Bringing epidemiology into the Semantic Web

João D Ferreira, Catia Pesquita, Francisco M Couto and Mário J Silva

Graz, July 23th
Introduction

Epidemiological data is heterogeneous

- Data is heterogeneous and from different fields of knowledge
Introduction
Epidemiological data is heterogeneous
Introduction

• Metadata is more *machine-friendly* when taken from controlled vocabularies

• Ontologies bring many advantages:
  – Inference in searching
  – Common vocabulary in shared resources
  – Semantic analysis
Introduction
Ontologies as source of vocabulary

• Searching for data on:
  – Infectious Diseases
  – In Europe

• Inference can help find the relevant resources
Introduction
Ontologies as source of vocabulary

- Common vocabulary
- Metadata in RDF can be resolved to known concepts

```
<rdf:Description rdf:about="resource_123">
  <em:disease rdf:resource="obo;DOID_8469" />
  <em:host rdf:resource="&mesh;D006801" />
...
</rdf:Description>
```
Introduction
Ontologies as source of vocabulary

• Exploration of technologies such as:
  – Semantic similarity
  – Ontology matching

leading to:
  – Pattern recognition
  – Knowledge creation
  – ...

SEMANTIC ANALYSIS
Epidemiology vocabulary

• We need concepts that are relevant in epidemiology
  – Not an ontology from scratch, but a reuse of concepts
NERO

The Network of Epidemiology-Related Ontologies

NERO

DISEASES

SYMPTOMS

GEOGRAPHICAL LOCATIONS

CHEMICAL COMPOUNDS

...
How to choose the ontologies?
• Inspiration from:
  – General epidemiological needs
• Inspiration from:
  – General epidemiological needs
  – The Epidemic Marketplace ([http://epimarketplace.net](http://epimarketplace.net))
NERO
Creation

Epidemiology-related ontologies

NERO

Epidemiological resources

feedback

applied on

EM
Epidemic Marketplace
NERO
NERO requirements

- Set of requirements that ensure:
  - interoperability
  - cohesion

- Requirements inspired on:
  - Epidemiological needs
  - W3C
  - OBO Foundry
NERO
NERO requirements

• Five examples:
  – Relevant domain
  – Textual definitions
  – Synonyms
  – Publicly available
  – Cross-references
Five examples:
- Relevant domain
- Textual definitions
- Synonyms
- Publicly available
- Cross-references

ATTENTION:
These requirements are guidelines
• Three types of ontologies:
  - 1) Ontologies specific to epidemiology
  - 2) Ontologies of generic scope
  - 3) Ontologies focused on a single domain
1) Ontologies specific to epidemiology
- BioCaster Ontology
- Epidemiology Ontology
- Dictionary of Epidemiology

These terminologies lack:
- structure, scope, depth, ...
2) Ontologies of generic scope
- UMLS
- SNOMED-CT
- MeSH

Branches are hard to choose
3) Ontologies focused on a single domain
   - Many ontologies in the OBO Foundry web page
     - Diseases
     - Drugs
     - Vaccines
     - GeoPlanet (from Yahoo!)
   - Insufficient coverage of the field of knowledge
Conclusions

● Current ontologies are not enough to represent epidemiological domain

● Best approach:

  single domain ontologies
  +
  some branches of
generic ontologies
## Conclusions

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Domain</th>
<th>Ref.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>In NERO?</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioCaster</td>
<td>Epidemiology</td>
<td>(Collier et al., 2008)</td>
<td>Y</td>
<td>N</td>
<td>±</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>No</td>
</tr>
<tr>
<td>Epidemiology Ontology</td>
<td>Epidemiology</td>
<td>(HuGE Net, 2007)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Yes</td>
</tr>
<tr>
<td>Dictionary of Epidemiology</td>
<td>Epidemiology</td>
<td>(Porta, 2008)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>±</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>No</td>
</tr>
<tr>
<td>UMLS</td>
<td>General</td>
<td>(Lindberg et al., 1993)</td>
<td>Y</td>
<td>Y-</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>N</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>MeSH</td>
<td>General</td>
<td>(Lipscumb, 2000)</td>
<td>Y</td>
<td>Y-</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>SNOMED-CT®</td>
<td>General</td>
<td>(Stearns et al., 2001)</td>
<td>Y</td>
<td>Y-</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>No</td>
</tr>
<tr>
<td>GeoPlanet™</td>
<td>Geography</td>
<td>(Yahoo!, 2011)</td>
<td>Y</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>GeoNames</td>
<td>Geography</td>
<td>(Geonames.org, 2011)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Geo-Net-PT</td>
<td>Geography</td>
<td>(Lopez-Pelllicer et al., 2009)</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>No</td>
</tr>
<tr>
<td>OBO ontologies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChEBI</td>
<td>Biochemistry</td>
<td>(de Matos et al., 2010)</td>
<td>Y</td>
<td>Y-</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>DOID</td>
<td>Diseases</td>
<td>(Osborne et al., 2009)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>ENVO</td>
<td>Environment</td>
<td>(EnvO developers, 2012)</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>HP</td>
<td>Symptoms</td>
<td>(Robinson and Mundlos, 2010)</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>IDO</td>
<td>Diseases</td>
<td>(Cowell and Smith, 2010)</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>NCBI Taxonomy</td>
<td>Taxonomy</td>
<td>(Wheeler et al., 2007)</td>
<td>Y</td>
<td>Y-</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>NCI Thesaurus</td>
<td>General</td>
<td>(Sioutos et al., 2007)</td>
<td>Y</td>
<td>Y-</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>SYMP</td>
<td>Symptoms</td>
<td>(Schriml et al., 2010)</td>
<td>Y</td>
<td>Y-</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>TRANS</td>
<td>Disease transmission</td>
<td>(Schriml et al., 2010)</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>±</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>VO</td>
<td>Vaccines</td>
<td>(Yang et al., 2011)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>
Contributions

- NERO as a *vocabulary* that can be used to characterize epidemiological resources

- Annotated epidemiological resources can be explored in the context of *semantic web*
  - Information Retrieval & Integration

- NERO enables *other technologies*:
  - Ontology Matching, Semantic Similarity, ...
Acknowledgments

XLDB / LaSIGE
Bringing epidemiology into the Semantic Web

João D Ferreira, Catia Pesquita, Francisco M Couto and Mário J Silva

Find me at: joao.ferreira@lasige.di.fc.ul.pt

Graz, July 23th