Ontology Virtualization for Smart Data

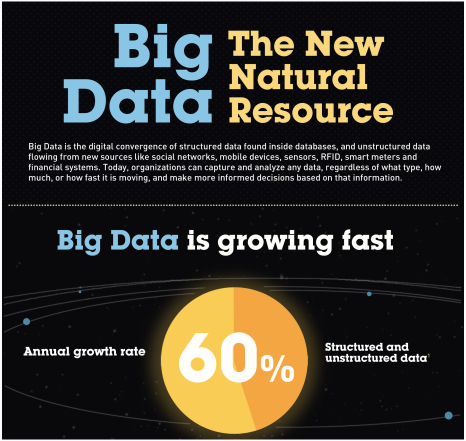
A Semantics Perspective on Open Data Sharing

Krzysztof Janowicz

STKO Lab, University of California, Santa Barbara, USA

RDA Metadata and Semantics Workshop, Feb. 23, 2015

Big Data As The New Natural Resource



Big Data The New Natural Resource

Big Data is the digital convergence of structured data found inside databases, and unstructured data flowing from new sources like social networks, mobile devices, sensors, RFID, smart meters and financial systems. Today, organizations can capture and analyze any data, regardless of what type, how much, or how fast it is moving, and make more informed decisions based on that information.

Big Data is growing fast

Annual growth rate 60% Structured and unstructured data[1]

http://www.ibmbigdatahub.com/infographic/big-data-new-natural-resource

Big Data As The New Natural Resource

Big Data The New Natural Resource

Big Data is the digital convergence of structured data found inside databases, and unstructured data flowing from new sources like social networks, mobile devices, sensors, RFID, smart meters and financial systems. Today, organizations can capture and analyze any data, regardless of what type, how much, or how fast it is moving, and make more informed decisions based on that information.

Big Data is growing fast

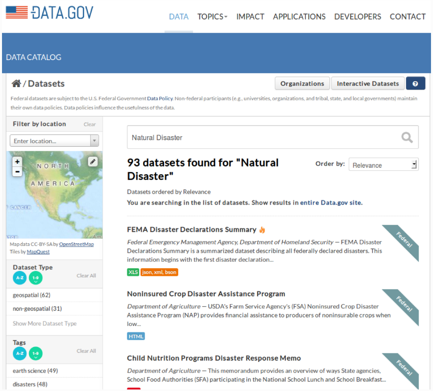
http://www.ibmbigdatahub.com/infographic/big-data-new-natural-resource

A suitable analogy?

- **Natural**, i.e., not man-made
- **Exhaustible**, finite quantity
- **Renewable**, replenishable
- **Consumed**, altered
- **Building block**

If we don't even **understand** what data are, how should we make **sense** out of them?

The Data Retrieval Problem Is Real



/ Datasets

Federal datasets are subject to the U.S. Federal Government Data Policy. Non-federal participants (e.g., universities, organizations, and tribal, state, and local governments) maintain their own data policies. Data policies influence the usefulness of the data.

Map data CC-BY-SA by OpenStreetMap
Tiles by MapQuest

geospatial (62)
non-geospatial (31)

earth science (49)
disasters (48)

Natural Disaster

93 datasets found for "Natural Disaster"

Datasets ordered by Relevance

You are searching in the list of datasets. Show results in entire Data.gov site.

FEMA Disaster Declarations Summary

Federal Emergency Management Agency, Department of Homeland Security — FEMA Disaster Declarations Summary is a summarized dataset describing all federally declared disasters. This information begins with the first disaster declaration...

XLS json, xml, bson

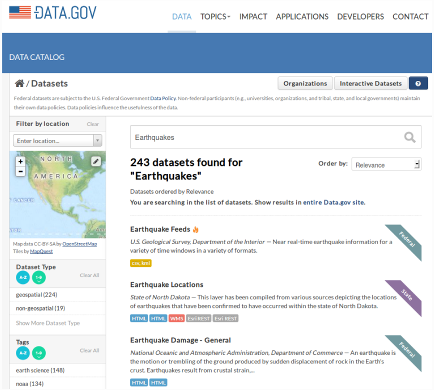
Noninsured Crop Disaster Assistance Program

Department of Agriculture — USDA's Farm Service Agency's (FSA) Noninsured Crop Disaster Assistance Program (NAP) provides financial assistance to producers of noninsurable crops when low...

HTML

Child Nutrition Programs Disaster Response Memo

Department of Agriculture — This memorandum provides an overview of ways State agencies, School Food Authorities (SFA) participating in the National School Lunch and School Breakfast...



/ Datasets

Federal datasets are subject to the U.S. Federal Government Data Policy. Non-federal participants (e.g., universities, organizations, and tribal, state, and local governments) maintain their own data policies. Data policies influence the usefulness of the data.

Map data CC-BY-SA by OpenStreetMap
Tiles by MapQuest

geospatial (224)
non-geospatial (19)

earth science (148)
noaa (134)

Earthquakes

243 datasets found for "Earthquakes"

Datasets ordered by Relevance

You are searching in the list of datasets. Show results in entire Data.gov site.

Earthquake Feeds

U.S. Geological Survey, Department of the Interior — Near real-time earthquake information for a variety of time windows in a variety of formats.

csv, kml

Earthquake Locations

State of North Dakota — This layer has been compiled from various sources depicting the locations of earthquakes that have been confirmed to have occurred within the state of North Dakota.

HTML HTML WMS Esri REST Esri REST

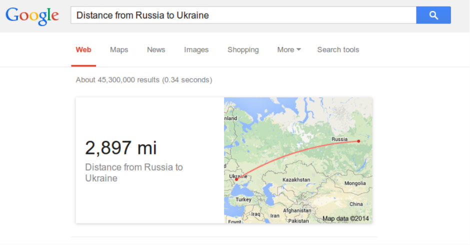
Earthquake Damage - General

National Oceanic and Atmospheric Administration, Department of Commerce — An earthquake is the motion or trembling of the ground produced by sudden displacement of rock in the Earth's crust. Earthquakes result from crustal strain,...

HTML HTML

Even the major data hubs such as Data.gov still rely on keyword-based search and have **unreliable, incomplete, and missing metadata**. For this type of retrieval problems, even '**a little semantics goes a long way**' (Hendler 1997).

Sensemaking is Difficult – Fitness for Puspose is Key

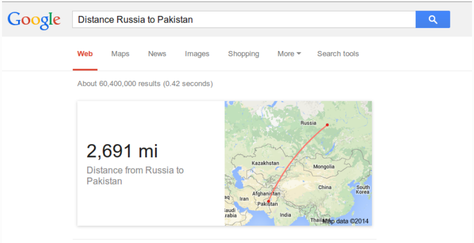


Distance from Russia to Ukraine

About 45,300,000 results (0.34 seconds)

2,897 mi

Distance from Russia to Ukraine



Distance Russia to Pakistan

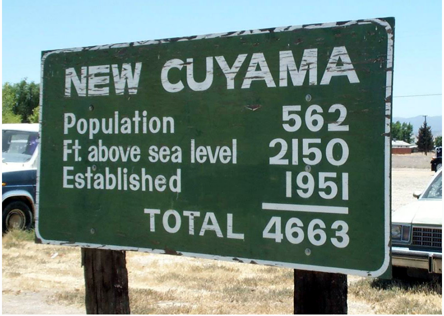
About 60,400,000 results (0.42 seconds)

2,691 mi

Distance from Russia to Pakistan

- There is no shortage of data, but finding data that is **fit for a certain purpose** is difficult.
- Data as **statements** not as truth.
- **Heterogeneity** is caused by cultural differences, progress in science, viewpoints, granularity, ...
- **semantics does not come for free.**
- Lack of **provenance** information
- **Sensemaking** requires more powerful semantic technologies and ontologies (compared to IR).

Meaningful Analysis and Synthesis is Difficult

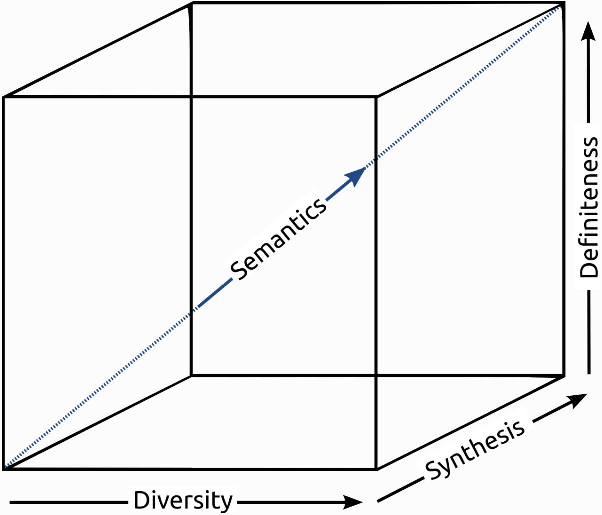


- Ensuring that data is **analyzed and combined** in a meaningful way is far from trivial.
- **What if the information on how to use the data would come together with these data?**
- Focus on **smart data** instead of (merely on) smart applications.
- The purpose of ontologies is not to agree on the meaning of terms but to make the data provider's **intended meaning explicit**.

A little experiment: The statement *all rivers flow into other water bodies* is not useful because it is **'true'**[1], but because...?

[1] It is not; rivers can flow into the ground or just dry up entirely before reaching another water body.

The Semantic Cube



http://goo.gl/fBHie6

Value Proposition

Why use Semantic Web, Linked Data, and Ontologies?

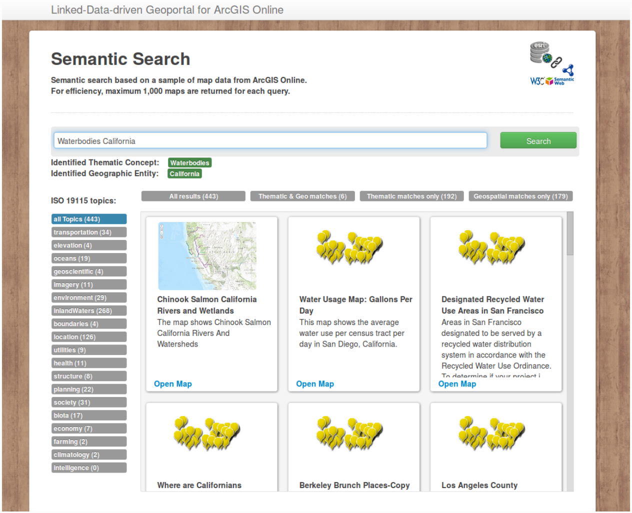
- **Federated queries** over multiple data sources
- **Unique global identifiers** easy conflation and deduplication
- **Transparent data model**; reduces the need for guessing
- **No data silos**, no API restrictions
- Many pre-defined **lightweight vocabularies** (ontologies)
- **Smart data** reduces the need for smart applications
- **Machine reasoning** support
- **No** need for agreement, ontologies make hidden assumptions explicit
- Does away with the **data – metadata** distinction!

The Smart Data Argument

One of the key arguments underlying the Linked Data paradigm is to ***make data smart****, not applications. Instead of developing increasingly complex software, the so-called business logic should be moved to the (meta)data. The rationale is that smart data will make all future applications more usable, flexible, and robust, while* ***smarter applications fail to improve data*** *along the same dimensions.*

(**http://goo.gl/fBHie6**)

Smart Data enables Semantic Search



Linked-Data-driven Geoportal for ArcGIS Online

Semantic Search

Semantic search based on a sample of map data from ArcGIS Online.
For efficiency, maximum 1,000 maps are returned for each query.

Waterbodies California | Search

Identified Thematic Concept: Waterbodies
Identified Geographic Entity: California

ISO 19115 topics:

All results (443) | Thematic & Geo matches (6) | Thematic matches only (192) | Geospatial matches only (179)

- all Topics (443)
- transportation (34)
- elevation (4)
- oceans (19)
- geoscientific (4)
- imagery (11)
- environment (29)
- inlandWaters (268)
- boundaries (4)
- location (126)
- utilities (9)
- health (11)
- structure (8)
- planning (22)
- society (31)
- biota (17)
- economy (7)
- farming (2)
- climatology (2)
- intelligence (0)

Chinook Salmon California Rivers and Wetlands
The map shows Chinook Salmon California Rivers And Watersheds
Open Map

Water Usage Map: Gallons Per Day
This map shows the average water use per census tract per day in San Diego, California.
Open Map

Designated Recycled Water Use Areas in San Francisco
Areas in San Francisco designated to be served by a recycled water distribution system in accordance with the Recycled Water Use Ordinance. To determine if your project i
Open Map

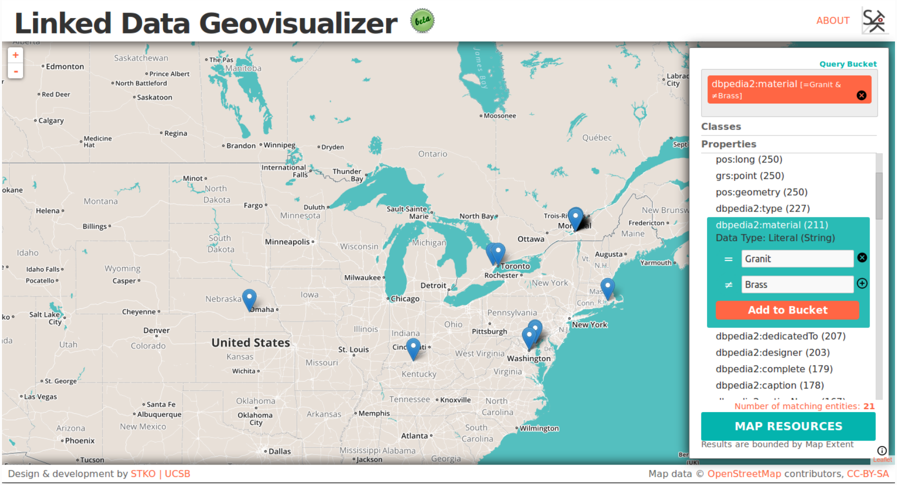
Where are Californians

Berkeley Brunch Places-Copy

Los Angeles County

Creating enriched, semantically-lifted **Linked Data** on top of Esri's ArcGIS Online

Flexible Pattern-based Data Exploration



Linked Data Geovisualizer

Query Bucket

dbpedia2:material [=Granit & ≠Brass]

Classes

Properties

- pos:long (250)
- grs:point (250)
- pos:geometry (250)
- dbpedia2:type (227)
- dbpedia2:material (211)
  Data Type: Literal (String)
  = Granit
  ≠ Brass
  Add to Bucket
- dbpedia2:dedicatedTo (207)
- dbpedia2:designer (203)
- dbpedia2:complete (179)
- dbpedia2:caption (178)

Number of matching entities: 21

MAP RESOURCES

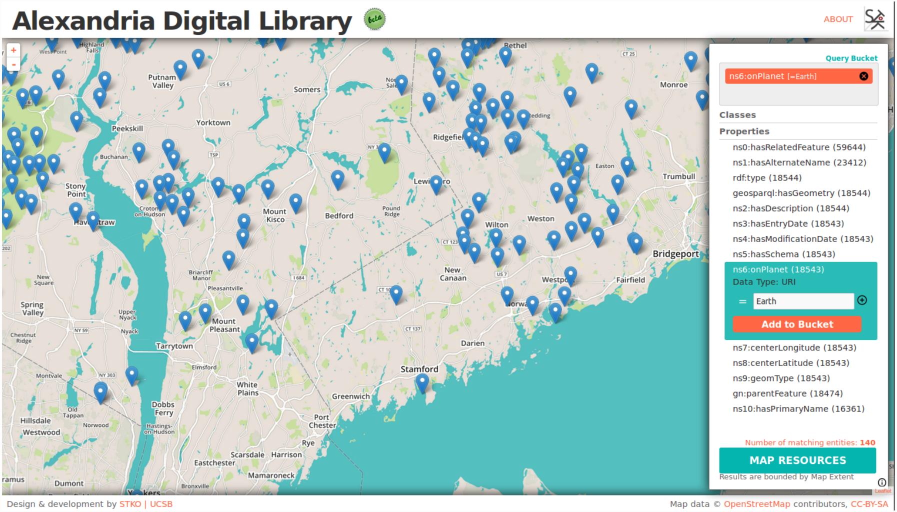
Results are bounded by Map Extent

Design & development by STKO | UCSB

Map data © OpenStreetMap contributors, CC-BY-SA

An user-friendly interface on top of the **DBpedia** SPARQL endpoints.

Flexible Pattern-based Data Exploration


Alexandria Digital Library

Query Bucket

ns6:onPlanet [=Earth] ⊗

Classes

Properties

- ns0:hasRelatedFeature (59644)
- ns1:hasAlternateName (23412)
- rdf:type (18544)
- geosparql:hasGeometry (18544)
- ns2:hasDescription (18544)
- ns3:hasEntryDate (18543)
- ns4:hasModificationDate (18543)
- ns5:hasSchema (18543)
- ns6:onPlanet (18543)
  Data Type: URI
  = Earth ⊕
  Add to Bucket
- ns7:centerLongitude (18543)
- ns8:centerLatitude (18543)
- ns9:geomType (18543)
- gn:parentFeature (18474)
- ns10:hasPrimaryName (16361)

Number of matching entities: **140**

MAP RESOURCES

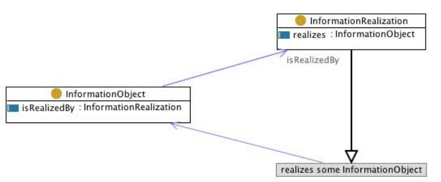
Results are bounded by Map Extent

Design & development by STKO | UCSB

Map data © OpenStreetMap contributors, CC-BY-SA

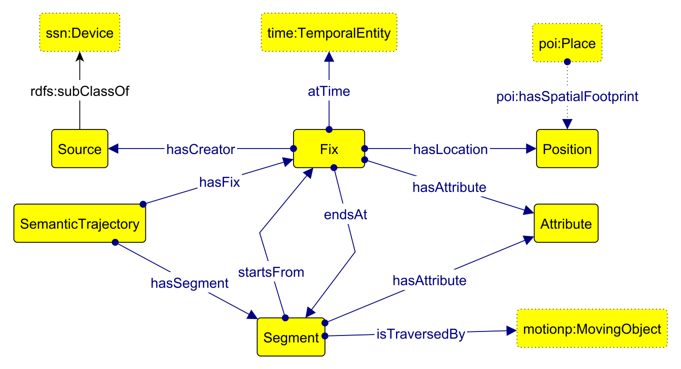
An user-friendly interface on top of the **ADL Gazetteer** SPARQL endpoints.

Ontology Design Pattern in a Nutshell


InformationRealization
realizes : InformationObject
isRealizedBy
InformationObject
isRealizedBy : InformationRealization
realizes some InformationObject

- **Modular** but **self-contained** building blocks
- Some patterns are **strategies**
- **Reusable** and **extendible**
- Even huge ontologies can be modularized using ODP (for example **DOLCE**)
- **No need** to import **full** ontology and **all** *ontological* ***commitments***
- Different **types** of patterns, e.g. **content** vs. logical
- How **many patterns** are there?

A (More Complex) Semantic Trajectory Pattern



A pattern for **discrete** trajectories of people, wildlife, vessels, and so forth.

Ontology Design Patterns Can Be Specialized

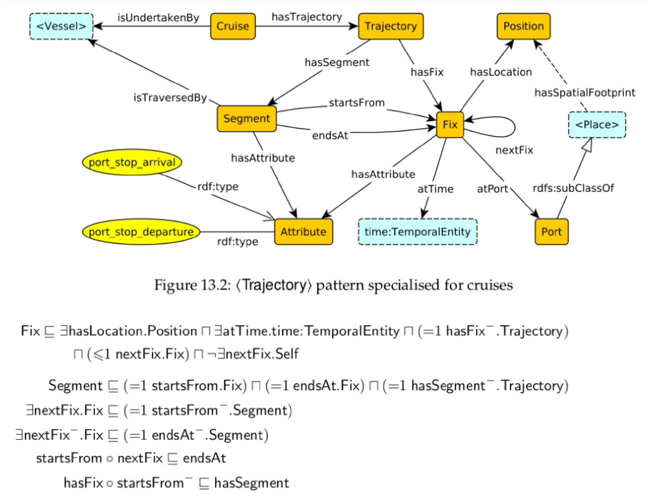


Figure 13.2: ⟨Trajectory⟩ pattern specialised for cruises

$$\mathsf{Fix} \sqsubseteq \exists \mathsf{hasLocation.Position} \sqcap \exists \mathsf{atTime.time{:}TemporalEntity} \sqcap (=1\ \mathsf{hasFix}^{-}.\mathsf{Trajectory}) \sqcap (\leqslant 1\ \mathsf{nextFix.Fix}) \sqcap \neg \exists \mathsf{nextFix.Self}$$

$$\mathsf{Segment} \sqsubseteq (=1\ \mathsf{startsFrom.Fix}) \sqcap (=1\ \mathsf{endsAt.Fix}) \sqcap (=1\ \mathsf{hasSegment}^{-}.\mathsf{Trajectory})$$

$$\exists \mathsf{nextFix.Fix} \sqsubseteq (=1\ \mathsf{startsFrom}^{-}.\mathsf{Segment})$$

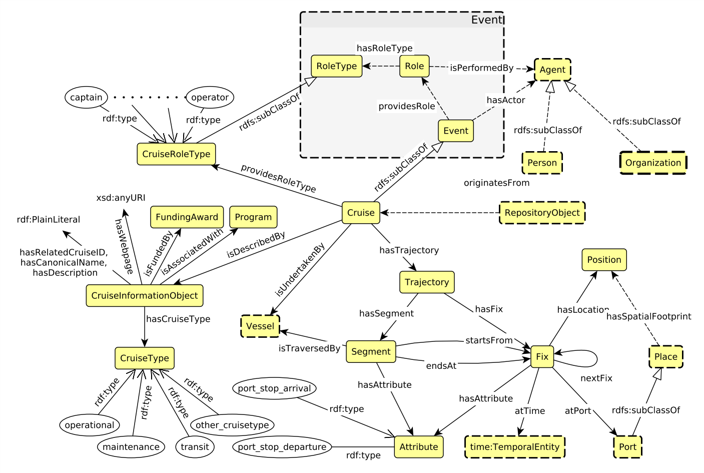
$$\exists \mathsf{nextFix}^{-}.\mathsf{Fix} \sqsubseteq (=1\ \mathsf{endsAt}^{-}.\mathsf{Segment})$$

$$\mathsf{startsFrom} \circ \mathsf{nextFix} \sqsubseteq \mathsf{endsAt}$$

$$\mathsf{hasFix} \circ \mathsf{startsFrom}^{-} \sqsubseteq \mathsf{hasSegment}$$

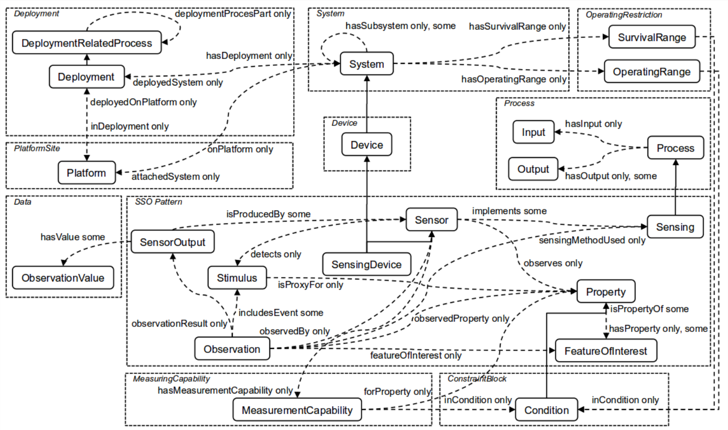
Trajectories that model the research **cruises** of scientific vessels

A Micro-Ontology for Cruises



Combining the InformationObject, Event, Vessel, and Trajectory patterns

W3C Semantic Sensor Network XG Ontology



The Ontology Standartization Argument

Given the early success of data format standardization, we assume that standardizing meaning (via ontologies) is less difficult and more persistent than aligning and translating local micro- ontologies. What if standardization is the more difficult task ?

(**http://goo.gl/2e751**)

Towards Ontology Virtualization

In analogy to hardware virtualization: given a set of ontology design patterns and their combination into micro-ontologies, we can abstract from the underlying axiomatization by:

- Dynamically reconfiguring patterns in a **plug&play** style
- **Bridging** between different patterns an micro-theories
- Providing ontological **views** and **semantic shortcuts** that suit particular provider, user, and use case needs by highlighting or hiding certain aspects of the underlying ontological model
- Map between major **modeling styles**,e.g., the use of instances versus classes

How do we handle different **ontological commitments?**
Quine: To be is to be the value of a (bound) variable
Example: Transportation is moving goods from one location to another.