Ontology-enabled Data Resources for Environmental Health Research and Decision Support

Deborah L. McGuinness
Tetherless World Senior Constellation Chair
Professor of Computer, Cognitive, and Web Science
Director RPI Web Science Research Center
RPI Institute for Data Exploration and Application Health Informatics Lead
dlm@cs.rpi.edu, @dlmcguinness,

Emerging Science for Environmental Health Decisions Workshop
Informing Environmental Health Decisions Through Data Integration Workshop
Good Decisions and Data Resources …
Integrated Understandable Data

Limited data integration without controlled vocabulary
Limited reproducibility without shared definitions
Difficulty in reuse without provenance

Ontologies can enhance integration, communication, reuse, and research impact
What is an Ontology?

An ontology specifies a rich description of the

- Terminology, concepts, nomenclature
- Relationships among concepts and individuals
- Sentences distinguishing concepts, refining definitions & relationships

relevant to a particular domain or area of interest.


*Based on AAAI ‘99 Ontologies Panel – McGuinness, Welty, Uschold, Gruninger, Lehmann*
Ontology Development Process

**Foundational Ontologies/Vocabularies**

- Use Cases
- Existing Ontologies & Vocabularies
- Data Reporting Templates
- Data Dictionaries / Codebooks
- Expert Interviews

**Labkey, Ontology Fragments**

- Semantic Extract Transform, Load (SETLr)
- Expert Guidance

**Ontology Curation (ongoing)**

- Generated Ontology
  - domain concepts
  - authoritative vocabularies
  - vetted definitions
  - supporting citations

- Human Aware Data Acquisition Framework
  - Ontology Browser

**Knowledge Graph Integration**

- Linking data and metadata content to domain terms
- Linking workflows based on semantic descriptions

**Repository Integration**

- Source Datasets
- Analytics source code
- Results
- Publications

**Knowledge-Enhanced Search**

- Finding what is there that might be of use

**Reviewers & Curators**

- Ontology Development Team
- Domain collaborators
- Invited experts
- "Consumers" (data analysts)

Exemplified by CHEAR

McGuinness NASEM 2/20/18

Erickson, McGuinness, McCusker, Chastain, Pinheiro, Rashid, Liang, Liu, Stingone, …
Support Browsing, Searching, Pooling, Deriving Values, Verification, …

- Ontology support for mapping and integration (e.g., education level)
- Ontology informs decisions about variables that may be combined, serve as proxy, or used to derive desired info (e.g., birth outcomes)
- Ontology Integrity constraints may help flag errors (e.g., APGAR > 10)
- Ontology helps expose implicit information and find links

<table>
<thead>
<tr>
<th>Mother’s Highest Education Level</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did not attend school</td>
<td>0</td>
</tr>
<tr>
<td>Elementary school</td>
<td>1</td>
</tr>
<tr>
<td>Technical post-primary</td>
<td>2</td>
</tr>
<tr>
<td>Middle school</td>
<td>3</td>
</tr>
<tr>
<td>Technical post-middle school</td>
<td>4</td>
</tr>
<tr>
<td>Highschool or junior college</td>
<td>5</td>
</tr>
<tr>
<td>Technical post-junior college</td>
<td>6</td>
</tr>
<tr>
<td>College</td>
<td>7</td>
</tr>
<tr>
<td>Graduate</td>
<td>8</td>
</tr>
<tr>
<td>Doesn’t know</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mother Education</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than High School</td>
<td>0</td>
</tr>
<tr>
<td>High School Graduate or More</td>
<td>1</td>
</tr>
</tbody>
</table>

McGuinness, McCusker, Pinheiro, Stingone, et. al. Funding: NIH/NIEHS 0255-0236-4609 / 1U2CES026555-01
Example Ontology and infrastructure (CHEAR)

**Hundreds** of concepts from the CHEAR ontology
- 258 Analytes
- 176 Epidemiological Attributes
- 28 Sample types
- 42 Assay types
- 122 Instrument types

**Thousands** of instances of CHEAR & foundational ontologies (e.g., subjects, samples, lab capabilities)

We use:
- **Labkey** to create, curate and maintain CHEAR concepts (ontology)
- **Labkey** to create and maintain CHEAR instances (knowledge graph)
- **SETLr** to build and publish the CHEAR ontology from CHEAR concepts
- **HADatAc** to connect CHEAR/foundational concepts and instances to CHEAR data
- **HADatAc** to browse/select/retrieve CHEAR data from CHEAR vocabulary

**Thousands** of concepts and relationships from foundational ontologies

- HAScO (Instruments/methods)
- SIO (Semantic Science Int Ont)
- PROV (Provenance)
- Disease Ontology
  - UBERON (Anatomy)
- Units Ontology
- CHEBI (Chemicals)
- RefMet (Metabolites)
- ENVO (Environment)
- UniProt (Proteins)
Ontology and Knowledge Graph (Behind the Scenes)

Concepts / relationships from foundational ontologies

Examples of concepts from the CHEAR ontology

Examples of terms from the CHEAR ontology

Instance of foundational ontology term

Epidemiological Measurements

McGuinness NASEM 2/20/18 Partially supported by: NIH/NIEHS 0255-0236-4609 / 1U2CES026555-01
Mapping Data to Meaning: Semantic Data Dictionaries


McGuinness NASEM 2/20/18  Partially supported by: NIH/NIEHS 0255-0236-4609 / 1U2CES026555-01
Laboratory Information Mgmt System (LIMS)-based backend integrated with Ontology.

Includes automatic ingest, access control, data governance, download, ...

Supports Search study, data sample, subject, ...

Enables statisticians to ask for content to support their studies e.g., find
Child: Birth Weight, Gender, Gestational Age at Birth
Mother: Age, BMI “early in pregnancy based on inclusion criterion for the particular study”, Parity, Education
Metals: As, CD, Mn, Mo, Pb

McGuinness NASEM 2/20/18

CHEAR Human Aware Data Acquisition Framework

Sample Question: Gennings

Partially supported by: NIH/NIEHS 0255-0236-4609 / 1U2CES026555-01
Ontology-Enabled Study Search / Precision Data Search and Download

Blood Biomarkers for Children’s Health (Study 1)

Institution:
Principal Investigator(s):
Number of Subjects:
Number of Samples:
Study Description:
Keywords:

Urine Biomarkers for Children’s Health (Study 2)

Institution:
Principal Investigator(s):
Number of Subjects:
Number of Samples:

Metabolomic Biomarkers for Children’s Health (Study 3)

Institution:
Principal Investigator(s):
Number of Subjects:
Number of Samples:
Study Description:
Keywords:
• Domain Science requires and is asking for ontologies for Findable, Accessible, Interoperable, Reusable (FAIR) support
• With tooling and processes, domain scientists can help build and maintain ontologies and ontology-enabled applications, ex. Epidemiologists are doing this
• The data center content in CHEAR with the human aware data collection framework knowledge infrastructure exemplifies ontology-enabled data ingest, integration, precision data search and download
• Moving forward these infrastructures can enable cognitive agent decision support currently exemplified with the IBM –RPI AI horizons network project
IBM and Rensselaer Team To Research Chronic Diseases With Cognitive Computing

IBM (NYSE: IBM) and Rensselaer Polytechnic Institute today announced the creation of the new Center for Health Empowerment by Analytics, Learning, and Semantics (HEALS). Located on the Rensselaer campus, the HEALS center is a five-year collaborative research effort aimed at researching how the application of advanced cognitive computing capabilities can help people to understand and improve their own health conditions.

“This collaboration between Rensselaer and IBM, which combines our significant research strengths in cognitive computing, could generate insights which will aid clinicians with more effective treatments for individual patients and overall efficiencies in the health care system,” said
Health Empowerment by Analytics, Learning, and Semantics

- How can we enhance population and individual health using information found inside and outside the traditional (E)HR?
- How can we develop precision medicine across the many levels of research from Genome to Phenotype to Population Health?
- How can we use IBM’s Watson Technology, augmented with Rensselaer’s semantics, learning, and analytics expertise to achieve these goals?

McGuinness NASEM 2/20/18 Partially supported through IBM Cognitive Network funding
Ontologies are an important piece; but are part of a larger integrated framework. In this case, enabling initially the clinical oncologist of the future.

Semantics-enabled Framework

**

Ontologies are an important piece; but are part of a larger integrated framework. In this case, enabling initially the clinical oncologist of the future.

Semantics/ SemNExt RPI team:
McGuinness, Bennett PIs with McCusker, Erickson, Seneviratne and the extended Research groups along with input from IBM HEALS collaborators and motivation from MANY projects, particularly from NIEHS/ Mount Sinai.
• Knowledge imported from drug, protein, and disease interaction databases.
• Each interaction given an evidence-driven probability.
• Find drugs that could affect melanoma, filtered by interaction probability.
• The best hypotheses were generated using the highest probabilities.
• Being expanded initially for genomics-aware cancer care applications.
Recommendations

- Ontologies enable FAIR (Findable, Accessible, Interoperable, Reusable) Data Resources
- They can support movement across levels of abstraction
- Use ontology-enabled architectures
- Do NOT build ontologies from scratch
- Selectively and thoughtfully reuse existing best practice ontologies/vocabularies
- Engage experts in choosing ontology (portions) and in designing the knowledge architecture
- Ecosystems and diverse teams are critical for success – community driven and maintained ontology-based systems are the future
- Lets enable the next generation of decision support together!

McGuinness NASEM 2/20/18
Questions?

• Contact: dlm@cs.rpi.edu

• Thanks to many: RPI Tetherless World team particularly McCusker, Erickson, Hendler, Pinheiro, Rashid, Liang, Liu, Chastain; RPI: Bennett, Dyson, Seneviratne; Mount Sinai particularly Teitelbaum, Stingone, Mervish, Gennings, Kovatch; IBM, particularly Das, Chen, Brown, ....

• Funding: NIEHS 0255-0236-4609 / 1U2CES026555-01, DARPA HR0011-16-2-0030, IBM-RPI HEALS, NSF ACI-1640840

• Forthcoming book: Ontology Engineering with Kendall
Initial data analysis aimed initially at the following public datasets:

**TCGA:** RNA expression, tumor mutation, protein expression, and clinical attributes (including staging, treatment, risk, and survival) on 32 cancer types in > 14,000 patients

**NHANES:** Cross-sectional biannual survey of the health and nutrition of the US population, including illness, environmental exposures, and risk exposures.

Additional analysis will include deidentified data in cancer topics.
Building the Knowledge graph:
Reusing Knowledge Sources to Bridge Abstractions

- **Already done**: COSMIC Gene Census, OMIM, DrugBank, iRefIndex
- **Pathway data**: KEGG, Reactome (small molecule interactions, curated interactions)
- **Gene Ontology**: protein localization in cell types and tissues, protein functions, biological process involvement
- **UniProt**: Protein families, including common binding sites
- **CAP Protocols**: Current cancer staging standards, NCCN… many of these evolve, e.g., breast cancer staging guidelines
- **Vocabularies**: SNOMED, NCI Thesaurus, NCI Metathesaurus, etc.
Value Propositions Matter to Get and Keep Collaborators

What will we be able to do that is hard or impossible today? One set of topics from an applied mathematician collaborator (Bennett)

- How to merge data from heterogeneous data sources for analysis
- What types of data are available for analysis
- What interesting analysis questions we are capable of asking
- Is a potential analysis question too broad or imprecise for the data
- Which adjustment covariates should be used for a given analysis question
- Which statistical and machine learning methods and workflows are appropriate
- What background information might be relevant for an analysis question
- If measurements are plausible and can be trusted
- Are there explanations of derived results/hypotheses in literature
- Are results similar to those of prior analyses
- What are appropriate ways to visualize and present results to user
- Should changes in data trigger a reanalysis/new analysis of questions of interest
Hypothesis:
*Does factor increase odds of disease?*

**User Specifies:**
- Data (NHANES Cohorts)
- Disease Definition
- Confounders (age, BMI)
- Factors (pesticides)

Agent dynamically applies standard risk analysis workflow based on log-odds

Applicable to any risk problem and data sets

**WORKFLOW**

- Ingest and Clean Data
- Specify Data
- Specify Model
- Knowledge Graph
- Semantic Browser
- Conduct Modeling
- Analyze and Visualize Results

Bennett, Erickson, McCusker, McGuinness et al

McGuinness NASEM 2/20/18
Obtain goals from all stakeholders

One analyst’s Goals:

- Integrate analytics with knowledge graphs to select germane data, discover relevant patterns, predict outcomes, and provide interpretations in response to queries from users or cognitive agents.
- Design and demonstrate semantic analytics workflows across the knowledge graph to support precision health inquiries.
- Discover new patterns and predict outcomes to create new knowledge and insights from the knowledge graph with the assistance of a cognitive computing agent.

Bennett, Erickson, McCusker, McGuinness, et al
Ontologies and the Data Life Cycle

Computer understandable specifications of meaning (semantics) support enhanced lifespan & impact of data
## Building and Evolving Ontologies

<table>
<thead>
<tr>
<th></th>
<th>Past</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design</strong></td>
<td>Knowledge Representation (KR) Expert with domain expert access</td>
<td>KR Expert(s) paired with domain experts <strong>AND community</strong></td>
</tr>
<tr>
<td><strong>Population</strong></td>
<td>KR expert learns domain and builds ontology with some external reuse</td>
<td>KR and domain experts determine seed vocabularies and <strong>HEAVILY</strong> leverage them</td>
</tr>
<tr>
<td><strong>Evolution</strong></td>
<td>KR expert heavily involved</td>
<td>KR expert involved in building / customizing tools that domain experts use; Input may include automatic techniques output (e.g., extraction)</td>
</tr>
<tr>
<td><strong>Tool Users</strong></td>
<td>Trained in Computer Science</td>
<td>Trained in Domain ScienceS</td>
</tr>
<tr>
<td><strong>Application Users</strong></td>
<td>Targeted well understood user base</td>
<td>Diverse and evolving user base</td>
</tr>
<tr>
<td><strong>Reuse</strong></td>
<td>Well thought out</td>
<td><strong>Expect the unexpected</strong></td>
</tr>
</tbody>
</table>

McGuinness NASEM 2/20/18
CHEAR Study-based Evolution Strategy

1. Identify terms that can be mapped to existing ontology
2. Identify terms to be added to ontology
3. Describe new terms with definitions and location within existing ontology
4. Compile new terms across multiple studies (e.g. Quarterly)
5. Review and revise updates with stakeholders
6. Incorporate new terms into existing ontology
7. Mappings (e.g. variable names) incorporated into knowledge graph
8. Data into knowledge graph after embargo period

New version Ontology X

Data Structures & Standards Working Group

Data Center

McGuinness NASEM 2/20/18  Partially supported by: NIH/NIEHS 0255-0236-4609 / 1U2CES026555-01
The Children’s Health Exposure Analysis Resource, or CHEAR, is a program funded by the National Institute of Environmental Health Sciences to advance understanding about how the environment impacts children’s health and development. CHEAR provides children’s health researchers access to laboratory analysis of environmental exposures and data analysis consultation at no cost to the investigator.

CHEAR is designed to expand the range of environmental exposures assessed in NIH-funded children’s health studies, including:

1. Studies wishing to expand their analysis to include environmental exposure analysis or assessments of interactions between genes and the environment
2. Studies that have collected environmental exposure data but seek more extensive analysis

Who might use CHEAR and why?