Open Health Data:
Link and Mashup Public Healthcare Data

Tim Lebo, Jim McCusker, Dominic Difranzo, Li Ding, Deborah McGuinness

Tetherless World Constellation
Rensselaer Polytechnic Institute
Outline

• Health data integration
• Open health data examples
  – Tobacco policy
  – Biomedical experiments
• Collaborations
• Future work
Healthcare Domain

• Diverse audiences
  Citizen consumers --- Policy makers
  Patients --- Health providers
  Researchers --- Funding organizations

• Diverse questions
  – Does raising tobacco taxes help reduce smoking?
  – Should the government spend more on Smoker Quit lines?
  – What is the life expectancy in my area?
  – Where can I live to avoid cigarette smoke?
  – Where can I live to get the most benefit from Medicare?
  – What kinds of organizations fund gene research?
Improve Health with Open Health Data

- Diverse audiences
- Diverse questions
- Diverse data
  - Data.gov provides 60+ raw datasets, e.g.
    - Veterans Health Administration 2008 Hospital Report Card - Medical Center Staffing (Department of Veterans Affairs)
    - Electronic Animal Drug Product Listing Directory (HHS)
    - OMHA Appeals Listed By State (HHS)
    - Medicare Cost Report Data (HHS)
    - SSA Disability Claim Data (SSA)
  - Other Federal/State Agencies, e.g. NIH, NCI, NLM, FDA
  - Linked Open Data, e.g. PubMed, Drug Bank, GeneID, …
  - Media, e.g. New York Times
  - Social Media, e.g. Nike+
State Tax per Pack of Cigarettes, 2000-2007

How has each state’s cigarette tax policy developed?

Source data: Orzechowski, Walker (2007). The Tax Burden on Tobacco

Tennessee (2003) $0.20

http://tw2.tw.rpi.edu/demos/lod-scigrid_tax/verbatim.html
Tobacco Tax and Unemployment

Should we consider unemployment in our new tobacco policy?

Orzechowski, Walker (2007). The Tax Burden on Tobacco

RDF: Linked (tax) Data

http://dbpedia.org/

<http://dbpedia.org/resource/Alabama>


owl:sameAs


State Tobacco Tax.xls

Workflow Provenance in Biomedical experiments

- 12,000 ArrayExpress experiments encoded in RDF
- Gene Ontology, NCI Thesaurus, et al. encoded in Simple Knowledge Organization System (SKOS)
- Identity ontology for various medical-specific senses of “same as”
  - Initial uses to infer gene function across species
- Next steps
  - Use provenance as a common model for integrating clinical, laboratory, experimental, and computational observations, data, and findings.
  - Provide automated evaluation of justifications for belief assertions.
Collaborations

• U.S. Department of Health & Human Services (HHS)
  – Community Health Data Initiative (CHDI)
    • http://www.cdc.gov/nchs/data_access/chdi.htm
  – Semantic version of National Library of Medicine’s MEDLINE
    • Ontology and SPARQL endpoint

• National Cancer Institute (NCI)
  – Population Science Grid (PopSciGrid)
Application Layer (e.g., Enhanced disease modeling, dashboards, data widgets…)

Grid Cyberinfrastructure

Common Vocabularies (Shared ontologies, common data elements)

Public Surveillance
- BRFSS
- HINTS
- NHIS
- Tax
- US Census,…

Grantees
- CECCRS
- CPHHD
- GEI
- TREC
- TTURCS

Biomedical
- Biological
- Genomic/proteomic

Clinical/Health System
- CRN
- caBIG PopSci SIG
- QCCC projects
- Registries (SEER)

Community/Contextual
- ‘Community health labs’
- GIS (geo-spatial data)
- Physical/Built environment
- Real-time data capture
Future Work

- Integrating additional data
- Providing additional presentation techniques
- Privacy management
- Dynamically expandable tools for providing customizable presentations.

http://data-gov.tw.rpi.edu/wiki/Open_health_data