Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery

Ali Hasnain[1], Maulik R. Kamdar[1], Panagiotis Hasapis[2], Dimitris Zeginis[3,4], Claude N. Warren, Jr[5], Helena F. Deus[6], Dimitrios Ntalaperas[2], Konstantinos Tarabanis[3,4], Muntazir Mehdi[1], and Stefan Decker[1]

[1] Insight Center for Data Analytics, National University of Ireland, Galway
{ali.hasnain,maulik.kamdar,muntazir.mehdi,stefan.decker}@insight-centre.org
[2] UBITECH Research, 429 Messogion Avenue, Athens, Greece
{phasapis,dntalaperas}@ubitech.eu
[3] Centre for Research and Technology Hellas, Thessaloniki, Greece
[4] Information Systems Lab, University of Macedonia, Thessaloniki, Greece
{zeginis,kat}@uom.gr
[5] Xenei.com claude@xenei.com
[6] Foundation Medicine Inc. Cambridge, MA hdeus@foundationmedicine.com

Abstract. The increase in the volume and heterogeneity of biomedical data sources has motivated researchers to embrace Linked Data (LD) technologies to solve the ensuing integration challenges and enhance information discovery. As an integral part of the EU GRANATUM project, a 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. The different components of the LBDS facilitate both the bioinformaticians and the biomedical researchers to publish, link, query and visually explore the heterogeneous datasets. We have extensively evaluated the usability of the entire platform. In this paper, we showcase three different workflows depicting real-world scenarios on the use of LBDS by the domain users to intuitively retrieve meaningful information from the integrated sources. We report the important lessons that we learned through the challenges encountered and our accumulated experience during the collaborative processes which would make it easier for LD practitioners to create such dataspaces in other domains. We also provide a concise set of generic recommendations to develop LD platforms useful for drug discovery.

Keywords: Linked Data, Drug Discovery, SPARQL Federation, Visualization, Biomedical Research

1 Introduction

Drug discovery entails the effective integration of data and knowledge from multiple disparate sources, the intuitive retrieval of vital information and the active involvement of domain scientists at all stages [33]. Biomedical data, encompassing a diverse range of spatial (gene ⇒ organism) and temporal (cell division ⇒ human lifespan) scales, is organized in separate datasets, each originally published to address a specific research problem. As a result, there are a large number of voluminous datasets available with varying representations, models, formats and semantics. Consequently, retrieving meaningful information for drug discovery-related queries, like *'List of molecules, with 5 Hydrogen bond donors, Molecular Weight <400 and effective against DNA Methyltransferase targets, referenced in any publications'*, becomes time-consuming and tedious as the scientist has to manually search and assemble results from several portals.

The advent of Linked Data (LD) technologies to solve the integrative challenges has opened exciting new avenues for scientific research in drug discovery [13]. These technologies not only facilitate the integration of various voluminous and heterogeneous data sources (i.e. experimental data, libraries, databases) but also provide an aggregated view of the biomedical data in a machine-readable and semantically-enriched way that enables re-use. However, domain users need to traverse a steep technical learning curve to use these technologies for addressing their research problems. Hence, the adoption of LD technologies by the actual beneficiaries of the integrated data sources is yet to be achieved.

An approach that facilitates the adoption of LD by the domain users was proposed by us, under the European FP7-funded GRANATUM project[7]. The project was conceived to semantically interlink knowledge and data for the design and execution of *in silico* experiments in the domain of cancer chemoprevention drug discovery. A Linked Biomedical Dataspace (LBDS) was developed as an integral part[8] of the GRANATUM project to offer a single-point, integrated access to multiple, diverse biomedical data sources for non-technical, domain users. We also provide a rich suite of tools to enable users publish, access and visualize their experimental datasets in conjunction with the LBDS. Our main motivation was to enable cancer researchers to retrieve information pertaining to their research questions. Previously, the domain experts have extensively evaluated the accuracy of our integration and the usability of our platform for information discovery [42,15,18]. During the development of the components we learnt important lessons by tackling the complex challenges associated with the complexity of biomedical data integration and discovery, and believe that our gained insights would be useful for LD practitioners.

The rest of this paper is structured as follows: Section 2 describes the related research carried out in this area. In Section 3, we provide a brief overview of the LBDS and its different components. In congruence with the domain experts, we outlined a set of questions (Table 1) which should be satisfactorily answered by the components. Section 4 showcases the use of different components to solve three research tasks associated with information discovery in cancer chemoprevention. Section 5 describes the evolution of our LBDS, summarizes the results of previous evaluations and compares our design decisions against some of the popular LD platforms developed for drug discovery. Finally, we report on the important lessons that we learned through the collective experience and challenges encountered, during the collaborative processes.

[7] www.granatum.org

[8] http://goo.gl/xo3KJB

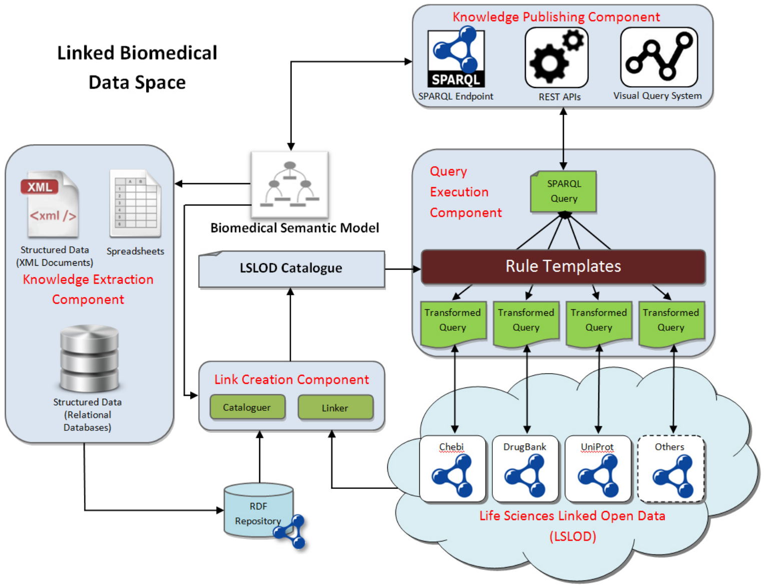


Fig. 1. Architecture of the Linked Biomedical Dataspace

2 Related Work

Initiatives, notably Bio2RDF [4] and Neurocommons [27], have been carried out for publishing biomedical resources using semantic web technologies. The Linking Open Drug Data (LODD) task force under the W3C Health Care and Life Sciences Interest Group (HCLS IG) has provided best practices and recommendations for transforming and exposing publicly available data about drugs in a LD representation [30]. Architectures like SQUIN [14] and FedX [32] could be configured for distributed querying across these data sources. Some projects have applied LD technologies for integrating and exploring biomedical data sources. OpenPHACTS [41], a pharmacological space, uses a bottom-up data-warehousing approach. DistilBio[9] was developed as a proprietary, graph-based, visual search platform for the life sciences. Health-e-child [5] employs knowledge resources and OLAP-based data normalization tools to build multi-dimensional semantic spaces from biomedical data collections. Linked2Safety [2], aims to accelerate clinical practice and medical research. Finally, Linked TCGA [29] enables evidence-based personalized prognosis for cancer.

3 Linked Biomedical Dataspace

The Linked Biomedical Dataspace (LBDS) enables the semantically-enriched representation, exposure, interconnection, querying and browsing of biomedical data and knowledge in a standardized and homogenized way. We envision the LBDS to comprise of four distinct components, namely: *i)* Knowledge Extraction

[9] http://distilbio.com/

(KEC), *ii)* Link Creation (LCC), *iii)* Query Execution (QEC), *iv)* Knowledge Publishing (KPC) (Fig. 1). A Biomedical Semantic Model is proposed as a common reference model and vocabulary for the synchronization of the four components. For LBDS, data is integrated from multiple sources including experimental datasets provided by the biomedical scientists and public repositories.

3.1 Biomedical Semantic Model

The role of the Biomedical Semantic Model is to unify the diverse and heterogeneous data sources scattered across the Life Sciences Linked Open Data (LSLOD) Cloud consistently. When the same concept (e.g. Molecule) is referred in two sources using different terms, the semantic model ensures that those terms are mapped appropriately (coreference). Furthermore, the semantic model is used for the creation of new links between entities in different data sources, the assembly of SPARQL queries and data browsing. A specific Cancer Chemoprevention semantic model (CanCO) was created for application in the cancer domain. The methodology for the CanCO development follows a "meet-in-the-middle" approach where the concepts emerged both in a bottom-up (i.e. analyze the domain) and a top-down (i.e. analyze ontologies/vocabularies) fashion [42].

3.2 Knowledge Extraction Component

Biomedical datasets are available in various formats like domain-specific CSV, XML (eXtended Markup Language) files or heterogeneous, structured databases. Representation using RDF allows standardized access and interlinking of data. The Knowledge Extraction Component (KEC) supports two main features :-

Extracting Knowledge from Dataset files : Specialized scripts were developed to transform the large datasets semi-automatically by using mapping rules, established in a simple declarative language. Any developer can easily map the structure of batch-produced XML or CSV files to concepts and properties derived from CanCO (Query Elements - Qe). Google RDF Refine[10] is also made available for the semantic enrichment of smaller files by domain users.

Extracting Knowledge from Relational Database : We have followed the D2RQ approach to expose any relational database as a virtual RDF graph and to make it available through a SPARQL endpoint [6]. The assignment of tables and columns into ontology terms, as well as the translation of SPARQL to SQL queries, is being handled via mappings expressed in the D2R language.

3.3 Link Creation Component

To assemble powerful queries traversing several SPARQL endpoints[11], it is first necessary to link the underlying data sources. A 'Cataloguer' explores and catalogues the schema used to represent data in more than 60 public LSLOD SPARQL endpoints, and a 'Linker' links the catalogued concepts and properties to CanCO Qe[12]. The Linker creates links using the following strategies: *i)* Naïve Matching/Syntactic Matching/Label Matching, *ii)* Named Entity Matching and *iii)* Manual and Domain-specific unique identifier Matching [15].

[10] `http://refine.deri.ie/`

[11] `http://srvgal78.deri.ie/RoadMapEvaluation/#Sparql_Endpoints`

[12] `http://srvgal78.deri.ie/arc/roadmap.php`

Table 1. Questionnaire

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?
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 the development of LBDS?
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?

3.4 Query Execution Component

The core component of our LBDS is a federated graph query engine, which reasons over the previously catalogued links - `{Concept_A subClassOf Qe}`, `{Concept_A void:uriRegexPattern stringPattern}` and `{sparqlEndpoint void:class Concept_A}`, to transform a simple query `{?s a Qe}` to a SPARQL construct `{{?s a Concept_A} UNION {?s a Concept_B}}` and execute the federated alternatives against the specific `sparqlEndpoint`. This ensures semantic interoperability as the formulated queries use the same semantic model and information retrieval is independent of the underlying schemas. An *ad hoc* module recursively monitors the latency of the SPARQL endpoints to 'smartly' determine which endpoints are available for querying. The query engine also provides a permission-based access to the RDFized experimental datasets.

3.5 Knowledge Publishing Component

The QEC is exposed as a SPARQL endpoint and as REST web services by the Knowledge Publishing Component (KPC). The KPC also provides a Visual Query System - ReVeaLD[13] (Real-time Visual Explorer and Aggregator of Linked Data) for facilitating non-technical biomedical users to intuitively formulate advanced SPARQL queries by interacting with a visual concept map representation of CanCO [18]. Results are aggregated from the LBDS and presented in a data browser with 'Smart Icons', which render domain-specific visualizations using a set of *Qe*-based Graphic Rules, and refer to additional information available on portals like ChemSpider [25] and PubChem [21].

4 Workflows

Our interactions with the domain experts during the development of the LBDS, allowed us to establish a set of questions (Table 1) which the components should satisfactorily address. As such, the identification of practices for addressing these is a necessary step to enable future practitioners to conceptualize dataspaces in other domains. We segregated three separate workflows where we present how the different components can be used in sequence to solve specific research problems.

[13] `http://srvgal78.deri.ie:8080/explorer`

We attempt to address the previous questions through these workflows. The users of our LBDS fall into two categories: a bioinformatician - a computer scientist with a biology background, who is responsible for data management, and a biomedical researcher who has no knowledge of computer science and uses the LBDS to query and explore the data (**Q1**). Workflow 4.1 is relevant only for the bioinformatician whereas 4.2 and 4.3 involves both users.

4.1 Discovering and cataloguing relevant sources from LSLOD

LBDS enables querying multiple, heterogeneous, distributed data sources through a single interface to address domain-specific problems. Two approaches are considered by a bioinformatician: "*a priori* integration", that uses the same vocabularies and ontologies, and "*a posteriori* integration", a methodology that defines mapping rules between different schemas, enabling the modification of the topology of queried graphs and the integration of data sources using alternative vocabularies. The steps taken for "*a posteriori* integration" are :-

1. There are multiple datasets in the LSLOD describing the concept `Molecule` - Bio2RDF KEGG `<kegg#Compound>`, DrugBank `<drugbank:Drug>`, ChEBI `<chebi#Compound>` and BioPAX `<biopax-level3.owl#SmallMolecule>` (**Q2**).
2. LSLOD SPARQL endpoints and the contained concepts and properties are catalogued. Sample instances and associated labels are also catalogued and linked to the corresponding concept using `void:exampleResource` predicate. Regular Expressions are used to identify the source of the instance (**Q5**).
3. Instances are assigned to new concepts through inference by identifying and creating a link that two concepts are similar (e.g. `owl:sameAs`, `rdfs:subClassOf`). Based on the nature of the data, the most appropriate linking process is decided using the aforementioned strategies (**Q5**,**Q6**).
4. SPARQL algebra rewrites the query at QEC to retrieve all `Molecules`.

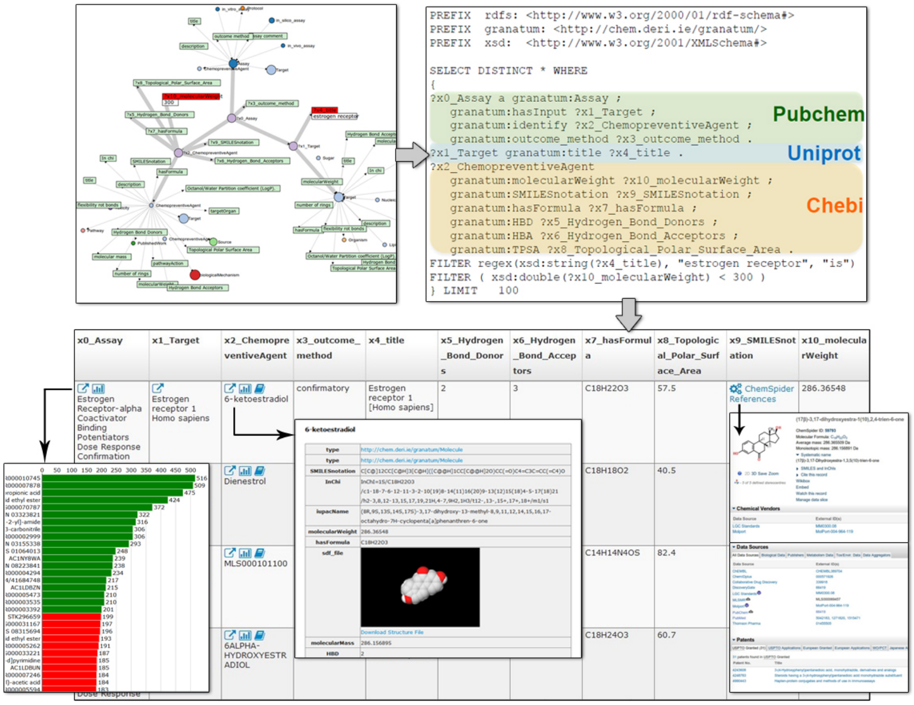
Listing 1. SPARQL Algebra to rewrite query at QEC

```
CONSTRUCT (bgp (triple?molecule a gr:Molecule)) UNION (
    SERVICE (<kegg/sparql>,<kegg/sparql>
        bgp(triple ?molecule rdf:type <kegg#Compound>))
    SERVICE (<chebi/sparql>,<chebi/sparql>
        bgp(triple ?molecule rdf:type <chebi#Compound>)))
```

4.2 Retrieving molecules, which interact with Estrogen receptors

One of the primary objectives[14] of the GRANATUM project was to identify molecules having a favorable binding affinity with Estrogen receptors-α and β for the prognosis of breast cancer drug therapy [36]. PubChem is a vast public repository cataloguing the potency of small molecules towards various biological targets, as determined by bioactivity assays (BioAssays) [21]. The central idea is to retrieve favorable agents (with `Molecular Weight`<300) targeting the Estrogen receptors from the PubChem BioAssays, and provide additional biological information of the resources (**Q1**). The steps taken were as follows :-

[14] `http://goo.gl/2OJePz`



```
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  granatum: <http://chem.deri.ie/granatum/>
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT * WHERE
{
?x0_Assay a granatum:Assay ;
   granatum:hasInput ?x1_Target ;
   granatum:identify ?x2_ChemopreventiveAgent ;
   granatum:outcome_method ?x3_outcome_method .
?x1_Target granatum:title ?x4_title .
?x2_ChemopreventiveAgent
   granatum:molecularWeight ?x10_molecularWeight ;
   granatum:SMILESnotation ?x9_SMILESnotation ;
   granatum:hasFormula ?x7_hasFormula ;
   granatum:HBD ?x5_Hydrogen_Bond_Donors ;
   granatum:HBA ?x6_Hydrogen_Bond_Acceptors ;
   granatum:TPSA ?x8_Topological_Polar_Surface_Area .
FILTER regex(xsd:string(?x4_title), "estrogen receptor", "is")
FILTER ( xsd:double(?x10_molecularWeight) < 300 )
} LIMIT   100
```

x0_Assay	x1_Target	x2_ChemopreventiveAgent	x3_outcome_method	x4_title	x5_Hydrogen_Bond_Donors	x6_Hydrogen_Bond_Acceptors	x7_hasFormula	x8_Topological_Polar_Surface_Area	x9_SMILESnotation	x10_molecularWeight
Estrogen Receptor-alpha Coactivator Binding Potentiators Dose Response Confirmation	Estrogen receptor 1 Homo sapiens	6-ketoestradiol	confirmatory	Estrogen receptor 1 [Homo sapiens]	2	3	C18H22O3	57.5	ChemSpider References	286.36548
		Dienestrol					C18H18O2	40.5		
		MLS000101100					C14H14N4OS	82.4		
Dose Response		6ALPHA-HYDROXYESTRADIOL					C18H24O3	60.7		

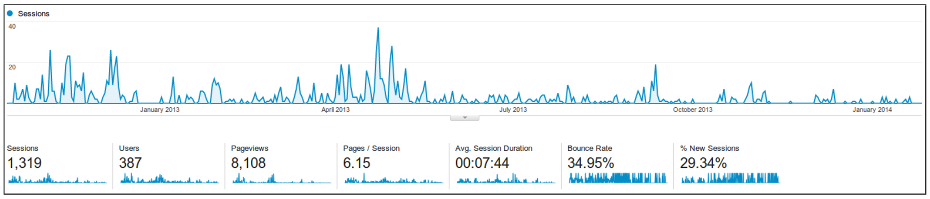
6-ketoestradiol

type	http://chem.deri.ie/granatum/Molecule
type	http://chem.deri.ie/granatum/Molecule
SMILESnotation	C[C@]12CC[C@H]3[C@H]([C@@H]1CC[C@@H]2O)CC(=O)C4=C3C=CC(=C4)O
InChi	InChI=1S/C18H22O3 /c1-18-7-6-12-11-3-2-10(19)8-14(11)16(20)9-13(12)15(18)4-5-17(18)21 /h2-3,8,12-13,15,17,19,21H,4-7,9H2,1H3/t12-,13-,15+,17+,18+/m1/s1
iupacName	(8R,9S,13S,14S,17S)-3,17-dihydroxy-13-methyl-8,9,11,12,14,15,16,17-octahydro-7H-cyclopenta[a]phenanthren-6-one
molecularWeight	286.36548
hasFormula	C18H22O3
sdf_file	Download Structure File
molecularMass	286.156895
HBD	2

Fig. 2. Using ReVeaLD to retrieve and visualize information on small molecules, identified for favorable binding activity towards Estrogen receptors

1. The bioinformatician realizes that the PubChem data source exposed as a SPARQL endpoint under Bio2RDF Release 1 experiences frequent query timeouts, making it unfeasible for integration. The datasets are downloaded through an FTP server[15] in CSV and XML formats. (**Q2**).
2. After discussing with the domain experts, the CanCO model is incremented by adding a new concept `AssayResult`, relationships `{Assay hasResult AssayResult}` and `{AssayResult mentionMolecule Molecule}`, and `AssayResult`-associated properties `outcomeMeasure` (EC_{50}, IC_{50}, $Potency$), `outcomeType` (Active or Inactive) and `outcomeValue`. (**Q10**,**Q13**)
3. The PubChem datasets are transformed using KEC and the extended CanCO model, and stored locally to ensure uninterrupted data availability. (**Q3**)
4. The advanced SPARQL query can be formulated by the biomedical researcher by clicking the concepts `Assay`, `Chemopreventive Agent (CMA)` and `Target` using ReVeaLD's concept map visualization, and setting a numerical filter (<300) on the `CMA:Molecular Weight` and a text filter (~estrogen receptor) on the `Target:title` properties (**Q7**). Additional biological properties of the CMAs could be retrieved by clicking the UI inputs.

[15] `ftp://ftp.ncbi.nih.gov/pubchem`



Sessions	Users	Pageviews	Pages / Session	Avg. Session Duration	Bounce Rate	% New Sessions
1,319	387	8,108	6.15	00:07:44	34.95%	29.34%

Fig. 3. Usage Statistics of ReVeaLD logged using Google Analytics

5. ReVeaLD's data browser replaces the RDF URIs with associated titles from the extracted dictionary. Entity information and domain-specific visualizations are accessed through 'Smart Icons' (Fig. 2) (**Q8**). Corrupt visualizations, due to deprecated structure file locations or unsupported libraries, are presented as text by default, making ReVeaLD interoperable (**Q9**).
6. ReVeaLD could transfer SMILES identifiers [39] of retrieved molecules to the ChemSpider REST API[16] to obtain information on patents and vendors, and to virtual screening platforms like LISIs [19] for *in silico* analysis (**Q11**).

4.3 Combining knowledge extracted from publications with LD

It is necessary to identify the adverse events associated with potential molecules (as discovered in assays and clinical trials) before selecting them. There is a huge wealth of knowledge stored in scientific publications, outlining the results of molecules tested previously. PubMed, an online search engine, is used by biomedical researchers globally. It comprises of citations for biomedical literature extracted from MEDLINE, Life Sciences Journals and books. Information in PubMed (publication metadata and open-access papers) is well-structured and maintained; however, the full potential of integrating this information with non-LD and LSLOD entities is yet to be realized (**Q1**). The steps taken are :-

1. The bioinformatician retrieves the XML files, regarding publication data, through PubMed Utilities (**Q2**). The KEC converts these files to RDF triples by using the *Qe* `Target`, `Molecule` and `Publication` concepts, and stores them locally to enhance query performance (**Q12**).
2. Databases of diseases and molecules, maintained by domain users, are identified and exposed as RDF Virtual Graphs using D2RQ [6]. The LCC creates links between the two aforementioned data sources. Only data sources of good granularity are selected as potential repositories to scan for links (**Q4**).
3. The QEC could perform queries upon an interlinked data sources as a single data graph. The biomedical researcher can select the `Publication` concept in ReVeaLD and request the SMILES information of the molecules, excluding those associated with adverse events harmful to human subjects (**Q7**,**Q11**).

5 Evolution and Evaluation

Since the launch of LBDS in October 2012, the bioinformaticians and biomedical researchers, associated with the project, have used the components to link newer

[16] `http://www.chemspider.com/AboutServices.aspx`

Table 2. Comparative Evaluation against Popular Linked Data Platforms

	GRANATUM	OpenPHACTS	Linked2Safety	DistilBio	Linked TCGA	Health-e-Child
Domain-specific model	✓	✗	✓	✗	✓	✗
Knowledge and Data Extraction	✓	✗	✗	✗	✗	✗
Query Federation	✓	✗	✓	✗	✓	✗
Data warehousing	✓	✓	✗	✓	✗	✓
Intuitive Querying	✓	✓	✗	✓	✓	✓
Domain-specific Visualization	✓	✗	✗	✗	✓	✓
Linked Open Data	✓	✓	✓	✓	✓	✓
Commercial Data	✗	✓	✗	✓	✗	✓

datasets and mine the LBDS (Section 3). We exposed a database, providing structural information on interesting molecules, using D2R, and converted the PubChem BioAssay XML datasets to RDF for secure access. Our discussions with the domain experts led to the inclusion of an auto-complete search input in ReVeaLD to allow single entity search, and the use of SMILES identifiers for *in silico* analysis. SPARQL endpoints under the Bio2RDF Release 2 were integrated later in May 2013 due to better uptime. To aid the non-technical users we created a screencast[17] outlining the different steps (Workflow II, III), which was made available on the project website and has been downloaded by ~175 users[18] by October 2013. Google Analytics tracking showed that ReVeaLD was accessed by 387 distinct users from 29 countries, up to February 2014 (Fig. 3), for querying and extending the CanCO semantic model. CanCO underwent 15 different changes, 9 of them were merged with the main model, whereas 6 are available as independent extensions on the GRANATUM platform.

As the LBDS evolved, we evaluated the different components separately. The expressivity, completeness, correctness, usability and simplicity of CanCO semantic model in the context of cancer chemoprevention domain was evaluated using an application-based and a human assessment methodology [42]. The links generated by LCC were evaluated both empirically and comparatively, as well as validated by the domain experts [15]. The usability and user experience of ReVeaLD was evaluated using the HCI-based 'Tracking Real-time User Experience (TRUE)' methodology [18]. Functional (`http://goo.gl/m67o03`) and non-functional (`http://goo.gl/dEZuUE`) requirements were evaluated later using questionnaires. Summarizing our results: *i)* CanCO fully covers the needs of the domain and facilitates easy usage, *ii)* existing linking strategies could not be used for LSLOD, and *iii)* a domain-specific model improves the intuitiveness of semantic search. A preliminary evaluation shows QEC to be the only federated query engine that ensures privacy and supports all SPARQL features [28].

We carried out an empirical comparative evaluation with some of the popular LD platforms, introduced in Section 2, enabling drug discovery (Table 2). In most cases, these initiatives are not yet user-driven or scalable and some approaches are too generic, whereas drug discovery is domain-specific [33]. OpenPHACTS [41], DistilBio and Health-e-child [5] platforms transform and store information from multiple providers (including commercial and private [13]) in semantic interoperable formats. Adoption by the actual users is impeded due to their

[17] `http://www.granatum.org/pub/bscw.cgi/d82084/3%20ReVeaLD.mp4`

[18] `http://goo.gl/hvKkQf`

use of a comprehensive ontology instead of a domain-specific model and they have emphasized the need for community-driven annotation and personalization. Linked2Safety [2] and Linked TCGA [29] are pursuing the domain-specific query federation approach towards data integration. However, the scalability of these platforms for integrating newer data sources is yet to be evaluated. Linked TCGA and Health-e-child also provide domain-specific visualizations [17].

6 Lessons Learned

While reviewing the state-of-the-art technologies and developing the LBDS components to address the questions (Table 1), we learned numerous lessons which may be useful for LD practitioners to develop such dataspaces in other domains.

Q1. What is the scope of Linked Biomedical Dataspace ?
The scope of the LBDS in general, and the semantic model in particular [37], should be determined initially before its conceptualization. The scope definition includes: *i)* the identification of the actual beneficiaries (end-users), *ii)* the identification of the potential use cases, and *iii)* the definition of the functional and non-functional requirements [42]. A well-defined scope will drive the whole design and development of the LBDS and facilitate subsequent decisions, like the selection of relevant resources i.e. models, ontologies, non-ontological resources, and the identification of the core *Qe*. The identification of the re-usable sections and the method of integration in the semantic model is also important [38].

Q2. What are the different types of relevant data sources integrated in the LBDS?
Due to the large number of data sources available dispersed across the web, it is crucial to determine the relevance of these sources with respect to the target domain before integration in the LBDS. The possible source types include ontologies (e.g. Gene Ontology - GO [3]), existing datasets from LSLOD (e.g. DrugBank, PubChem, PubMed), data dumps, SPARQL endpoints, user-provided data (e.g. Excel files, experimental data). A starting point of investigation could be the BioPortal [40], Bio2RDF [4] and Neurocommons [27].

Q3. How would you confirm uninterrupted data availability from integrated sources?
The latency and functionality of public SPARQL endpoints affects the quality of the retrieved query results and the domain users may not be able to get information from an important data source (e.g. PubChem). Most endpoints aggregate all the SPARQL results and push back to the client in bulk, instead of buffering them, making it difficult to determine if the endpoint has timed out or is still collecting the results. Moreover, databases like ZINC [16] are very useful for structure-based virtual screening, but are not available as SPARQL endpoints or RDF Dumps. Data warehousing approaches could be used, but the maintenance, storage and continuous updating is rate-limiting and may necessitate manual intervention [10]. Specialized applications like SPARQLES could be used to recursively monitor the availability of public SPARQL endpoints to determine query federation and make the data publishers conscious [8] .

Q4. How can one manage Linked Data sources that are of bad quality?
Curated data sources in LSLOD suffer from lack of accuracy, incompleteness,

temporal inconsistency or coverage. We found issues like: *i)* Different namespaces used by the same provider, e.g. `<http://bio2rdf.org/kegg_vocabulary:xGene>`, `<http://bio2rdf.org/ns/biopax#pathway>`, `<http://bio2rdf.org/ns/ns/bind#interactionPart>`, and `<http://bio2rdf.org/ns/ns/ns/pubchem#Molecular_Formula>`, *ii)* URL-encoded labels, e.g. `pdb:1%2C1%2C5%2C5tetrafluorophosphopentylphosphonicAcidAdenylateEster`, *iii)* non-dereferenceable URIs, e.g. `kegg_vocabulary:bpm+BURPS1710b_1815+BURPS1710b_A0336`, and *iv)* Alphanumeric URIs, for which no labels were defined, e.g. `so:0000436` [15]. Possible solutions include using partial snapshots of the endpoints (not whole RDF dumps) or mechanisms to assess the quality of LD repositories during link creation.

Q5. What should be the link types, granularity, format, size and structure of the catalogue?
As different data catalogues exist to serve distinct purposes, one should decide how well the chosen catalogue fulfills the requirements. When data linking is a key requirement it is prudent to compile a catalogue from scratch. Existing vocabularies e.g. VoID [1], DCAT[19], Dublin Core (DC)[20], and FOAF[21] can be used to describe data in the catalogue. The selection of a vocabulary depends upon the purpose of the catalogue and the granularity under consideration. For example, the PROV Namespace[22] can be used when the user wants to record the provenance information in the catalogue. The overall structure of the catalogue and its format is an important design factor. If Query Transformation Rules are to be derived from the catalogue, it should be conceived to suit considered linking approaches. *Qe* in the catalogue could be linked using link types with completely different semantics (e.g. `rdfs:subClassOf`, `owl:sameAs`).

Q6. What are the available linking and aligning strategies, approaches and tools?
Linking and aligning the semantic model with other models and ontologies plays a pivotal role in ensuring semantic interoperability and addressing data heterogeneity. However alignment of ontologies is generally suited when the data has been structured as a hierarchy which is not always the case [11]. Vocabularies e.g. WordNet [23], and Unified Medical Language System (UMLS) [7] can be used to achieve automated similarity and relatedness scores. As these vocabularies and available linking tools e.g. SILK and LIMES are very generic for LSLOD, limited success is obtained (non-specific, unrealistic and redundant links) [15].

Instance Alignment i.e. identifying the same entity referenced using different URIs, is currently very difficult to achieve at run-time and query results often contain duplicates. There is no set of common properties and unique identifiers may be encoded using different nomenclatures. For example, *Aspirin* (DrugBank), also referred as *Acetylsalicylic Acid* (ChEBI), is an interesting compound for *in silico* studies of colorectal cancer [31]. However, there is a marked difference in their InChi and SMILES representations (`smilesStringIsomeric` versus

[19] `http://www.w3.org/TR/2012/WD-vocab-dcat-20120405/`

[20] `http://dublincore.org/documents/dcmi-terms/`

[21] `http://xmlns.com/foaf/spec/`

[22] `http://www.w3.org/ns/prov#`

smilesStringCanonical). Molecular Weights and Formulas could not be used, as stereo-isomers have similar values for these attributes but are drastically different from a biological perspective (e.g. *D-Glucose* and *L-Glucose*). Approaches like [24] could be delved into deeper and tested for LSLOD.

Q7. How can the domain users intuitively search information from the LBDS?
Semantic search applications allow the formulation of highly expressive queries but SPARQL is the least usable *modus operandi* for biomedical users who may not have technical knowledge of LD Technologies. Even for a skilled LD practitioner it is difficult to assemble federated queries. An interface, which effectively lowers the barrier between Usability (*Natural*) and Expressivity (*Formal*), should be developed [20]. Such an interface evolves through 5 distinct stages - SPARQL, VQS, Single entity search, Keyword search and Google-like NL-queries. Instead of using standard ontologies a semantic model devised by the domain experts increases the intuitiveness as users are familiar with the *Qe* [42]. Concept maps augment translation of any knowledge graph, to solve a domain-specific problem, into a formal representation [22]. ReVeaLD allows visual interaction through a concept map, but still shows an extreme reliance on the CanCO *Qe*, e.g. compulsory selection of the Drug concept to retrieve information on *Aspirin* [18]. Primarily, an exhaustive dictionary summarizing all types of 'biological entities' should be compiled using machine-learning term extraction [34] and the gap could then be bridged further by proposed methodologies [12,20].

Q8. How could the retrieved information be presented in a human-readable, domain-specific format?
Although RDF representations are more suitable for semantic reasoning, RDF URIs are confusing for the biomedical researcher. Fresnel Vocabulary [26] could be used to provide a more human-readable representation. Most biomedical data sources expose REST APIs which provide structural information on any entity (i.e. 3D structures, pathway maps, etc.) and native web technologies makes it relatively easy to develop and integrate visualization libraries. ReVeaLD searches for specific triple patterns (Graphic Rules) to provide a domain-specific outlook e.g. `drugbank:targets/844 drugbank:pdbIdPage <http://www.pdb.org/pdb/explore/explore.do?structureId=1IVO>` [18]. However, many entities in the LSLOD do not have values for the predicates `rdfs:label` and `dc:title`, or the required triple patterns (`drugbank:pdbIdPage`) for the Graphic Rules.

Q9. How are the limitations of the LBDS, in terms of the availability, scalability and interoperability across different platforms addressed?
The scalability of our LBDS is directly impacted by: *i)* Number of desirable SPARQL endpoints to be queried by the QEC (current threshold is 105 endpoints), *ii)* The size and complexity of the datasets to be RDFized, and limitations of the existing tools of KEC, and *iii)* Visualization of a larger number of results (>10000) and computing facets for data navigation. A rule-based reasoning-enabled QEC for *Qe*-specific queries (i.e. DrugBank and ChEBI for Molecule) may alleviate this but the processing time would differ between the *Qe* i.e. retrieving information on Molecules is more taxing than Assays. The

reliance of ReVeaLD on the configuration of the client system (graphics card, system RAM and browser version) affects the interoperability across different platforms [18]. Some technologies, like WebGL, are only supported by modern browsers, necessitating backward compatibility. Libraries like Modernizr[23] could be used to detect which browser-based features are supported in real-time.

Q10. What is the role of domain experts during the development of LBDS? Domain experts should be actively involved throughout all stages of the development, especially during conceptualization of the semantic model, since they would be the final users. The existing methodologies for building ontologies and semantic models lack interaction with the domain experts which results in a well-construed ontology that may be not be useful for the end-users [42]. We found the collaborative decision-making between the computer scientists and domain experts essential for: *i)* Model development, by identifying the scope, relevant data sources and core *Qe*, *ii)* Validation of the links generated by LCC, *iii)* Prototyping of ReVeaLD [18] and *iv)* Evaluation of the LBDS. However, domain experts need a stronger motivation for active participation. We obtained their input and feedback through brainstorming, interviews and questionnaires.

Q11. What are the possible uses of the LBDS demonstrated in real scenarios? The main application of the LBDS would be to significantly reduce the time and costs of current drug discovery techniques. The LBDS enables domain scientists to strategically and informatively isolate ~100 biological compounds of biological 'relevance' from >300,000 compounds (Workflows II, III). These compounds can be virtually screened using *in silico* methods like Protein-Ligand Docking [35], to obtain around 10 potential compounds for *in vivo* analysis. LBDS could also be used for the discovery of biological interactions (protein-protein or gene-drug interactions) by integrating *'-omics'* datasets with GO or PubChem.

Q12. Should external links to Linked Data sources be locally materialized to enhance query responses? RDF entities existing in repositories are subject to changes, data unavailability or are badly-curated. As interfaces request data from a federated query engine, which executes queries to remote repositories, the user experience or semantic reasoning by agents is disrupted in such situations. A potential solution can be the partial materialization of RDF triples from remote resources to local repositories [9]. The query engine could first try to resolve a query locally and if it is not possible, the query can be forwarded to external repositories. The selection of triples to be cached, as well as the refresh mechanisms is subject to a lot of parameters that could be solved by weighted-equations.

Q13. How would the LBDS address emerging user needs? Even if the model seems to fully represent an area of interest (e.g. cancer chemoprevention) at the time of its creation, new needs might emerge in the future (e.g. new *Qe*) for end-users. The LBDS has to provide a maintenance mechanism that satisfies these demands. An incrementation tool was integrated with

[23] http://modernizr.com/

ReVeaLD to enable users to extend or merge the semantic model by adding new *Qe*. A naive versioning is enabled for domain users to maintain and share different modifications of their extensions.

7 Recommendations

We summarize a set of generic recommendations that initiatives developing LD platforms for drug discovery might find useful.

1. End-users (i.e. domain experts) should be involved at all stages (from conceptualization to evaluation) of the LBDS development.
2. Developers must use a domain-specific semantic model for the homogenisation of the data sources and the integration of the LBDS components.
3. Quality and availability of the RDF data sources should be taken into consideration when discovering datasets.
4. SPARQL endpoints must be monitored constantly for availability and interoperability, and feedback should be used to inform data publishers.
5. Caching mechanisms must be incorporated at the data sources and QEC.
6. Data publishers must ensure that the RDF URIs are HTTP-dereferenceable.
7. User-driven tools for data extraction and annotation must be provided.
8. Retrieved information from the LBDS should be made more human-readable and personalized to meet the needs of the domain.
9. Concept maps must be used for knowledge visualization, to enable preliminary users to interpret and formulate domain problems.
10. HCI-based (Human-computer interaction) evaluations of semantic web applications must be carried out to enhance user experience and usability.

8 Conclusion

In this paper, we present the important lessons learned during the collaborative development of a Linked Biomedical Dataspace (LBDS) for supplementing drug discovery. We provided a brief overview of the different components and the state-of-the-art technologies which could be integrated to publish, interlink, access and visualize LD. We emphasize the collaborative involvement of domain users in all the decision-making processes of the LBDS development. Three workflows showcase how the LBDS can be exploited by bioinformaticians and biomedical researchers for cancer chemoprevention drug discovery. We compare the main features of our LBDS against some of the popular LD platforms available for drug discovery. Our experiences and the challenges encountered have helped us outline the important lessons and summarize generic recommendations for LD practitioners to create such dataspaces in other domains.

Acknowledgments

This research has been supported in part by the EU FP7 GRANATUM project, ref. FP7-ICT-2009-6-270139, Science Foundation Ireland under Grant Number SFI/12/RC/2289 and SFI/08/CE/I1380 (Lion 2). The authors would like to acknowledge the members of the GRANATUM Consortium, for providing essential inputs during the development stages of the different components of the LBDS.

References

1. Alexander, K., Cyganiak, R., et al.: Describing linked datasets. In: LDOW (2009)
2. Antoniades, A., Georgousopoulos, C., Forgo, N., et al.: Linked2Safety: A secure linked data medical information space for semantically-interconnecting EHRs advancing patients' safety in medical research. In: 12th International Conference on Bioinformatics & Bioengineering (BIBE). pp. 517–522. IEEE (2012)
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., et al.: Gene Ontology: tool for the unification of biology. Nature genetics 25(1), 25–29 (2000)
4. Belleau, F., Nolin, M.A., et al.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics 41(5), 706–716 (2008)
5. Berlanga, R., et al.: Exploring and linking biomedical resources through multidimensional semantic spaces. BMC bioinformatics 13(Suppl 1), S6 (2012)
6. Bizer, C., Seaborne, A.: D2RQ-treating non-RDF databases as virtual RDF graphs. In: Proceedings of the 3rd international semantic web conference (ISWC) (2004)
7. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 32(suppl 1), D267–D270 (2004)
8. Buil-Aranda, C., et al.: SPARQL Web-Querying Infrastructure: Ready for Action? In: 12th International Semantic Web Conference, pp. 277–293. Springer (2013)
9. Castillo, R., Leser, U.: Selecting materialized views for RDF data. In: Current Trends in Web Engineering, pp. 126–137. Springer (2010)
10. Cheung, K.H., Frost, H.R., Marshall, M.S., et al.: A journey to semantic web query federation in the life sciences. BMC bioinformatics 10(Suppl 10), S10 (2009)
11. Euzenat, J., Meilicke, C., et al.: Ontology alignment evaluation initiative: six years of experience. In: Journal on data semantics XV, pp. 158–192. Springer (2011)
12. Freitas, A., Curry, E., et al.: Querying linked data using semantic relatedness: a vocabulary independent approach. Internet Computing IEEE pp. 24–33 (2012)
13. Goble, C., Gray, A.J., Harland, L., Karapetyan, K., Loizou, A., et al.: Incorporating commercial and private data into an open linked data platform for drug discovery. In: The Semantic Web–ISWC 2013, pp. 65–80. Springer (2013)
14. Hartig, O., Bizer, C., Freytag, J.C.: Executing sparql queries over the web of linked data. In: The Semantic Web - ISWC 2009, pp. 293–309. Springer (2009)
15. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences LOD Cloud. In: 1st International Workshop on Ontology Engineering in a Data-driven World at EKAW12 (2012)
16. Irwin, J.J., Shoichet, B.K.: ZINC-a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling 45(1), 177–182 (2005)
17. Kamdar, M.R., Iqbal, A., Saleem, M., Deus, H.F., Decker, S.: GenomeSnip: Fragmenting the Genomic Wheel to augment discovery in cancer research. In: Conference on Semantics in Healthcare and Life Sciences (CSHALS). ISCB (2014)
18. Kamdar, M.R., Zeginis, D., Hasnain, A., Decker, S., Deus, H.F.: ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of Biomedical Informatics 47(0), 112 – 130 (2014)
19. Kannas, C., Achilleos, K., Antoniou, Z., Nicolaou, C., Pattichis, C., et al.: A workflow system for virtual screening in cancer chemoprevention. In: 12th International Conference on Bioinformatics & Bioengineering (BIBE). pp. 439–446. IEEE (2012)
20. Kaufmann, E., Bernstein, A.: Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases. Web Semantics: Science, Services and Agents on the World Wide Web 8(4), 377–393 (2010)

21. Li, Q., Cheng, T., Wang, Y., Bryant, S.H.: PubChem as a public resource for drug discovery. Drug discovery today 15(23), 1052–1057 (2010)
22. Markham, K.M., et al.: The concept map as a research and evaluation tool: Further evidence of validity. Journal of research in science teaching 31(1), 91–101 (1994)
23. Miller, G.A., Beckwith, R., Fellbaum, C., et al.: Introduction to WordNet: An on-line lexical database. International journal of lexicography 3(4), 235–244 (1990)
24. Nikolov, A., Uren, V., Motta, E., De Roeck, A.: Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution. In: The Semantic Web, pp. 332–346. Springer (2009)
25. Pence, H.E., Williams, A.: ChemSpider: an online chemical information resource. Journal of Chemical Education 87(11), 1123–1124 (2010)
26. Pietriga, E., Bizer, C., et al.: Fresnel: A browser-independent presentation vocabulary for RDF. In: The semantic web-ISWC 2006, pp. 158–171. Springer (2006)
27. Ruttenberg, A., Rees, J.A., et al.: Life sciences on the semantic web: the Neurocommons and beyond. Briefings in bioinformatics 10(2), 193–204 (2009)
28. Saleem, M., Khan, Y., Hasnain, A., Ermilov, I., et al.: A fine-grained evaluation of SPARQL endpoint federation systems. Semantic Web Journal (2014)
29. Saleem, M., et al.: Big linked cancer data: Integrating linked TCGA and PubMed. Web Semantics: Science, Services and Agents on the World Wide Web (2014)
30. Samwald, M., Jentzsch, A., et al.: Linked open drug data for pharmaceutical research and development. Journal of cheminformatics 3(1), 19 (2011)
31. Sandler, R.S., Halabi, S., Baron, J.A., Budinger, S., Paskett, E., et al.: A randomized trial of aspirin to prevent colorectal adenomas in patients with previous colorectal cancer. New England Journal of Medicine 348(10), 883–890 (2003)
32. Schwarte, A., et al.: FedX: Optimization techniques for federated query processing on linked data. In: The Semantic Web–ISWC 2011, pp. 601–616. Springer (2011)
33. Searls, D.B.: Data integration: challenges for drug discovery. Nature reviews Drug discovery 4(1), 45–58 (2005)
34. Shi, L., Campagne, F.: Building a protein name dictionary from full text: a machine learning term extraction approach. BMC bioinformatics 6(1), 88 (2005)
35. Sousa, S.F., et al.: Protein-ligand docking: current status and future challenges. Proteins: Structure, Function, and Bioinformatics 65(1), 15–26 (2006)
36. Speirs, V., Parkes, A.T., et al.: Coexpression of Estrogen Receptor α and β Poor Prognostic factors in Human Breast Cancer? Cancer research 59(3), 525–528 (1999)
37. Uschold, M., Gruninger, M.: Ontologies: Principles, methods and applications. The knowledge engineering review 11(02), 93–136 (1996)
38. Visser, P.R., Jones, D.M., Bench-Capon, T., Shave, M.: An analysis of ontology mismatches; heterogeneity versus interoperability. In: AAAI 1997 Spring Symposium on Ontological Engineering, Stanford CA., USA. pp. 164–72 (1997)
39. Weininger, D.: SMILES, a chemical language and information system. Journal of chemical information and computer sciences 28(1), 31–36 (1988)
40. Whetzel, P.L., Noy, N.F., et al.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic acids research 39(suppl 2), W541–W545 (2011)
41. Williams, A.J., Harland, L., Groth, P., Pettifer, S., et al.: Open PHACTS: semantic interoperability for drug discovery. Drug discovery today 17(21), 1188–1198 (2012)
42. Zeginis, D., et al.: A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web (2013)