ENABLING TAILORED THERAPEUTICS WITH LINKED DATA

Oktie Hassanzadeh  University of Toronto

Join work with
Anja Jentzsch  Freie Universität Berlin
Bo Andersson  AstraZeneca R&D Lund
Susie Stephens  Eli Lilly and Company
Christian Bizer  Freie Universität Berlin

Linking Open Drug Data (LODD) Task Force
http://esw.w3.org/topic/HCLSIG/LODD
Outline

- Linking Open Drug Data Project
  - Objectives and Status
- Published linked data sources
- Interlinking of the data sources
- Business use cases
- Conclusion and Future Work
An HCLSIG task force
- Started October 1st, 2008

Primary Objectives
- Survey publicly available data sets about drugs
- Publish and interlink these data sets on the Web
- Explore interesting questions that could be answered if the data sets are linked
Survey of Existing Data sets

- http://esw.w3.org/topic/HCLSIG/LODD/Data
Drug Data Sources

Source: Mark Sharp, et al. (AMIA 2008)
A Framework for Characterizing Drug Information Sources
Extending LOD cloud
HCLS in LOD cloud

As of March 2009
LODD in LOD cloud

- Published Data Sets
  - LinkedCT
  - Diseasome
  - DailyMed
  - DrugBank

As of March 2009
LODD in LOD cloud

- Interlinked to:
  - DBpedia/YAGO
  - Symbol
  - CAS
  - HGNC
  - KEGG
  - OMIM
  - ChEBI
  - GeneID
  - PubMed
  - UniProt
  - Pfam
  - PDB

As of March 2009
Published Datasets

- **LinkedCT**
  - [http://linkedct.org](http://linkedct.org)
  - From ClinicalTrials.gov
    - Online registry of clinical trials conducted in the United States and around the world
    - Published in XML
  - More than 60,000 trials
  - 7,011,000 triples

- **DrugBank**
  - [http://www4.wiwiss.fu-berlin.de/drugbank/](http://www4.wiwiss.fu-berlin.de/drugbank/)
  - A repository of almost 5000 FDA-approved small molecule and biotech drugs
    - Published as DrugBank DrugCards
  - 1,153,000 triples
Published Datasets

- **DailyMed**
  - [http://www4.wiwiss.fu-berlin.de/dailymed/](http://www4.wiwiss.fu-berlin.de/dailymed/)
  - High quality information about marketed drugs
    - Published by the National Library of Medicine
    - In a flat file representation
  - 124,000 triples

- **Diseasome**
  - [http://www4.wiwiss.fu-berlin.de/diseasome/](http://www4.wiwiss.fu-berlin.de/diseasome/)
  - Information about 4,300 disorders and disease genes linked by known disorder–gene associations
    - Published in Structured Product Labeling (an XML-based standard for exchanging medication information)
  - 88,000 triples
Interlinking Datasets

- Two classes of links
  - Based on common identifiers
    - Links present in the source data sets
  - Based on link discovery and record linkage techniques
    - String matching
    - Semantic matching

As of March 2009
Interlinking Datasets

- Link discovery techniques
  - String matching
    - Linking LinkedCT and Diseasome
      - E.g., "Alzheimer's disease" in LinkedCT was matched with "Alzheimer_disease" in Diseasome
  - Semantic matching
    - Many drugs and diseases have multiple names
      - E.g., "Varenicline" has the synonym "Varenicline Tartrate" and the brand names "Champix" and "Chantix"
## Interlinking Statistics

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkedCT</td>
<td>290,000 links; 50,000 of them inside the LODD cloud</td>
</tr>
<tr>
<td>DrugBank</td>
<td>23,000 links; 8,500 of them inside the LODD cloud</td>
</tr>
<tr>
<td>DailyMed</td>
<td>29,600 links; all of them inside the LODD cloud</td>
</tr>
<tr>
<td>Diseasome</td>
<td>23,000 links; 8,400 of them inside the LODD cloud</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>365,600 links; 8.4 million triples</td>
</tr>
</tbody>
</table>
Business Use Cases

- [http://esw.w3.org/topic/HCLSIG/LODD/Business](http://esw.w3.org/topic/HCLSIG/LODD/Business)

- **Example competitive intelligence use case**
  - A neuroscience focused business manager interested in seeing an update on new clinical trials by competitors on Alzheimer’s Disease (AD).
    - A phase III trial by Pfizer for a drug called Varenicline
      - [http://data.linkedct.org/resource/trials/NCT00744978](http://data.linkedct.org/resource/trials/NCT00744978)
    - More information about the drug on DBpedia, DailyMed and DrugBank
      - [http://dbpedia.org/resource/Varenicline](http://dbpedia.org/resource/Varenicline)
      - [http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01273](http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01273)
    - DailyMed indicates the drug is already on the market for Nicotine addiction
      - Possible side effects are listed for the typical dose
      - According to LinkedCT, the dose in the trial is no more than the typical dose
Why a nicotine addiction drug might work for AD?

- DrugBank allows the manager to find drug targets “Neuronal acetylcholine receptor subunit alpha-4” and “Neuronal acetylcholine receptor subunit alpha-7” and associated gene names.

- Diseasome, however, indicates that the corresponding genes are only important in nicotine addiction, rather than AD.

- This suggests that there is a more complex relationship between the diseases, than just sharing a drug target.

- Extending the browsing to the SWAN Knowledgebase* shows that there are hypotheses relating AD to nicotinic receptors through amyloid beta.

* http://hypothesis.alzforum.org/swan/
Conclusion and Future Work

- Extending the LOD cloud with HCLS datasets
  - Focus on clinical and pharmaceutical data sources
- Identify missing datasets and linkage points
  - By developing business use cases by pharmaceutical researchers
- Interlinking of the datasets
  - Using novel link discovery tools and frameworks including Silk and LinQuer
- Evaluating linked data exploration interfaces
The End

☐ Thank you!