Incorporating Commercial and Private Data into an Open Linked Data Platform for Drug Discovery

Carole Goble[1], Alasdair J G Gray[1], Lee Harland[2], Karen Karapetyan[3], Antonis Loizou[4], Ivan Mikhailov[5], Yrjänä Rankka[5], Stefan Senger[6], Valery Tkachenko[3], Antony J Williams[3], and Egon L Willighagen[7]

[1] School of Computer Science, University of Manchester, UK.
[2] Connected Discovery, UK.
[3] Royal Society of Chemistry, UK.
[4] Department of Computer Science, VU University of Amsterdam, The Netherlands.
[5] OpenLink Software, UK.
[6] GlaxoSmithKline, UK.
[7] Department of Bioinformatics - BiGCaT, Maastricht University, The Netherlands.

Abstract. The Open PHACTS Discovery Platform aims to provide an integrated information space to advance pharmacological research in the area of drug discovery. Effective drug discovery requires comprehensive data coverage, i.e. integrating all available sources of pharmacology data. While many relevant data sources are available on the linked open data cloud, their content needs to be combined with that of commercial datasets and the licensing of these commercial datasets respected when providing access to the data. Additionally, pharmaceutical companies have built up their own extensive private data collections that they require to be included in their pharmacological dataspace. In this paper we discuss the challenges of incorporating private and commercial data into a linked dataspace: focusing on the modelling of these datasets and their interlinking. We also present the graph-based access control mechanism that ensures commercial and private datasets are only available to authorized users.

1 Introduction

Drug discovery requires integrating data from multiple sources about pharmacology: understanding the (malfunctioning) biological process or pathway that is causing disease, identification of the target (protein) on that pathway which can be manipulated without causing side effects, and finally identifying drugs (small chemical compounds) that interact with that target in an attempt to restore the normal biological behavior. Data on the interaction of a drug with a target is key to drug design.

Much of the pre-competitive drug discovery data is available in open public data repositories such as ChEMBL [9], ChemSpider [19], WikiPathways [15], and UniProt [22]; although some impose restrictions for commercial use of the data,

e.g. BRENDA [20] and KEGG [18]. Effective drug discovery requires comprehensive coverage of the pharmacological space, i.e. the assembly of as many datasets as possible [21]. Additionally, pharmaceutical companies have built up their own private, commercial intellectual property about compounds, targets and their interactions which they need to combine with the openly available data.

The Open PHACTS project[8] is a public-private partnership aimed at addressing the problem of public domain data integration for both academia and major pharmaceutical companies [26]. The key goal of the project is to support a variety of common tasks in drug discovery through a technology platform, the Open PHACTS Discovery Platform[9] [10] (Section 4), that will integrate pharmacological and other biomedical research data using open standards such as RDF. A key driver of the project is to address concrete pharmacological research questions and integrate with the workflows and applications already used within the drug discovery pipeline. A major requirement from the pharmaceutical companies is the ability to incorporate both commercial datasets for which they hold licenses and their own private data. Thus there is a requirement to limit access to datasets based on license restrictions and subscriptions as well as the user's credentials.

This paper presents:

- A discussion on the privacy issues around advertising the descriptions of commercial and private datasets (Section 4.1) and the deposition of chemical compounds into a registry and validation service (Section 4.2);
- The challenges of converting commercial and private datasets into linked data and combining them into a linked data platform (Section 5);
- A graph-level approach to ensure privacy of private and commercial datasets (Section 6.1), even when they are linked into the open data cloud.

2 Motivating Use Case and Requirements

The aim of any data integration system is to provide the user with a *fuller picture* of a particular dataspace than is possible by any single dataset. Such efforts are critical in pharmacology where the aim is to fully understand the effects that one or more man-made chemical molecules may have on a biological system. Such chemicals are very often designed to inhibit or activate one specific protein, yet in practice this is rarely the case. Indeed, most drugs exhibit "polypharmacology" [6] whereby they interact and perturb multiple targets in the body to different extents. The selection of a chemical for further study or commercial development is directly influenced by these profiles, assessing the risk that these unwanted effects may have on the outcome. Naturally, there have been many attempts to produce models that predict polypharmacology based on statistics generated from large pharmacology databases such as ChEMBL [9] and PubChem [25].

[8] `http://www.openphacts.org/` accessed July 2013
[9] `https://dev.openphacts.org/` accessed July 2013

Thus, access to as complete a dataset as is possible is critical at both the individual user level (exploring a particular chemical or target) and to data mining efforts, such as those building predictive pharmacology models. However, there is no one single database that holds all known pharmacology data. Moreover, the public domain systems are also complementary to commercial pharmacology databases [21] which are essential resources for many drug discovery companies. Recently, initiatives such as Pipeline Data Integrator[10] by Thompson Reuters have sought to close this gap by providing mechanisms to incorporate internal and public data along side the provider's resource. However, one might wish to integrate multiple commercial databases and/or other biological and chemical data. Therefore, in the Open PHACTS project we have undertaken a task to create a vendor-neutral, secure space whereby multiple commercial vendor datasets can sit alongside public ones; with the commercial data only accessible to authorised users, i.e. those who hold a license.

An immediate and critical question concerns whether such integration should be achieved by combining datasets within one database (i.e. data warehousing), or through web-services (i.e. federation). The Open PHACTS Discovery Platform supports both approaches. Copies of each dataset are cached into a single database in order to provide interactive responses to queries that integrate the data: the data is left in its original form. At the same time, operations such as chemical similarity search are dispatched to specialist remote web services. The nature of the queries which our users wish to perform are data intense and require searches across multiple datasets. Results from one dataset may have an affect on the data required from the other datasets in the system. For example, if the user requests the 'top ten most frequent proteins for which this chemical is active', a protein may only appear in this list given a suitable number of aggregated data points from across the resources. Thus, our approach was to design a system which would integrate commercial and public data within a dataspace. The requirements for such a platform were:

- Metadata about commercial datasets should be available to all, however private datasets should remain hidden except to those who are authorized to access the data;
- Only authorized users should be able to access the commercial and private datasets;
- Commercial and private data should be seamlessly integrated with open data.

3 Pilot Commercial Datasets

For this pilot study we obtained data from three commercial systems – GOSTAR from GVK Biosciences[11], Integrity from Thompson Reuters[12], and the AurSCOPE

[10] `http://thomsonreuters.com/pipeline-data-integrator/` accessed July 2013
[11] `https://gostardb.com/gostar/` accessed July 2013
[12] `http://integrity.thomson-pharma.com/` accessed July 2013

databases from Aureus Sciences[13] (now part of Elsevier). These datasets have already been licensed to many pharmaceutical companies for use on their internal IT infrastructure. Thus, commercial data providers already trust third parties to secure their data. Sample data were provided based on a number of specifically selected pharmacological targets in order to demonstrate the utility of combining these data and identifying the challenges, both technical and social, in including commercial and private data in an open linked data platform.

The commercial datasets differ in both their sources of data and other properties they capture. Public resources such as PubChem and ChEMBL tend to focus on published data, either from journal articles or direct from laboratories themselves. For instance, ChEMBL collates information for 1.3 million bioactive drug-like small molecules mainly extracted from over 50,000 journal articles by expert curation. In contrast, the GVK GOSTAR database includes millions of structures sourced from the patent literature as well as scientific literature. Finally, the Thompson Reuters Integrity database supplements patent and journal bioactivity with rich information on key drug discovery elements such as pharmacokinetics, company pipelines and clinical progression. Thus, a true picture of the "bioactivity space" is only available by combining all of these resources, i.e. open, commercial and private datasets.

4 Open PHACTS Discovery Platform

Building upon the Open PHACTS Discovery Platform that is detailed in [10] we discuss the incorporation of private and commercial data into an open linked data platform. The Open PHACTS Discovery Platform, depicted in Fig. 1, exposes a domain specific web service API to a variety of end user applications. The domain API co-ordinates the access to a series of services that enable the desired functionality. Briefly, the *Domain Specific Services* enable chemical structure and similarity searches as well as providing a chemical registration service (see Section 4.2); the *Identity Resolution Service* maps textual strings to concepts denoted with a URI; the *Identity Mapping Service* (IMS) supports the management of multiple URIs denoting the same concept; and the *Linked Data Cache* provides a triplestore that contains a local copy of each of the datasets. Data is cached locally for performance reasons.

The following steps have been identified for incorporating new data into the Open PHACTS Discovery Platform.

1. Define the use cases for which the data will be used, this is led by the research questions which drive the development of the Open PHACTS Discovery Platform [2].
2. Work with the data providers to generate RDF with dataset descriptions (see Sections 5.1 and 4.1).
3. Create instance level mappings from the new data source to existing data sources (see Section 5.2).

[13] `http://www.aureus-sciences.com/` accessed July 2013

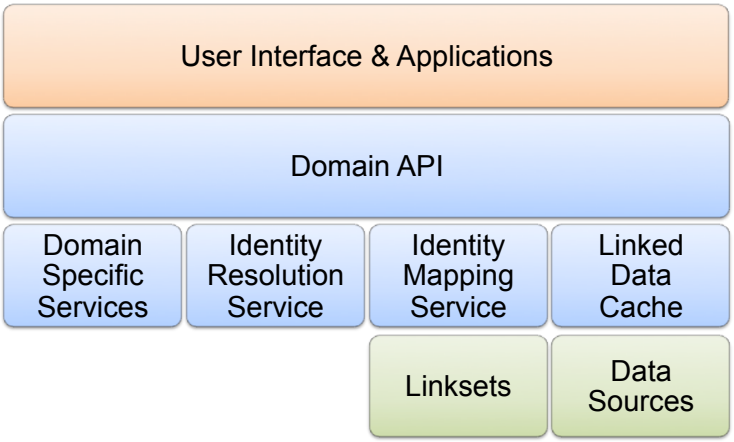

Fig. 1. Main components of the Open PHACTS Discovery Platform. Components in blue represent the core platform which exposes a domain specific API for application developers and relies on existing published data.

4. Index data for text to URI resolution.
5. Load RDF into data cache.
6. Identify data access paths required and extend or create SPARQL queries for API calls (see Section 5.3).

In the following we will discuss the privacy issues we encountered when incorporating private and commercial data into the Open PHACTS Discovery Platform. Section 5 discusses the technical challenges of modelling and linking the data.

4.1 Dataset Descriptions

Open PHACTS have specified a minimal information model for dataset descriptions [11] based on VoID [1]. The dataset description enables the dataset to be discovered, license information to be known, and for provenance of results to be returned to user requests. As per the fourth principle of [27], we believe that it is desirable that the dataset descriptions are open and accessible to all. This supports the discovery of data and can bring additional revenue to commercial dataset providers: it can be seen as advertisement for the product. However, this openness directly conflicts with the privacy requirements of the private datasets of the pharmaceutical companies.

Currently the dataset description guidelines require a substantial level of detail about the creation, sources, and release of the dataset. It also requests that statistics about the dataset, e.g. the number of concepts, are made available. With regard to commercial datasets, a balance needs to be found between the amount of information that can be exposed and the perceptions of the dataset. There are advantages to data providers in advertising the availability of a dataset in a dataset description; pointing to the provider's website allowing potential

customers to discover the existence of commercial pharmacology data without being able to access it until they subscribe. There are also potential downsides. Providers might be more wary of releasing detailed dataset statistics as this could affect the perception of their product (e.g. reporting a lower number of records than their competitor, even though these records might be of higher quality). We are in the process of revising the Open PHACTS dataset description guidelines for open and commercial datasets so that enough provenance information can be provided to the end users whilst addressing the concerns of the commercial data providers.

With regard to private datasets, it is imperative that such descriptions are not available to all: the knowledge of the existence of a dataset on a given topic is deemed as a commercial secret. However, dataset descriptions are still required in order that applications can display information correctly, the platform can decide about access to the data, and provenance about query answers can be provided to users of that data thus enabling them to verify the sources of data used to compute their query results. Therefore for private datasets we expect a minimal set of properties to be provided. These include the title of the dataset for use by applications built on top of the Open PHACTS Discovery Platform, licence used to help decide who can access the data, publisher, and issued date/version number. This metadata would only be used to respond to queries where valid credentials have been used and provides a minimal provenance trail for the data.

4.2 Chemical Registration Service

It is common for compounds in different datasets to be represented differently and this can lead to various challenges when comparing and interlinking data. To ensure data quality for the representation of chemical compounds, the Open PHACTS Discovery Platform provides a chemical registration service [14]. The chemical registration service reads a standard chemical structure information file (SD File) [8] and performs validation and standardization of the representations of the compound. The validation step checks the chemical representation for chemistry issues such as hypervalency, charge imbalance, absence of stereochemistry, etc; while the standardization step uses a series of rules, generally those associated with the US Food and Drug Administration's Substance Registration System [23], to standardize the chemical representations including the generation of charge neutral forms of the compound, non-stereo forms of the chemical, etc.

From the input SD file the chemical registration service generates an RDF representation of the data, with each distinct chemical structure being given its own identifier (URI). Various properties are computed including a unique string representing the compound (InChI) [17] together with a hash representation (InChI Key), and properties that can be derived from the canonical structure, e.g. SMILES strings and various physicochemical properties. Based on the standard InChI representation, the chemical registration service is able to collapse and aggregate the open chemical datasets used in the Open PHACTS Discov-

ery Platform, e.g. ChEBI, ChEMBL, and DrugBank; and thus generate linksets from the chemical registration data to each of these datasets.

The chemical registration service has been developed specifically to meet the privacy needs of the private and commercial data providers. It is a requirement that compounds in these datasets are not inserted into open datasets such as ChemSpider when they are deposited into the Open PHACTS Discovery Platform. Such inserts to open datasets would enable pharmaceutical companies to discover the compounds of interest of their competitors. Another consideration are the results returned to a chemical structure search. A compound is returned if and only if the molecule exists in one or more data sources to which the user has access rights. There is an interesting question of when a molecule exists in an open, commercial and private dataset, should the user know it is in the commercial and private dataset even though they do not have access rights? For private datasets, the answer is clearly no; otherwise commercially sensitive information about the dataset is passed on. However, for commercial datasets it could be seen as an advertisement for the dataset; as the user is unable to access the commercial data associated with the molecule, i.e. the value added data. Currently we are following an opt-in policy whereby the commercial providers need to choose to have their data returned to such searches. Thus, the data generated by the chemical registration service is given the same privacy level as the incoming data.

5 Converting Commercial Data to Linked Data

Converting proprietary data to linked data is quite similar to converting open data to linked data, and similar problems occur. One important aspect is interpreting the meaning of the incoming data, e.g. property names in relational schema are often not documented. However, this problem is not a consequence of the open or closed nature of the data. Instead, it is one of being able to get answers from the data providers; indeed, if the provider of open data is unwilling to provide answers, the outcome is identical.

It may be noted, however, that one should expect the context of the proprietary data may lie in data that cannot be shared. For example, the data may use internal ontologies to classify objects.

Another important aspect is that the dataset description should clearly state what users can and cannot do with the data. This may be less clearly specified with proprietary data where non standard licenses are used.

5.1 Data Modelling

The Open PHACTS project have provided guidelines [12], as a how-to guide, for the creation of five star linked data [4] for use within the Open PHACTS Discovery Platform. Here we discuss the conversion of an existing commercial or private dataset into RDF.

The original data is provided in some format: typically a database dump or an SD file for chemical data. Chemical data is passed through the chemical registration service (see Section 4.2) in order to ensure basic properties are available in the Open PHACTS Discovery Platform together with links to relate the compounds to other datasets. The data is converted into RDF and loaded into the data cache. An important aspect of modelling the data in RDF is removing any details of the underlying relational database, e.g. tables, keys, and indices. These relationships should be captured through the ontology that will be used to represent the data and the properties that it provides, i.e. they are replaced by the scientific notion they represent.

Data Structure. The three commercial datasets in this pilot study include binding data for targets with compounds. Due to the similarity with the public ChEMBL data, the triple structure used by the ChEMBL-RDF data structure [29] was used, as a *de facto* standard for encoding such data. However, compared to this approach, we here use the BioAssay Ontology for semantically annotation activities with the biological end points against which measurements were made [24]. Examples, include the IC_{50}. Various data sources use different string representations ("IC50", "IC_50", "IC-50", etc), and normalization further improves how we can mine the data.

Proprietary Ontologies. Some of the data in the commercial datasets refers to internal (implicit) ontologies. For example, the input data provided by the commercial partners includes controlled vocabularies, often including internal database identifiers. Some of these have been converted during the process into an OWL ontology. For example, such internal vocabularies have been detected in the data for the systems targeted in the experiments (which may be proteins, but also more complex biological structures) as well as pharmacological modes of action, and diseases.

However, these vocabularies are currently not further used during the integration process, and touch upon key intellectual property of the partners beyond the example data provided to us. Moreover, converting such vocabularies into more formal ontologies is a task in itself, and outside the scope of the work presented here.

Units. The activity data provided by ChEMBL mostly involves data normalized to a set of units. However, the data found in these proprietary databases do not provide normalized values. This stresses the importance of using ontologies for units, so that such normalization can be done automatically during conversion to RDF. The jQUDT library was used for this, as was used in Open PHACTS before [28], because it uses the unit conversions defined in the QUDT ontology itself, therefore effectively applying ontological reasoning.

5.2 Data Mapping

The Open PHACTS Discovery Platform requires that the Identity Mapping Service (IMS) contains information about mapping the identifiers of concepts across datasets. This is provided by a series of pairwise linksets that relate instances in the datasets. For example, the ChemSpider record for aspirin is related to the ChEMBL record for aspirin as they share the same chemical structure. When a new dataset is added, links are required to one or more existing datasets in order that the queries that power the domain API calls can return data from the new dataset when the API method is given an entry URI from another dataset.

For the interlinking of data about chemical compounds, the IMS is loaded with the linksets that are generated by the chemical registration service. The chemical registration service ensures that chemical compounds are mapped across the key Open PHACTS datasets.

With regard to biological targets (e.g. proteins), there is no equivalent service to the Open PHACTS chemical registration service. However, datasets tend to include either links to other datasets, e.g. ChEMBL and UniProt, or the enzyme commission number which can be linked directly to these datasets.

We do not concern ourselves with the issues of private and commercial data while linking the data. These are deferred to the graph-based access control used when querying the data (Section 6.1). This simplifies the mapping approach and is permissible providing that the IMS is not publicly accessible. In the Open PHACTS Discovery Platform, the IMS is only available through the domain API which is deployed on a secure web server and requires user credentials to gain access.

5.3 Querying Data

The commercial datasets considered in this pilot study are similar in their content to the ChEMBL database. As such, the RDF representation of ChEMBL-RDF was used to model the data. By adopting the same structure, the existing SPARQL queries used to respond to the domain API method calls could be used with only adding additional graph clauses to cover the commercial data. That is, each of the commercial datasets is loaded into its own named graph and these need to be addressed in the query. The benefit of loading each dataset into a graph is that we can rely on the graph-based access control of the underlying triplestore, see Section 6.1.

6 Implementation and Validation

In this section we give details of the graph-based access control employed to secure the access to the commercial and private data, and detail the validation of the approach we have applied.

```
-- Create group
DB.DBA.GRAPH_GROUP_CREATE('http://example.org/group/private');
-- Insert into group
DB.DBA.GRAPH_GROUP_INS('http://example.org/group/private',
    'http://example.org/graph/a');
DB.DBA.GRAPH_GROUP_INS('http://example.org/group/private',
    'http://example.org/graph/b');
```

Fig. 2. Creating a graph group with two members.

6.1 Graph-based Access Control

The Open PHACTS Discovery Platform uses the commercial edition of Virtuoso 7 for its triplestore which provides graph-based access control[14]. Each of the datasets used in the Open PHACTS Discovery Platform is loaded into a separate named graph. The queries that are used to respond to the domain API method calls are separated into graph blocks which control the properties that come from each of the datasets. Due to these design decisions we are able to employ the graph-based access control in Virtuoso to ensure that only authorised users are given access to commercial and private datasets.

Graph Groups. To make authorization manageable when dealing with a large number of graphs, Virtuoso introduces the concept of *graph groups*. A graph group has an IRI which represents a number of graphs. The commands for creating a graph group are given in Fig. 2.

The SPARQL query processor will "macroexpand" a graph group IRI in the dataset defined by a FROM clause into a list of its respective graphs if the executing user has permission to access the list of members of said graph group. The SPARQL query language implementation is also extended with NOT FROM and NOT FROM NAMED clauses to restrict the dataset in a query. This exclusion may also be defined through runtime parameters passed with the SPARQL query.

Authentication. Virtuoso triplestore inherits its user management from the underlying SQL database. Any query, including SPARQL queries, execute with the privileges of a valid SQL user. In case of the unauthenticated endpoint, the executing user[15] and privileges thereof are defined by the virtual directory settings for said endpoint in the internal web server. Besides the standard SPARQL protocol endpoint, one can use separate pre-defined endpoints [16] for RFC2617[17],

[14] `http://docs.openlinksw.com/virtuoso/rdfgraphsecurity.html` accessed July 2013

[15] Default user is "SPARQL"

[16] /sparql-auth, /sparql-oauth, /sparql-webid

[17] `http://www.ietf.org/rfc/rfc2617.txt` accessed July 2013

```
DB.DBA.USER_CREATE('John','VerySecretPassword');
GRANT SPARQL_SELECT TO "John";
GRANT SPARQL_UPDATE TO "John";
```

Fig. 3. Creating a user, granting read-write permissions.

OAuth[18], and WebID[19] authentication. For added security, one can use TLS for encryption. This is especially important if RFC2617 basic authentication is used, as plaintext-equivalent credentials would be passed by the client otherwise.

Customized authentication protocols can be added by declaring an authentication hook function for the internal web server's virtual directory hosting the endpoint. Authentication functions are Virtuoso stored procedures with full access to the incoming request's URL, headers and body. Hence, custom user table lookups may be performed or credentials validated through an external web service using the built-in client. Upon successful validation, the function may set the session's effective SQL user and return a value signaling the server to proceed with processing the request. The function may produce a reply (re)requesting client authentication, and cancel any further processing, should the validation fail.

Authorization. Once a user has been authenticated, there remain two levels of authorization: On the top level we have the SQL privileges mechanism – any SPARQL operations on behalf of the user require SQL privileges `SPARQL_SELECT`, and possibly `SPARQL_UPDATE` to have been granted to said user (Fig. 3). On the second level we have graph-level authorization, where a user can be granted (additional) access to individual graphs or *graph groups*.

Permissions. Graph permissions are sets of

$$\{u, g, p\},$$

where u is a valid SQL user, g is a graph or graph group IRI, and p is an integer value representing a bit vector as seen in Table 1. A simple API is provided for managing the permissions. See example in Fig. 4.

6.2 Validation

We have instantiated a test prototype of the Open PHACTS Discovery Platform to meet the needs of the pilot study to support commercial and private data. The main challenges, as reported in Sections 4 and 5, have been around modelling and interlinking the commercial data. The commercial data prototype Open PHACTS Discovery Platform correctly responds to method calls. For example,

[18] `http://www.ietf.org/rfc/rfc5849.txt` accessed July 2013
[19] `http://www.w3.org/2005/Incubator/webid/spec/` accessed July 2013

Mask	Permission
0x1	allow read access
0x2	allow write access via SPARUL
0x4	allow write access via Sponger
0x8	allow retrieval of list of members in a graph group

Table 1. Graph permission bits.

```
-- be very restrictive by default
DB.DBA.RDF_DEFAULT_USER_PERMS_SET ('nobody', 0);
-- Create user John
DB.DBA.RDF_DEFAULT_USER_PERMS_SET ('John', 0);
-- John can read this group
DB.DBA.RDF_GRAPH_USER_PERMS_SET
    ('http://example.org/group/private', 'John', 9);
-- Read-write access to own graph
DB.DBA.RDF_GRAPH_USER_PERMS_SET
    ('http://example.org/people/john','John', 7);
```

Fig. 4. Setting premissions for a graph.

in responding to a pharmacology by target method call we received additional query answers when credentials that were allowed to access the commercial data were used.

Security. The Open PHACTS Discovery Platform is accessible through standard security approaches to secure the data and provide access to it, e.g. HTTPS for web service access, API keys from 3scale[20], and graph-based access control. The graph-level security subsystem of Virtuoso is equipped with an audit procedure that checks the consistency of security rules and integrity of security-related data. These approaches have satisfied the commercial data providers involved in this pilot study.

Data. The scripts generated and used to convert the commercial data into RDF have been validated and discussed with the relevant data publisher. For each dataset, a report has been generated outlining the scripts and the rationale for their approach for generating RDF. The reports also include, in their appendices, the source code for the scripts and the generated data. The commercial data providers have been satisfied with the accuracy of the RDF data conversion.

[20] `http://www.3scale.net/` accessed July 2013

7 Related Work

There is a considerable body of work on the conversion of datasets into RDF and making them accessible as linked data [13]. Specifically within the life sciences the Banff manifesto [3] provides six rules of thumb for generating linked data and recommended best practices have been identified by the W3C Health Care and Life Sciences (HCLS[21]) interest group [16]. These community guidelines have been followed in the Bio2RDF conversion of many life sciences open datasets [5]. They are also the basis on the Open PHACTS RDF "how-to" guidelines [12] which have been used in the generation of the open and commercial datasets used in this pilot study.

Related to the generation of the data is the metadata description of the data. While VoID [1] has gained widespread use in the linked data community, there are no required properties and thus a large variation in the amount and quality of the metadata provided. Within the Open PHACTS project, we have specified a checklist of properties to provide [11]; these enable API responses to be augmented with appropriate levels of provenance information. This work has considered how these guidelines should be employed for commercial and private datasets.

In [7] the authors identify the research challenges and discuss a range of business models for linked closed data, i.e. commercial data. Cobden *et al.* focus on the sustainability of open data and a variety of business models, e.g. using advertising, to cover the hosting costs. However, this is considered on a per dataset basis. The focus of this work has been on incorporating private and commercial data into an open linked data platform to provide an integrated dataspace.

8 Conclusions

This pilot study has investigated the issues and challenges of incorporating commercial and private datasets into a linked open data platform. Samples of three commercial datasets were used to identify the challenges in converting the data and ensuring appropriate access control mechanisms. Apart from these technical issues, we also encountered social challenges around incorporating private and commercial data into an open system. These were centred around openly publishing metadata about the datasets, required for providing provenance to method calls, and registering chemical compounds in a central service. We adopted a stance whereby open and commercial dataset descriptions should be public, although possibly with different levels of granularity, while the descriptions of private datasets should remain private. With regard to the data generated by the chemical registration service, these retain the same privacy level as their source. A similar approach has been adopted for the linksets between datasets.

A key concern of the data providers is *trusting* someone else with their valuable datasets. They require strong guarantees that such data will be safe in

[21] `http://www.w3.org/blog/hcls/` accessed July 2013

the hands of a third party such as Open PHACTS. The security mechanisms employed address these concerns.

The Open PHACTS Discovery Platform[22] has been released in April 2013 and is already seeing take-up by the pharmaceutical companies as well as academic researchers. Commercial data will be included in a release in late 2013, based on both the technical and social outcomes of this pilot.

Acknowledgements. The research has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement number 115191, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007- 2013) and EFPIA companies' in kind contribution. Support was also received from the UK EPSRC myGrid platform grant (EP/G026238/1).

We would like to thank the three companies for providing us with the sample data: GOSTAR from GVK Biosciences, Integrity from Thompson Reuters, and the AurSCOPE from Aureus Sciences, now part of Elsevier.

References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with the void vocabulary. Note, W3C, `http://www.w3.org/TR/void/` (Mar 2011)
2. Azzaoui, K., Jacoby, E., Senger, S., Rodríguez, E.C., Loza, M., Zdrazil, B., Pinto, M., Williams, A.J., de la Torre, V., Mestres, J., Pastor, M., Taboureau, O., Rarey, M., Chichester, C., Pettifer, S., Blomberg, N., Harland, L., Williams-Jones, B., Ecker, G.F.: Scientific competency questions as the basis for semantically enriched open pharmacological space development. Drug Discovery Today `http://dx.doi.org/10.1016/j.drudis.2013.05.008` (To appear)
3. Banff manifesto. `http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff_Manifesto` (May 2007)
4. Berners-Lee, T.: Linked data. Technical report, W3C (2006), `http://www.w3.org/DesignIssues/LinkedData.html`
5. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2rdf release 2: Improved coverage, interoperability and provenance of life science linked data. In: ESWC. pp. 200–212. Springer, Montpellier, France (May 2013)
6. Chen, B., Wild, D., Guha, R.: Pubchem as a source of polypharmacology. Journal of Chemical Information and Modeling 49(9), 2044–55 (Sep 2009)
7. Cobden, M., Black, J., Gibbins, N., Carr, L., Shadbolt, N.: A research agenda for linked closed dataset. In: Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011). CEUR Workshop Proceedings, Bonn, Germany (2011)
8. Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A., Laufer, J.: Description of several chemical structure file formats used by computer programs developed at molecular design limited. Journal of Chemical Information and Modeling 32(3), 244 (1992)

[22] `https://dev.openphacts.org/` accessed July 2013

9. Gaulton, A., Bellis, L., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Akhtar, R., Atkinson, F., Bento, A., Al-Lazikani, B., Michalovich, D., Overington, J.: ChEMBL: A large-scale bioactivity database for chemical biology and drug discovery. Nucleic Acids Research. Database Issue. 40(D1), D1100–D1107 (2012)
10. Gray, A.J.G., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C., Burger, K., Chichester, C., Evelo, C.T., Goble, C., Harland, L., Pettifer, S., Thompson, M., Waagmeester, A., Williams, A.J.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semantic Web Journal `http://semantic-web-journal.net/sites/default/files/swj258.pdf` (To appear)
11. Gray, A.: Dataset descriptions for the open pharmacological space. Working draft, Open PHACTS, `http://www.openphacts.org/specs/datadesc/` (Oct 2012)
12. Haupt, C., Waagmeester, A., Zimmerman, M., Willighagen, E.: Guidelines for exposing data as RDF in Open PHACTS. Working draft, Open PHACTS (Aug 2012), `http://www.openphacts.org/specs/rdfguide/`
13. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, Synthesis Lectures on the Semantic Web: Theory and Technology, vol. 1. Morgan & Claypool, 1st edn. (2011)
14. Karapetyan, K., Tkachenko, V., Batchelor, C., Sharpe, D., Williams, A.J.: Rsc chemical validation and standardization platform: A potential path to quality-conscious databases. In: 245th American Chemical Society National Meeting and Exposition. New Orleans, LA, USA (Apr 2013)
15. Kelder, T., van Iersel, M., Hanspers, K., Kutmon, M., Conklin, B., Evelo, C., Pico, A.: WikiPathways: building research communities on biological pathways. Nucleic Acids Research 40(D1), D1301–D1307 (2012)
16. Marshall, M.S., Boyce, R., Deus, H.F., Zhao, J., Willighagen, E.L., Samwald, M., Pichler, E., Hajagos, J., Prud'hommeaux, E., Stephens, S.: Emerging practices for mapping and linking life sciences data using RDF - a case series. Journal of Web Semantics 14, 2–13 (2012)
17. McNaught, A.: The IUPAC international chemical identifier: InChI. Chemistry International 28(6) (2006)
18. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 27(1), 29–34 (1999)
19. Pence, H.E., Williams, A.: Chemspider: An online chemical information resource. Journal of Chemical Education 87(11), 1123–1124 (Aug 2010)
20. Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D.: Brenda, the enzyme database: updates and major new developments. Nucleic Acids Research 32(Database issue), D431–3 (2004)
21. Southan, C., Várkonyi, P., Muresan, S.: Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. Journal of Cheminformatics 1(10) (2009)
22. The UniProt Consortium: Update on activities at the universal protein resource (UniProt) in 2013. Nucleic Acids Research 41(D1), D43–D47 (2013)
23. US Food and Drug Administration: Food and Drug Administration Substance Registration System Standard Operating Procedure, 5c edn. (Jun 2007), `http://www.fda.gov/downloads/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/ucm127743.pdf`

24. Vempati, U.D., Przydzial, M.J., Chung, C., Abeyruwan, S., Mir, A., Sakurai, K., Visser, U., Lemmon, V.P., Schürer, S.C.: Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the BioAssay ontology (BAO). PLoS ONE 7(11), e49198+ (Nov 2012)
25. Wang, Y., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B., Suzek, T., Wang, J., Xiao, J., Zhang, J., Bryant, S.: An overview of the pubchem bioassay resource. Nucleic Acids Research 38(Database issue), D255–66 (Jan 2010)
26. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today 17(21-22), 1188–1198 (2012)
27. Williams, A.J., Wilbanks, J., Ekins, S.: Why open drug discovery needs four simple rules for licensing data and models. PLoS Computational Biology 8(9) (Sep 2012)
28. Willighagen, E.: Encoding units and unit types in RDF using QUDT. Working draft, Open PHACTS (Jun 2013)
29. Willighagen, E.L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A.J., Tkachenko, V., Hastings, J., Chen, B., Wild, D.J.: The ChEMBL database as linked open data. Journal of Cheminformatics 5(23) (2013)