Semantic Web for Earth Science

Mike Dean
mdean@bbn.com

EarthScienceOntolog Panel Session 1
23 August 2012

Overview

- Semantic Web technologies appear to be widely applicable to large scale earth science data management and applications
- General
  - Ontologies
  - Linked Data
- Specific emerging technologies
  - GeoSPARQL
  - RDF Data Cube Vocabulary
  - RDB to RDF
  - Provenance

Ontologies

- Capture terms and relationships in a form amenable to automation
- "Schema on steroids"
- Models the world, not the data
  - E.g. every Person has 1 father, vs. DBMS integrity constraint 0 or 1 fathers
  - Can accommodate inferred and unknown data
- Ideally, fully distinguishes subclasses via their properties
- Generally useful to capture domain knowledge, even if it isn't initially used, e.g. someone can't be their own father

Linked Data

- Large collection of interlinked data sets using Semantic Web standards
  - 295+ data sets
  - 31+ billion RDF statements
- Includes DBpedia, Geonames, LinkedGeoData, lots of life science data, etc.
- Increasing focus on authoritative sources
  - OrdnanceSurvey, USGS, IGN
- Provides URIs for many common objects
- http://linkeddata.org - cloud diagram
- http://linkeddatabook.com - principles and best practices

GeoSPARQL

- New Open Geospatial Consortium standard for representing and querying geospatial information
- Supports multiple
  - Geometries (point, lines, polygons)
  - Coordinate reference systems
  - Qualitative relations (within, intersects, etc.)
- Preferred vocabulary for publishing new geospatial data
- http://www.opengeospatial.org/standards/geosparql
- Parliament GeoSPARQL is an open-source implementation

RDF Data Cube Vocabulary

- Vocabulary for publishing multi-dimensional data, such as statistics, as Linked Data
- Supports units of measure and slices
- Could presumably be extended for “stand off” annotation of large datasets
- http://www.w3.org/TR/vocab-data-cube/

RDB to RDF

- Much of the data on the (Semantic) Web resides in relational databases
- W3C has 2 Proposed Recommendations for accessing such data
  - RDB to RDF Mapping Language (R2RML)
  - Direct Mapping of Relational Data to RDF
- These or similar approaches could be used to dynamically access other forms of structured data

Provenance

- Traceability of data from its source through various processing transformations is important
- W3C PROV addresses
  - Entities (e.g. documents), including Alternates
  - Activities (e.g. creation)
  - Agents (e.g. people, organizations, software)
  - Roles (e.g. editor)
  - Plans (e.g. workflows)
  - Derivation and Revision
  - Timestamps
- http://www.w3.org/2011/prov/wiki/Main_Page
  - Start with the PROV Primer
  - Several documents are Last Call Working Drafts