Scientific Lenses to Support Multiple Views over Linked Chemistry Data

Colin Batchelor[1], Christian Y.A. Brenninkmeijer[2], Christine Chichester[3], Mark Davies[4], Daniela Digles[5], Ian Dunlop[2], Chris T. Evelo[6], Anna Gaulton[4], Carole Goble[2], Alasdair J.G. Gray[7], Paul Groth[8], Lee Harland[9], Karen Karapetyan[1], Antonis Loizou[8], John P. Overington[4], Steve Pettifer[2], Jon Steele[1], Robert Stevens[2], Valery Tkachenko[1], Andra Waagmeester[6], Antony Williams[1], and Egon L. Willighagen[6]

[1] Royal Society of Chemistry, UK
[2] School of Computer Science, University of Manchester, Manchester, UK
[3] Swiss Institute for Bioinformatics, Switzerland
[4] European Molecular Biology Laboratory European Bioinformatics Institute, Hinxton, UK
[5] University of Vienna, Department of Pharmaceutical Chemistry, Vienna, Austria
[6] Maastricht University, Maastricht, The Netherlands
[7] Heriot-Watt University, Edinburgh, UK
[8] VU University of Amsterdam, The Netherlands
[9] Connected Discovery, UK

Abstract. When are two entries about a small molecule in different datasets the same? If they have the same drug name, chemical structure, or some other criteria? The choice depends upon the application to which the data will be put. However, existing Linked Data approaches provide a single global view over the data with no way of varying the notion of equivalence to be applied.

In this paper, we present an approach to enable applications to choose the equivalence criteria to apply between datasets. Thus, supporting multiple dynamic views over the Linked Data. For chemical data, we show that multiple sets of links can be automatically generated according to different equivalence criteria and published with semantic descriptions capturing their context and interpretation. This approach has been applied within a large scale public-private data integration platform for drug discovery. To cater for different use cases, the platform allows the application of different *lenses* which vary the equivalence rules to be applied based on the context and interpretation of the links.

1 Introduction

Links between datasets are generally defined by the data providers using the `owl:sameAs` predicate [1]. However, Halpin *et al.* [2] have shown that `owl:sameAs` is widely misused to capture different degrees of equivalence, i.e. its practical use is not limited to the case where two resources are truly identical (implying logical equivalence) but instead capture some application scenario where the

P. Mika et al. (Eds.) ISWC 2014, Part I, LNCS 8796, pp. 98–113, 2014.

two distinct data entries can be treated as being *operationally equivalent*. This is because datasets frequently capture alternative views of the world at different levels of granularity. For example, in the case of chemical datasets used for drug discovery the focus can be on the molecular structure (e.g. ChemSpider [3]) or the drug (e.g. DrugBank [4]), which are not necessarily the same thing. This can lead to multiple ways to equate the entries in these datasets, e.g. for some applications the fact that the entries share the same drug name is enough to consider them operationally equivalent whereas for other applications a stricter criteria may be required such as having the same chemical structure or being one of the many variations possible for the compound. This means that the data can be linked in a variety of ways to satisfy different application needs depending upon the perspective of the user for a particular task.

We argue that the application using the Linked Data should decide upon the operational equivalence to apply between entries in different datasets by using a suitable *scientific lens* [5] — a set of rules that modifies the links between datasets according to some notion of operational equivalence. For the chemical example, distinct sets of links should be created for each of the different equivalence interpretations. To enable the lenses approach, the *meaning* of the link between two resources needs to be published together with the mapping. We call this the context of the link.

The work presented in this paper has been conducted as part of the Open PHACTS project [6]; a public-private partnership that has built and deployed a large scale drug discovery information space supporting several applications[1]. The Open PHACTS Discovery Platform [7] provides a domain specific Linked Data API [8] through which drug discovery data can be retrieved. The development of the platform was guided by research questions provided by drug discovery researchers both in academia and industry [9]. A requirement for the platform drawn from these research questions was to provide a mechanism through which different notions of equivalence between data sources could be supported.

This paper presents

- an approach to capturing the meaning of links which is compatible with existing published Linked Data (Section 3) and demonstrates that these can be automatically generated for chemical datasets (Section 4);
- chemical lenses that change the links between entities in different datasets based on the chemical alignments that are deemed to represent equivalent concepts under different assumptions (Section 5);
- an evaluation of the use of the chemistry lenses within the Open PHACTS Discovery Platform (Section 7).

2 Multiple Identifiers, But Are They the Same?

Scientific data is messy. It is stored in multiple datasets, each of which has been created with its own focus. For example information about drugs can be

[1] `http://www.openphactsfoundation.org/apps.html` accessed July 2014.

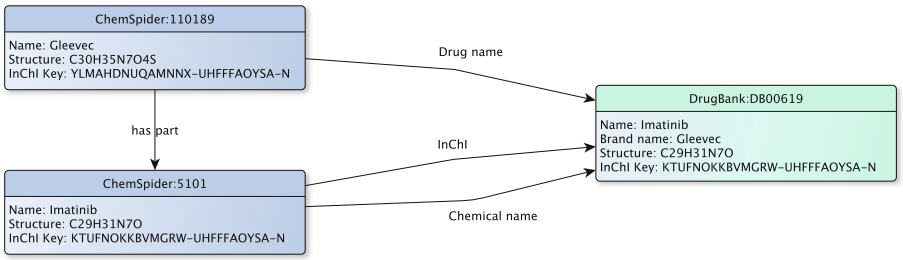

Fig. 1. Example showing the different links for relating the ChemSpider entries for imatinib and gleevec to the DrugBank record for gleevec. The equivalence encoded by each link has been provided. It also provides an equivalence relationship between the two ChemSpider records.

retrieved from DrugBank [4] while data about the chemical substances that compose the drug are available from ChEMBL [10] and ChemSpider [3]. The entries in these datasets do not align neatly, or in the ways that the scientists who need an integrated view of the data expect. The datasets use their own identifier schemes and do not always follow best practice for representing their data, e.g. representing the chemical structure with full details of charges and stereochemistry [11,12]. The challenge is identifying when two entries should be considered equivalent to meet specific scientific needs, particularly when these needs change on a per use case basis.

Consider the entries for the drug gleevec–the chemical substance imatinib mesylate– shown in Figure 1. The ChemSpider entry (`ChemSpider:110189`[2]) has the name field set to gleevec and the chemical structure for imatinib mesylate. The entry on the right is from DrugBank (`DrugBank:DB00619`[3]). It has its chemical name set to imatinib, the drug name field shows gleevec and the chemical structure is that of imatinib. Note that imatinib mesylate is a salt-form of imatinib, shown by the `has_part` relationship between the two ChemSpider records. Are the ChemSpider and DrugBank records for gleevec the "same"? For a scientist interested in the biological and medical effects of gleevec they would be, but not for a scientist interested in the physicochemical properties of imatinib mesylate.

Many datasets contain links to other related datasets. For example, UniProt [13] includes links to several related datasets. However, the nature of these links are not captured; in the case of the RDF export of UniProt they are all stated as `rdfs:seeAlso`. This is to avoid making inaccurate claims about the links, but reduces the knowledge conveyed. At the other extreme, the datasets in the Linked Data Cloud widely use the predicate `owl:sameAs` [1]; typically they do not intend the strict semantics of `owl:sameAs` [2].

For users and applications to interpret and reuse links between datasets, they need to understand what notion of equivalence is being expressed by the link.

[2] `http://www.chemspider.com/110189` accessed July 2014.
[3] `http://www.drugbank.ca/drugs/DB00619` accessed July 2014.

They need to distinguish between (i) two entries that capture different aspects of the same real-world concept, e.g. the ChemSpider and ChEMBL entries for imatinib mesylate, (ii) two entries that are highly related, e.g. the ChemSpider and DrugBank records for gleevec, and (iii) an entry that is a relevant reference but not the same real-world concept, e.g. the protein target that gleevec interacts with in the body. It is therefore hard to automatically reuse such links due to the differing natures of the datasets and meaning of the link. As such, existing links need to be used with caution in many application domains, particularly in science. To overcome this, we argue that the context of the links, i.e. the setting in which the operational equivalence between the data entries holds true, should be captured in the metadata of the link.

3 Describing Datasets and Their Links

The power of Linked Data comes from the links that relate the entities represented by the data resources. In many integration scenarios, including that of the Open PHACTS Discovery Platform, these links represent an equivalence relationship, i.e. stating that the two linked entities can be considered "the same". For example, consider the co-reference links available through the sameas.org service[4].

To enable the reuse of the links between datasets, the link consumer – a human user or an application – needs to understand what have been linked and in which context. That is, the consumer needs to know which datasets, and in particular which version of a dataset, has been linked, and what were the reasons for claiming the mapping relationship, e.g. the entities can be considered an exact match as they share the same chemical structure. (The notion of exact match is defined in SKOS [14].)

We use the approach of a VoID linkset to capture the context of the links [15]. A VoID linkset contains a collection of link triples that relate the entries in a pair of datasets through a single mapping relationship. The linked datasets are themselves described using VoID. For the purposes of the Open PHACTS Discovery Platform, we have defined a checklist of properties that must be provided, e.g. the license and version number, and those that are optional to provide, e.g. the location of a SPARQL endpoint containing the data [16].

As shown by the example in Figure 1, there can be many reasons to equate entries across datasets. The VoID linkset metadata captures details of the datasets linked, i.e. the context, and the link relationship. However, the link predicate tends to be a generic mapping relationship such as `owl:sameAs` or one from SKOS which does not convey the *reason why* the entries are equivalent, i.e. the *justification* for the equivalence relationship.

One approach to capture the equivalence relationship conveyed by a link between two data entries is to define a domain specific predicate. For example, one could define a mapping predicate that states that two linked chemical entries are considered operationally equivalent because they have the same drug

[4] `http://sameas.org/` accessed July 2014.

name. This new mapping predicate could be declared as a sub-property of the `skos:exactMatch` predicate. This would allow standard inferencing rules to be applied. However there is a major social barrier to such an approach – gaining consensus on the required linking predicates. Additionally, there is the burden of updating the existing links in the datasets to use these new link predicates; a human intensive task. Such an approach is unlikely to gain traction.

An alternative is to continue using existing link predicates such as `owl:sameAs` and those in SKOS, and annotate the linkset descriptions with additional contextual data that captures the equivalence criteria used to generate the links. This enables the use of the existing links unchanged. Thus, lowering the barrier to uptake as the annotations can be retrofitted to the existing links.

We term this additional metadata the *justification* for the linkset; the notion captured is the scientific interpretation of the operational equivalence applied by the linkset. For example, the linkset relating ChemSpider and DrugBank because they have the same InChI representation of the chemical[5] would express the justification in the linkset VoID header with the triples

```
:Chemspider-Drugbank_Linkset void:linkPredicate skos:exactMatch ;
             bdb:linksetJustification cheminf:CHEMINF_000059 .
:cheminf:CHEMINF_000059 rdfs:label "InChIKey" .
```

where `:Chemspider-Drugbank_Linkset` is the resource that describes the linkset, the link predicate is declared to be `skos:exactMatch` using the VoID predicate, and the justification is specified using the BridgeDb vocabulary (namespace `bdb`[6]) with the value for InChI Key taken from the Chemical Information Ontology (namespace `cheminf`[7]). The set of supported justifications within the Open PHACTS Discovery Platform can be found in [16]; the subset relating to chemistry data are included in Tables 1 and 2. A key advantage of this approach is that it extends rather than changes the existing data, i.e. the metadata can be added later on with minimal effort.

4 Linked Chemistry Data

There are a large number of datasets (openly) available that contain information about chemicals. However, differences in scientific or technical approaches to molecular structure representation mean that data sources will not always be in agreement in the chemical structure for a given substance. Various efforts are ongoing to link entries for the same chemical between databases, for example, to link metabolites [18,19].

[5] InChI is a standardised string representation for chemical compounds, the hash value of which is called the InChI Key [17].

[6] `http://vocabularies.bridgedb.org/ops` to appear soon.

[7] `http://semanticscience.org/resource/` accessed May 2014.

Table 1. Predicates used to capture the justification of chemical linksets and the operational equivalence that is interpreted. `sio`: Semantic Science Integrated Ontology, `cheminf`: Chemical Information Ontology, and `chebi`: Chemical Entities of Biological Interest Ontology.

Term	Justification
Chemical entity `sio:SIO_010004`	The concepts linked represent the same chemical entity.
InChI Key `cheminf:CHEMINF_000059`	The concepts linked have the same InChI Key.
Has part `chebi:has_part`	Used to indicate the relationship between part and whole.
Is tautomer of `chebi:is_tautomer_of`	Used to denote that the related chemical entities are tautomers.

4.1 Chemistry Registration Service

It is common for compounds in separate datasets to be represented differently and this can lead to various challenges when comparing and interlinking chemical data. To ensure data quality for the representation of chemical compounds, the Open PHACTS Discovery Platform provides a Chemical Registration Service [20], which reads a standard chemical structure information file (SD File) [21] 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. The standardization step uses a series of rules, based on those of the US Food and Drug Administration's Substance Registration System [22], to standardize the chemical representations.

The Chemical Registration Service identifies the *chemical counterparts* of each molecule—these are representations of the substance stripped of their stereo bonds, salts and charge. These counterparts provide a resource for relating representations across datasets. Previously, ChEBI [23] had the richest set of relationships between molecular structures, including parthood relations, relations between enantiomers (opposite stereo forms) and relations between tautomers (rapidly interconverting forms of a molecule such as the ring and chain forms of glucose) [24], and of course the subclass relation relating a more-completely specified structure to a less-completely specified structure (in terms of, for example, stereochemistry or isotopic composition). However, ChEBI does not distinguish between different subclass relations, indicate which of the forms of a tautomer are in the majority under physiological conditions, or indeed relate structures to structures that have been normalized according to the Open PHACTS rules. Thus we have for the moment, after discussion with ChEBI about adding more relationships, extended CHEMINF with the concepts and relationships given in Table 2. The ChEBI team will consider them for future inclusion in their

Table 2. Additional predicates for representing chemical equivalences. The `cheminf` namespace refers to `http://semanticscience.org/resource/`.

Term	Description
has uncharged counterpart `cheminf:CHEMINF_000460`	Connects a molecule to molecule with identical heavy-atom connectivity which is neutral. It is not a subclass relation.
has component with uncharged counterpart `cheminf:CHEMINF_000480`	Connects a molecular substance, say a mixture containing ions, with a neutral form of one of the ions.
has stereoundefined parent `cheminf:CHEMINF_000456`	Subclass relation between a class that has stereochemistry defined and an otherwise identical class that does not.
has isotopically unspecified parent `cheminf:CHEMINF_000459`	Subclass relation between a class that has isotopes specified, for example D2O or 14C-urea, and an otherwise identical class that does not, for example water or urea.
has major tautomer at pH 7.4 `cheminf:CHEMINF_000486`	*A* exists in an equilibrium with *B* at pH 7.4 and physiological temperature and *B* is the dominant isomer.
has OPS normalized counterpart `cheminf:CHEMINF_000458`	This connects a molecule to its normalized counterpart according to the OPS specification.

ontology [25]. These predicates can be used in addition to those in Table 1 as the justification property of the linkset descriptions to capture the equivalence condition applied. For example, the relationship between the two ChemSpider records in Figure 1 would use the justification `chebi:has_part`.

4.2 Generating Linked Chemistry Data

From the input SD file the Chemical Registration Service generates an RDF representation of the data, with each distinct chemical structure having its own Open PHACTS identifier (URI). Various properties are computed including its InChI representation [26] and properties that can be derived from the canonical structure, e.g. SMILES strings and various physicochemical properties such as molecular weight. Based on the InChI representation, the Chemical Registration Service is able to collapse and aggregate the source dataset representations, and thus generate linksets from the Chemical Registration Service data to each of these datasets, e.g. ChEBI, ChEMBL and DrugBank. Note that mol V2000 is used for the internal representation for chemicals [27].

The generation of linksets has been implemented as part of the data processing pipeline of the Chemical Registration Service. Each dataset and linkset has a metadata description conforming to the specification in [16]. The metadata

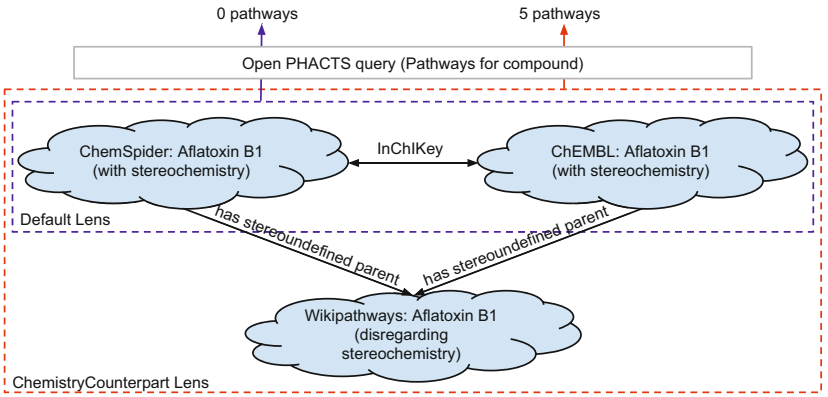

Fig. 2. A graphical depiction of the Aflatoxin B1 example. The blue dashed box encompasses the linksets activated under the Default lens while the red dashed box encompasses the additional linksets activated under the ChemistryCounterpart lens. The top of the figure states the number of pathways discovered under each lens when querying for Aflatoxin B1 with details of its stereochemistry.

description captures the *context* of the linkset, i.e. which specific version of a dataset has been loaded into the Chemical Registration Service on which date, as well as the *justification* for the links, i.e. the equivalence criterion used to generate the linkset.

5 Chemistry Lenses

Users of data integration systems such as the Open PHACTS Discovery Platform expect answers to their queries despite discrepancies in the underlying data that make aligning the data difficult. For example, consider the Wikipathways [28] entry WP699[8] representing the cellular pathway for the human metabolism of aflatoxin B1. When mappings are based on entries sharing the same InChI, the pathway is not returned when searching for pathways containing the compound aflatoxin B1 as represented by the ChemSpider entry `ChemSpider:162470`[9], represented by the blue box in Figure 2. This is due to one, or more, of the underlying data sources not containing details of the stereochemistry – it is common for datasets to not include details of the stereochemistry as it is simply unknown in many cases. However, the users expect that the pathway would be returned for the call since they loosen their notion of equivalence to include stereoisomers, corresponding to the red box in Figure 2. We propose the use of lenses to enable such functionality, i.e. to vary the equivalence criteria applied for a given query by applying a different lens.

[8] `http://www.wikipathways.org/instance/WP699` accessed May 2014.
[9] `http://www.chemspider.com/162470` accessed July 2014.

A *lens* defines a conceptual view over the data that varies the links between datasets based on the operational equivalence to be applied. Lenses are modelled in RDF and consist of the following:

- Identifier: Each lens is given a URI to identify it.
- Title (`dct:title`): Each lens is given a short descriptive title.
- Description (`dct:description`): Each lens has a textual description that explains the effect of the lens to a domain scientist.
- Documentation link (`dcat:landingPage`): A link to further explanation with illustrative examples of the effects of the lens.
- Creator (`pav:createdBy`): A link to a resource that represents the person that created the lens.
- Creation date (`pav:createdOn`): Timestamp of when the lens was created.
- Equivalence rules (`bdb:linksetJustification`): A set of URIs identifying the justifications that hold under the lens.

At present, we capture minimal provenance information (creator and creation date), using properties from the PAV ontology [29]. We have found it necessary to provide detailed documentation of the effects of each of the lenses deployed on the Open PHACTS Discovery Platform. This documentation demonstrates the effects of the lens using examples to show the changes in the results returned.

Within the Open PHACTS consortium, we are testing two lenses. The first encapsulates a set of default expected behaviours. This lens equates chemicals that have the same InChI representation or where the datasets equate their identifiers. This lens provides the primary linking between chemical compounds and matches the behaviour of existing integration strategies and in particular that of the Open PHACTS Discovery Platform prior to the introduction of lenses.

The second lens, called the ChemistryCounterpart lens exploits the full set of relationships generated by the Chemical Registration Service, i.e. the justifications captured in Tables 1 and 2. It is very permissive in its notion of equivalence, relating all entries that are variations of charge, isotopes, stereochemistry, salt forms, tautomers, and compounds in a mixture.

Additional lenses that only activate one of these variations, e.g. a stereochemistry lens, could easily be added from a technical perspective – the infrastructure and data exist to provide the lens. However, considerable effort is required to explain the behaviour of a given lens to the scientific users of the system.

For the cellular pathway example, the users of the Open PHACTS API benefit from the use of lenses. Under the default lens, no pathways are returned due to the datasets containing different stereoisomers. Using the ChemistryCounterpart lens, five pathways are returned including the Wikipathways one.

6 Identity Mapping with Lenses

The lenses functionality is provided within the Open PHACTS Discovery Platform by the Identity Mapping Service (IMS). The IMS provides a lookup service to return "equivalent" URIs for a given URI. The notion of equivalence can be

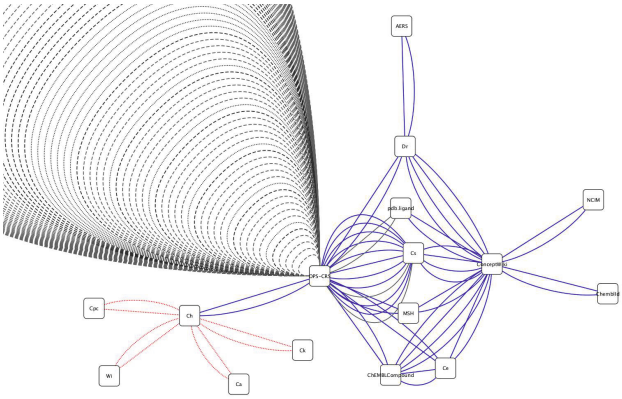

Fig. 3. Visualisation showing the interlinking of the 16 chemistry datasets. Blue edges depict InChI equivalences, red edges depict the same chemical entity equivalences, and grey depict the ChEBI and CHEMINF equivalences from Tables 1 and 2. Solid lines are `skos:exactMatch` links, dashed lines are `skos:closeMatch` links and dotted lines are `skos:relatedMatch` links.

varied by supplying the URI for the lens to be applied. The IMS implementation is an extended version of the BridgeDb framework that maps database identifiers [30]. The IMS implementation supports cross-references over Linked Data sources, i.e. supporting the use of URIs to represent entries in datasets and loading mapping data from VoID linksets. The source code is available from `https://github.com/openphacts/IdentityMappingService` and the service is accessible through the Open PHACTS API, `https://dev.openphacts.org/`.

6.1 Interconnected Data

The linked chemistry data consists of 130 linksets containing 13,970,556 links that connect the Chemical Registration Service to each of its source datasets, generating a hub of data shown in Figure 3. To answer the queries behind the Open PHACTS API methods we require links directly between the various source datasets[10]. These can be computed by the IMS using custom inference chains. However, this process needs to consider the justifications associated with the linksets.

Based on the justification of linksets, we can compute inferred linksets. For example, we can generate a linkset between datasets A and C through some intermediary dataset B if there is a linkset between A and B and one between B and C such that both linksets have the same justification. Definition 1 formally gives the rule for computing inferred linksets based on their justification. We denote a linkset between datasets A and B with the link predicate p and justification j as $A \xrightarrow[p]{j} B$. Note that we do not require that the linksets have

[10] In order that these queries run efficiently, the links are materialised in the IMS.

the same link predicate when inferring linksets. The resulting inferred linkset is given the weaker of the two link predicates with a hierarchy of

$$\texttt{owl:sameAs} \preceq \texttt{skos:exactMatch} \preceq \texttt{skos:closeMatch} \preceq \texttt{rdfs:seeAlso}.$$

Thus, if p was the link predicate `owl:sameAs` and r the link predicate `rdfs:seeAlso`, the computed linkset $A \xrightarrow[r]{j} C$ would have the link predicate `rdfs:seeAlso`.

Definition 1 (Inferring linksets based on justifications). *Given datasets A, B, and C, linksets $A \xrightarrow[p]{j} B$ and $B \xrightarrow[r]{j} C$ both with the justification j and link predicates p and r respectively then we can generate the linkset*

- $A \xrightarrow[r]{j} C$ *if* $p \preceq r$;
- $A \xrightarrow[p]{j} C$ *if* $r \prec p$.

By iteratively applying the rule given in Definition 1 it is possible to compute chains of linksets that use the same justification. This can be seen as materialising the network of 'follow-your-nose' links in the data for a given equivalence type. It is possible to enter an infinite cycle while computing these links; thus the IMS implementation prevents a dataset being revisited in a chain. As part of the provenance of the computed linkset, the linksets that are used to compute it are tracked and reported in the resulting VoID description of the linkset.

By inferring the network of links over the Open PHACTS datasets, the deployed IMS contains 51,168,586 links from 40,802 linksets. Note that the link materialisation is independent of the lenses applied. The materialisation mechanism computes every possible inferred linkset based on the justification and link predicate.

6.2 Lens Implementation

The IMS responds to a request for equivalent URIs by performing a lookup in its internal database. Since the network of interlinks is pre-computed, the implementation of lenses is straightforward. The API call is extended with a new parameter to pass in the lens URI. This URI is used to retrieve the justifications that are enabled by the lens. The equivalent URI lookup query has additional conditions added which ensure that only links with enabled justifications are returned.

A lookup for data through the Open PHACTS Discovery Platform must provide the URI of the entity of interest. However, the user does not need to know the equivalent URIs in all of the datasets used by the Discovery Platform. This is handled by the IMS which is called by the workflow that fulfils the API call. We have previously shown that this adds a small overhead to the execution time of a method call, but that a user will not perceive this [31]. We believe the advantage of enabling the user to select their operational equivalence conditions outweighs this small performance hit.

7 Evaluation

The effects of the chemistry lenses on the answers returned by the Open PHACTS Discovery Platform were analysed by two pharmacology researchers. The researchers used a set of 22 chemicals which were chosen for the different chemical features they exhibit[11], viz. stereochemistry (15), tautomers (10), isotopes (3), charge (2), salts (3). One compound acted as a control as it did not contain any of the above features. Note that a single chemical may exhibit multiple features unless it is in the control group. This resulted in an extensive number of relationships that were systematically compared to verify their correctness by the pharmacology researchers. (A very labourious task.)

For each chemical, identified using its ChemSpider URI[12], the evaluator executed the `mapUri` API call[13], which returns the set of equivalent URIs for the given seed value under the supplied lens. The calls were made first using the Default lens, which matches the behaviour of earlier releases of the Open PHACTS Discovery Platform, and then with the ChemistryCounterpart lens.

The results of each call were analysed. First the images of the chemicals returned by the call were visually inspected against the associated image of the seed chemical. This visual inspection was used to determine that the returned substances were related to the seed substance, e.g. as a charge neutral parent chemical. Next, the result set was inspected to ensure that each of the relevant parent chemicals were returned when the ChemistryCounterpart lens was invoked, i.e. if a chemical exhibits stereochemistry and is a salt we would expect that the stereo parent as well as the salt base and the base chemical would be returned. The lenses were found to work as expected.

The linkset data and the lens enabled IMS have been deployed in the Open PHACTS Discovery Platform. The Linked Data API of the platform has been extended to enable the lens parameter to be passed in; if no URI is supplied then the Default lens is applied. The Open PHACTS Discovery Platform receives over 2 million hits a month providing further assurance of the correctness of the lenses and the underlying linksets.

8 Related Work

Data integration has been widely studied both in the relational database and the semantic web communities [32]. Integration systems expose a single view of the world to users and require the work of a domain expert to interrelate the datasets to be integrated. Dataspaces [33] aim to lower the up-front cost by starting with rough relationships that can be refined automatically through

[11] Values in brackets indicates the number of chemicals that exhibit that property.

[12] ChemSpider chemicals are indexed using the InChI code set to the standard settings with the exception of the reconnected layer, so distinguish the various forms that a substance can take.

[13] `https://beta.openphacts.org/1.3/mapUri` accessed July 2014. To create a free API access key and read documentation see `https://dev.openphacts.org/`

user feedback. The Open PHACTS Discovery Platform takes a similar approach; integration is achieved through queries and the relationships between datasets are captured in our global-as-view queries. However, we enable multiple views over the data by varying the active equivalence relationships for the instance URIs through the use of lenses.

Lenses rely on the availability of multiple links between datasets which provide different equivalence relationships. Several tools have been developed for generating links between datasets [34]. Since 2009 there has been an instance matching track[14] in the annual ontology matching competition[15] to compare such tools. The most recent results are available from [35]. These are general purpose link generators that look for similarities between resources in two datasets. In general, they generate one set of links based on the matching algorithms applied and a threshold value for closeness. The Chemical Registration Service exploits domain knowledge, viz. properties of the chemicals, to generate multiple linksets, each based on different equivalence criteria. Other efforts are ongoing to link entries for the same chemical between databases, for example, to link metabolites [18,19,36], but these are focused on linking database entries and do not consider the need to support multiple linkages for different use cases. We are investigating similar approaches for proteins and other entities of interest.

There are two approaches in the literature for managing the multiple URI problem. The first approach recognises that the same logical resource can be given multiple URIs, e.g. when a dataset is served by multiple mirrors, or that some entities may be unambiguously identified. Services such as the Identifiers.org [37] which addresses the multiple data mirrors problem and the Entity Name System [38] which addresses the disambiguation problem provide a URI for the concept that can be used unambiguously. However, this is not the problem addressed by the lenses proposed in this paper. The second approach are co-reference services that provide links between entities in different datasets. This is the problem addressed by the lenses. Co-reference services such as sameas.org[16] [39] provide a service by which equivalent URIs can be obtained. sameas.org harvests `owl:sameAs` links from publicly available datasets on a wide range of topics. These existing co-reference services do not consider under what conditions the equivalence holds. The data loaded into the IMS is curated and comes with a justification for the equivalence. We believe that these third party co-reference services are an underutilized but key part of developing practical semantic web applications.

9 Conclusions

In this paper, we have shown the importance of understanding the nature of how links between datasets are created in order to effectively answer scientific

[14] `http://www.instancematching.org/oaei/imei2013/results.html` accessed May 2014.

[15] `http://oaei.ontologymatching.org/` accessed May 2014.

[16] `http://www.sameas.org` accessed May 2014.

questions. We describe a process for generating such domain specific links and techniques for applying them. Our approach is deployed on a live system that has been used as the basis for a variety of drug discovery applications[17]. Moreover, expert users have verified the validity of the results of our system. Lenses have practical benefits in allowing users to vary how the data is exposed under integration.

While the technical implementation of lenses is relatively straightforward and indeed the overall concept of a lens is easy to grasp, the effects of applying a lens requires considerable training and educating of the users. To this end, we are endeavouring to supply suitable user-oriented documentation for each lens deployed in the Open PHACTS Discovery Platform.

Given the broad capabilities of the scientific lenses approach, we are still discussing which set of lenses will be included in future versions of the platform. The division of the chemical features between the Default lens and other lenses remains to be decided. There is some interest in including the tautomers in the default and dividing the other chemical features (stereochemistry, salt forms, *etc.*) into their own specific lenses rather than one lens which contains all features. This may simplify the results returned, but increases the choice presented to applications and users.

Finally, we are looking at expanding our lenses approach to the other types of datasets needed for drug discovery, viz. proteins, splice variants, cross-species relationships. We are also looking at how lenses can be used to vary the quality associated with the links, e.g. curated versus non-curated links.

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.

References

1. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool (2011)
2. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 305–320. Springer, Heidelberg (2010)
3. Pence, H.E., Williams, A.J.: ChemSpider: an online chemical information resource. Journal of Chemical Education 87(11), 10–11 (2010)
4. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C., Wishart, D.S.: DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Research 39(Database issue), D1035–D1041 (2011)

[17] `http://www.openphactsfoundation.org/apps.html` accessed July 2014.

5. Brenninkmeijer, C.Y.A., Evelo, C., Goble, C., Gray, A.J.G., Groth, P., Pettifer, S., Stevens, R., Williams, A.J., Willighagen, E.L.: Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision. In: Proc. Linked Science, Boston, MA, USA. CEUR-WS.org (2012)
6. 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)
7. Gray, A.J.G., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C.Y.A., Burger, K., Chichester, C., Evelo, C.T., Goble, C.A., Harland, L., Pettifer, S., Thompson, M., Waagmeester, A., Williams, A.J.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semantic Web 5(2), 101–113 (2014)
8. Groth, P., Loizou, A., Gray, A.J.G., Goble, C., Harland, L., Pettifer, S.: API-centric Linked Data Integration: The Open PHACTS Discovery Platform Case Study. Journal of Web Semantics (2014)
9. 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 18(17-18), 843–852 (2013)
10. Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., Overington, J.P.: The ChEMBL bioactivity database: an update. Nucleic Acids Research 42(Database issue), D1083–D1090 (2014)
11. Williams, A.J., Ekins, S.: A quality alert and call for improved curation of public chemistry databases. Drug Discovery Today 16(17-18), 747–750 (2011)
12. Williams, A.J., Ekins, S., Tkachenko, V.: Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discovery Today 17(13-14), 685–701 (2012)
13. The UniProt Consortium: Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Research 41(Database issue), D43–D47 (2013)
14. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. Recommendation, W3C (2009), `http://www.w3.org/TR/skos-reference`
15. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary. Note, W3C (2011)
16. Gray, A.J.G.: Dataset descriptions for the Open Pharmacological Space. Working draft, Open PHACTS (2013)
17. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., Pletnev, I.: InChI-the worldwide chemical structure identifier standard. J. of Cheminformatics 5(1), 1–9 (2013)
18. Wohlgemuth, G., Haldiya, P.K., Willighagen, E., Kind, T., Fiehn, O.: The chemical translation service a web-based tool to improve standardization of metabolomic reports. Bioinformatics 26(20), 2647 (2010)
19. Haraldsdóttir, H.S., Thiele, I., Fleming, R.M.: Comparative evaluation of open source software for mapping between metabolite identifiers in metabolic network reconstructions: application to recon 2. Journal of Cheminformatics 6(1), 2 (2014)
20. 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 (2013)

21. 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. US Food and Drug Administration: Food and Drug Administration Substance Registration System Standard Operating Procedure. 5c edn. (2007), http://www.fda.gov/downloads/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/ucm127743.pdf
23. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research 36, D344–D350 (2008)
24. Sayle, R.A.: So you think you understand tautomerism? Journal of Computer-Aided Molecular Design 24, 485–496 (2010)
25. Hastings, J.: Personal communication
26. McNaught, A.: The IUPAC international chemical identifier: InChI. Chemistry International 28(6) (2006)
27. 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 Computer Sciences 32(3), 244–255 (1992)
28. Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R., Evelo, C.: WikiPathways: pathway editing for the people. PLoS Biol. 6(7), e184 (2008)
29. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J.G., Goble, C., Clark, T.: PAV ontology: Provenance, Authoring and Versioning. Journal of Biomedical Semantics 4(37) (2013)
30. van Iersel, M.P., Pico, A.R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B.R., Evelo, C.T.: The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics 11(5) (2010)
31. Brenninkmeijer, C.Y.A., Goble, C., Gray, A.J.G., Groth, P., Loizou, A., Pettifer, S.: Including Co-referent URIs in a SPARQL Query. In: 4th International Workshop on Consuming Linked Data, Sydney, Australia (2013)
32. Doan, A., Halevy, A., Ives, Z.: Principles of Data Integration. Elsevier (2012)
33. Halevy, A.Y., Franklin, M.J., Maier, D.: Principles of dataspace systems. In: PODS 2006, Chicago (IL, USA), pp. 1–9. ACM (2006)
34. Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEEE Trans. Knowl. Data Eng. 25(1), 158–176 (2013)
35. Cuenca Grau, B., Dragisic, Z., Eckert, K., Euzenat, J., Ferrara, A., Granada, R., Ivanova, V., Jiménez-Ruiz, E., Kempf, A.O., Lambrix, P., Nikolov, A., Paulheim, H., Ritze, D., Scharffe, F., Shvaiko, P., Trojahn, C., Zamazal, O.: Results of the Ontology Alignment Evaluation Initiative 2013. In: Ontology Matching (2013)
36. Galgonek, J., Vondrasek, J.: On InChI and evaluating the quality of cross-reference links. Journal of Cheminformatics 6(1), 15+ (2014)
37. Juty, N., Le Novère, N., Laibe, C.: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Research 40(Database issue), D580–D586 (2012)
38. Bouquet, P., Stoermer, H., Bazzanella, B.: An Entity Name System (ENS) for the Semantic Web. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 258–272. Springer, Heidelberg (2008)
39. Glaser, H., Jaffri, A., Millard, I.: Managing Co-reference on the Semantic Web. In: WWW 2009 Work. Linked Data Web, Madrid, Spain (2009)