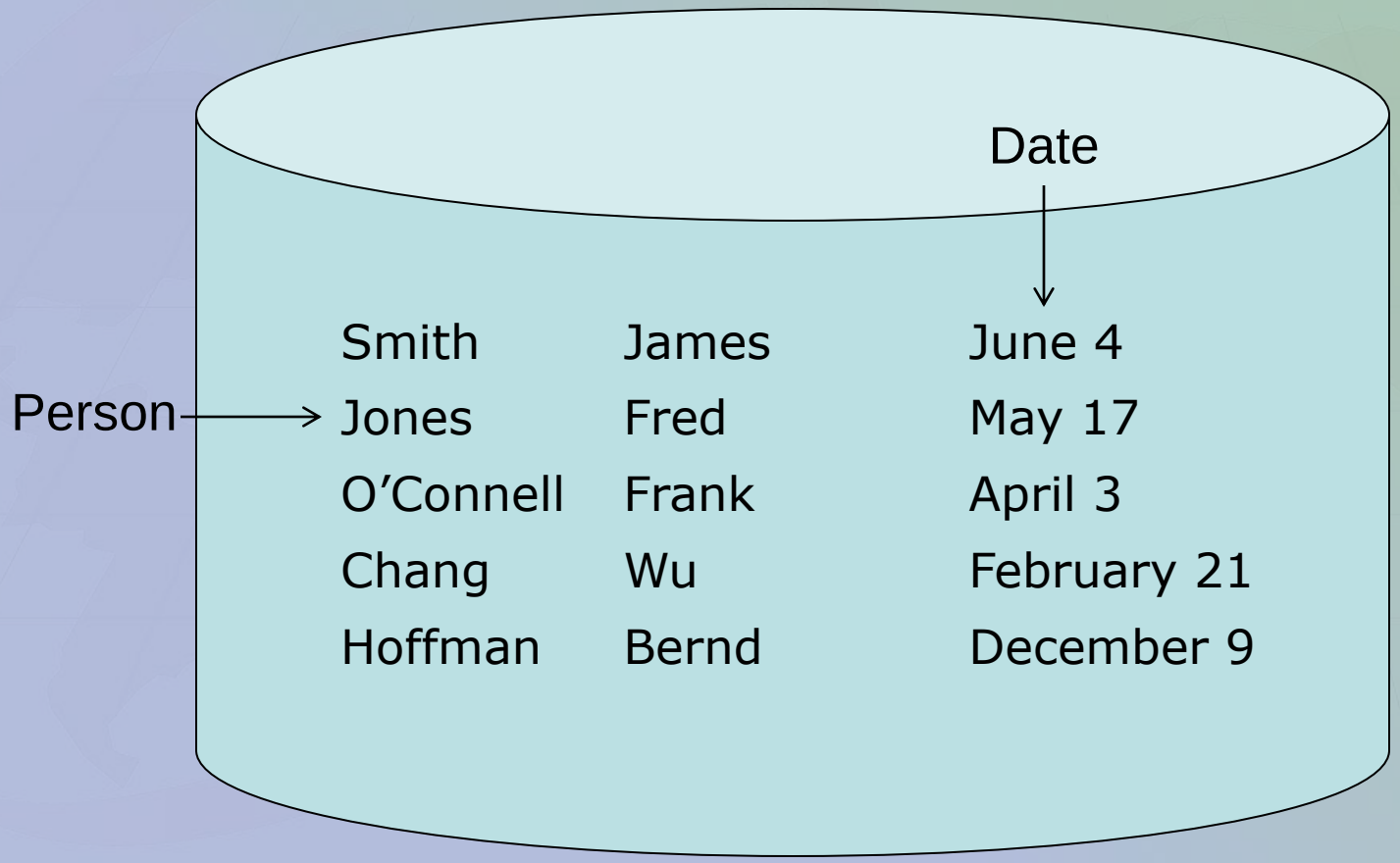
EU moving in right direction

Tetherless World Constellation

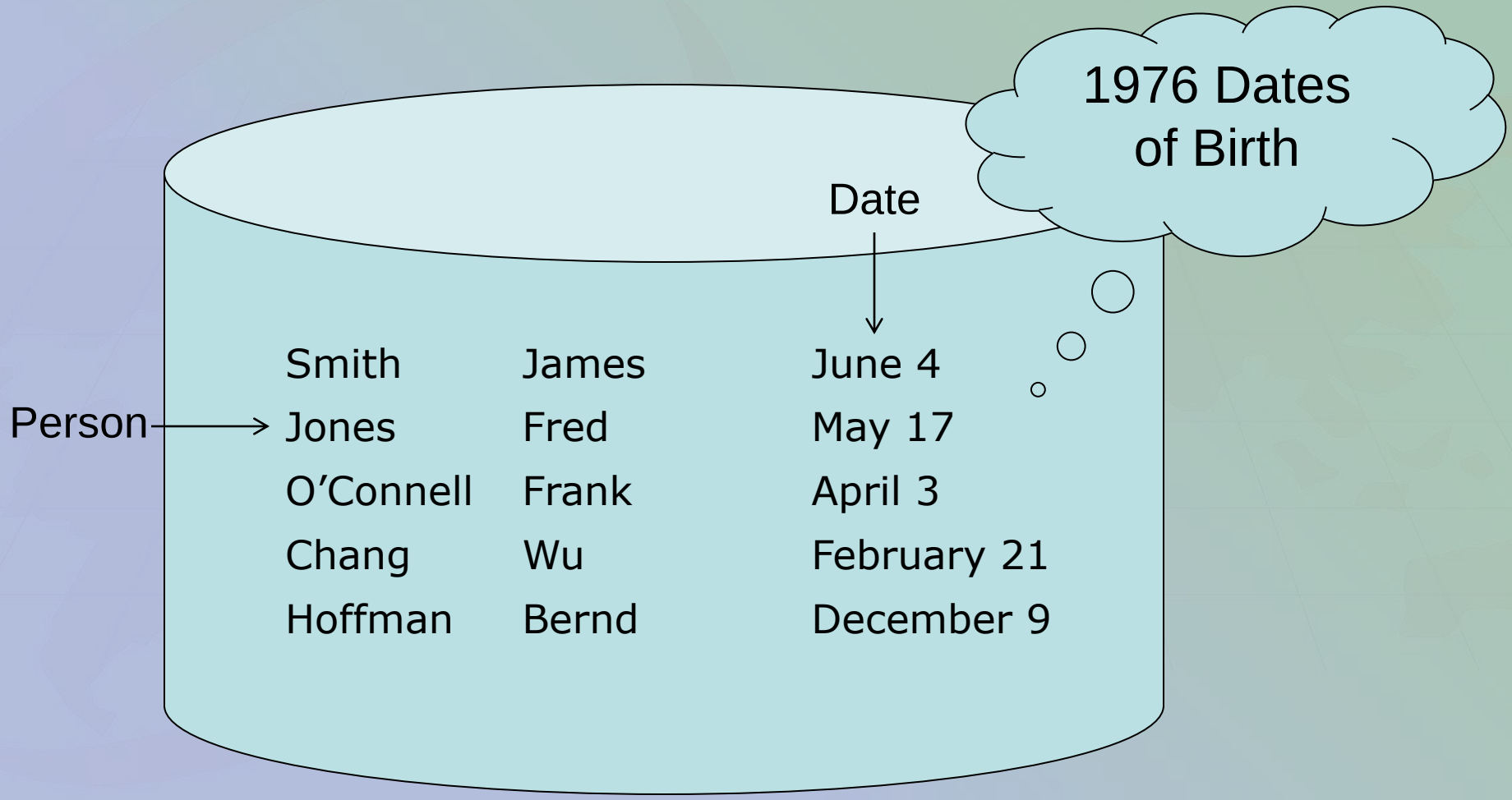
ADMS



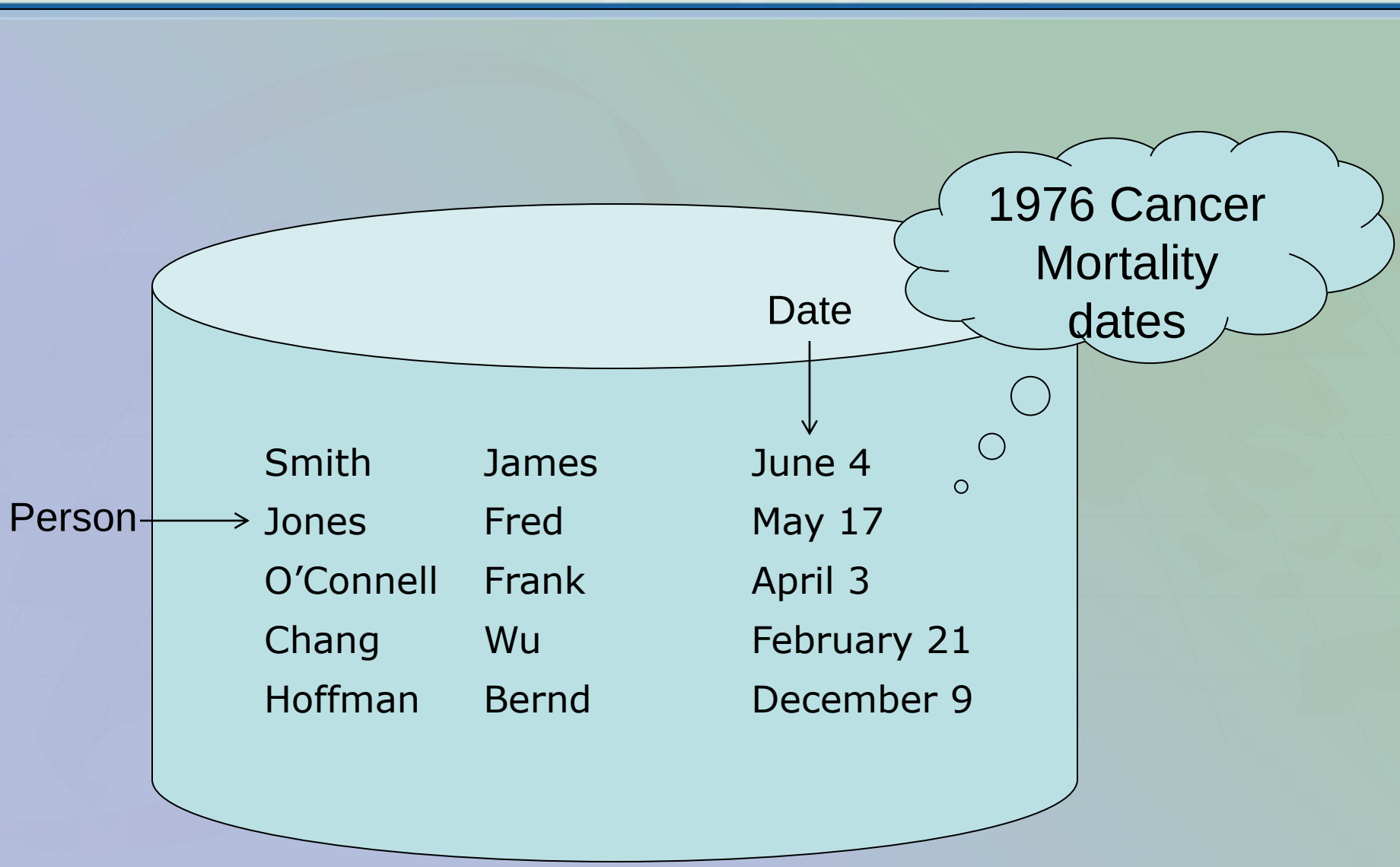
It’s not enough just to describe the data elements…



Describing a dataset … requires a context



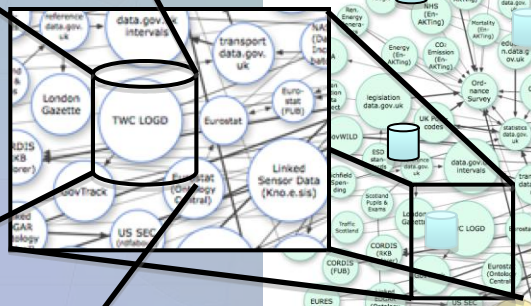
Describing a dataset … requires a context
How do we capture more of this information?



Challenge: Linked Metadata
“A Small World of Big Data”

Tetherless World Constellation

Data



MetaData

Linked Semantic Web “metadata” documents can be used to link very large databases in distributed data systems. This leads to orders of magnitude reduction in information flow for large-scale distributed data problems.

- Open data is becoming "Broad Data"
  - World Wide Web trend towards more and more varied data
    - In many domains
      - E-commerce, Open Govt, many more (cf. Health/Medical care)
- Broad data requires thinking outside the "Database" box
  - DIVE: discover, integrate, visualize, explain
- Broad data requires
  1. Modern, Web-oriented metadata
  2. LINKING the metadata, not the data