Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery

Ali Hasnain et al.
Motivation

• Linked Biomedical Dataspace (LBDS) was developed to semantically interlink data from multiple sources and augment the design of in silico experiments for cancer chemoprevention drug discovery.

• LBDS provides a single-point, integrated access to multiple, diverse biomedical data sources for non-technical, domain users.

• Successful delivery of system motivated collaborators to report Important Lesson Learned.

• Questions and their applicability in real world workflows enable developers to provide the concise set of recommendations to develop LD platforms useful for drug discovery.
Linked Biomedical DataSpace
A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources

Dimitris Zeginis, Ali Hasnain, Nikolaos Loutas, Helena F. Deus, Ronan Fox, Konstantinos Tarabanis

Journal of Web Semantics 2013
Link Creation Component
Cataloguing and Linking Life Sciences LOD Cloud.

Ali Hasnain, Ronan Fox, Stefan Decker and Helena F. Deus
18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 8 - 12 October 2012), Galway, Ireland
Query Execution Component

- Query Engine
- SPARQL 1.1

Query Transformation Rules (CONSTRUCTS)

Query1 -> Query2 -> Query3 -> Query4

- Chebi
- DrugBank
- PubChem
- DailyMed

Transformed query

http://srvgal78.deri.ie:8007/graph/Granatum
Knowledge Publishing Component

ReVeaLD: REAL-TIME VISUAL EXPLORER AND AGGREGATOR OF LINKED DATA

Availability: http://srvgal78.deri.ie:8007/explorer
Knowledge Publishing Component

ReVeaLD: REAL-TIME VISUAL EXPLORER AND AGGREGATOR OF LINKED DATA

Availability: http://srvgal78.deri.ie:8007/explorer

ReVeaLD: A user-driven domain specific interactive search platform for biomedical research

Maulik R. Kamdar, Dimitris Zeginis, Ali Hasnain, Stefan Decker and Helena F. Deus

Journal of Biomedical Informatics
Questions to Consider

Q1 What is the scope of Linked Biomedical Dataspace (LBDS)?
Q2 What are the different types of relevant data sources integrated in the LBDS?
Q3 How would you confirm uninterrupted data availability from integrated sources?
Q4 How would you deal with bad quality Linked Data sources?
Q5 What should be the link types, granularity, format, size and structure of the catalogue?
Q6 What are the available linking and aligning strategies, approaches and tools?
Q7 How can the domain users intuitively search information from the LBDS?
Questions to Consider (cont…)

Q8 How could the retrieved information be presented in a human-readable, domain-specific format?
Q9 How are the limitations of the LBDS, in terms of the availability, scalability and interoperability across different platforms addressed?
Q10 What is the role of domain experts during different stages of the LBDS development?
Q11 What are the possible uses of the LBDS demonstrated in real scenarios?
Q12 Should external links to Linked Data sources be locally materialized to enhance query responses?
Q13 How would the LBDS address emerging user needs?
Workflows

Discovering and cataloguing relevant sources from LSLOD

Retrieving molecules, which interact with Estrogen receptors

Combining knowledge extracted from publications with LD
Retrieving molecules, which interact with Estrogen receptors
## Evolution and Evaluation

<table>
<thead>
<tr>
<th></th>
<th>GRANATUM</th>
<th>Open PHACTS</th>
<th>Linked2 Safety</th>
<th>DistilBio</th>
<th>Linked TCGA</th>
<th>Health-e-Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain-specific Model</td>
<td>√</td>
<td>✗</td>
<td>√</td>
<td>✗</td>
<td>√</td>
<td>✗</td>
</tr>
<tr>
<td>Knowledge &amp; Data Extraction</td>
<td>√</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Query Federation</td>
<td>√</td>
<td>✗</td>
<td>√</td>
<td>✗</td>
<td>√</td>
<td>✗</td>
</tr>
<tr>
<td>Data warehousing</td>
<td>√</td>
<td>√</td>
<td>✗</td>
<td>√</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Intuitive Querying</td>
<td>√</td>
<td>√</td>
<td>✗</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Domain-specific Visualization</td>
<td>√</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Linked Open Data</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Commercial Data</td>
<td>✗</td>
<td>√</td>
<td>✗</td>
<td>√</td>
<td>✗</td>
<td>√</td>
</tr>
</tbody>
</table>
Lesson Learned

The scope definition should include the identification of the:

- actual beneficiaries (end-users)
- potential use cases
- functional and non-functional requirements
Lesson Learned (cont…)

Relevant data sources integrated in the LBDS?

• determine the relevant sources.
Lesson Learned (cont…)

Relevant data sources integrated in the LBDS?

• determine the relevant sources.
Lesson Learned (cont…)

Relevant data sources integrated in the LBDS?

• determine the relevant sources.
Lesson Learned (cont...)

Relevant data sources integrated in the LBDS?

• determine the relevant sources.
Lesson Learned  (cont…)

Uninterrupted data availability from integrated sources?

• Specialized applications like SPARQLES could be used to recursively monitor the availability of public SPARQL Endpoints.
Lesson Learned  (cont…)

Quality of Linked Data sources?
Different namespaces used by the same provider.

<http://bio2rdf.org/kegg vocabulary:xGene>
<http://bio2rdf.org/ns/biopax#pathway>
<http://bio2rdf.org/ns/ns/bind#interactionPart>
<http://bio2rdf.org/ns/ns/ns/pubchem#MolecularFormula>

URL-encoded labels.

<http://bio2rdf.org/pdb:1%2C1%2C5%2C5tetrauorophosphopentylphosphonicAcidAdenylateEster>

non-dereferenceable URIs.

<http://bio2rdf.org/kegg vocabulary:bpm+BURPS1710b
1815+BURPS1710b A0336>

Alpha-numeric URIs, for which no labels were defined.

<http://bio2rdf.org/so:0000436>
Lesson Learned (cont…)

Link types, granularity, format, size and structure of the catalogue

• Catalogue format is an important design factor.
• If used to derive Query Transformation Rules it should be conceived to suit considered linking approaches.
• Links definition amongst Qe and Concepts (e.g. rdfs:subClassOf, owl:sameAs).
Lesson Learned (cont…)

Available linking and aligning strategies, approaches and tools

Vocabularies

Linking tools
Intuitively search information from the LBDS:
Such an interface evolves through 5 distinct stages:

1) SPARQL Query
2) Visual Query System
3) Single Entity Search (Aspirin)
4) Keyword Search
5) Google-like NL-queries.
Lesson Learned (cont…)

Information be presented in a human-readable, domain-specific format?

- RDF URIs are confusing for the biomedical researcher.
- Fresnel Vocabulary could provide more human-readable representation.
- Structural information on any entity (i.e. 3D structures, pathway maps), etc makes it easy to develop and integrate visualization libraries.
Lesson Learned  (cont…)  

Limitations of the LBDS, in terms of the availability, scalability and interoperability across different platforms

The scalability of LBDS is directly impacted by:

• Number of desirable SPARQL Endpoints to be queried.
• The size and complexity of the datasets to be RDFized, and limitations of the existing tools of KEC, and
• Visualization of a larger number of results (>10000) and computing facets for data navigation.
Lesson Learned (cont…)

LBDS address emerging user needs:
The LBDS has to provide a maintenance mechanism that satisfies emerging demands.
Lesson Learned  (cont…)  

Role of domain experts:

Collaborative decision-making is essential for:

- Model development, by identifying the scope, relevant data sources and core Qe,
- Validation of the links generated by LCC,
- Prototyping of ReVeaLD
- Evaluation of the LBDS.

However, domain experts need a stronger motivation for active participation.
Lesson Learned  (cont…)

Possible uses of the LBDS demonstrated in real scenarios:

• To strategically and informatively isolate 100 biological compounds of biological `relevance' from >300,000 compounds that can be virtually screened using in silico methods like Protein-Ligand Docking, to obtain around 10 potential compounds for in vivo analysis.
Recommendations

- **End-users (i.e. Domain experts) Involvement.**
- **Domain-specific semantic model** for the homogenisation of the data sources and the integration of the LBDS components.
- **Quality and availability** of the RDF data sources.
- SPARQL Endpoints must be monitored constantly for **availability and interoperability**.
- **Caching mechanisms** must be incorporated at the QEC.
- Data publishers must ensure the RDF **URIs are dereferenceable** across HTTP.
Recommendations (cont....)

- *User-driven tools* for data extraction and annotation must be provided.
- Retrieved information from the LBDS should be made more *human-readable and personalized*.
- Concept maps must be used for *knowledge visualization*, to enable preliminary users to interpret and formulate domain problems.
- *HCl-based evaluations of semantic web applications* must be carried out to enhance user experience and usability.
Acknowledgements
Thank You