Customisable Query Resolution in Biology and Medicine

Peter Ansell[1], **James Hogan**[1] **and Paul Roe**[1]

[1] School of Information Technology
Queensland University of Technology,
2 George Street, Brisbane, Queensland, 4000
Email: p.ansell@qut.edu.au, j.hogan@qut.edu.au and p.roe@qut.edu.au

Abstract

Scientists and healthcare workers regularly use data from a number of sources as part of their research and professional work. The current mechanisms for providing combined access to multiple datasources are either closed or not easily extensible, with some requiring users to locally load and query each datasource independently. In this work we introduce a new model for transparent querying across multiple datasources which relies on the single unifying format of RDF to merge information before returning it to users. The use of normalised, resolvable URI's, combined with the SPARQL RDF query language, enables common queries to be executed across multiple public and private datasources, including those not initially designed or represented using RDF. In order to accommodate a range of users, the implemented system has been set up to enable customisation of queries and datasources based on RDF formatted configuration files. This breadth of data and configurability allows scientists and healthcare workers to more efficiently find and communicate semantic references, supporting research, professional practice and dissemination of knowledge across communities and disciplines.

1 Introduction

The areas of science and medicine rely on information transfer between organisations to make sure each organisation is taking advantage of the latest innovations and discoveries. These sources of information are varied in nature and diverse in location, with users sometimes requiring data from many locations to make the best decisions. The ability to cross between disciplines, for example from medicine through to genomics and chemistry, generally requires a large amount of expertise, including an understanding of each of the relevant data formats. Particularly in medicine, organisations need to be able to utilise both external and local knowledge bases as part of their decision making processes. The combination of distributed, cross-discipline, and potentially private knowledge provides a case for a system designed around a single knowledge representation format. In this system users who are unfamiliar with a particular discipline can utilise a single query method to explore the information and decide on the importance of the information without requiring the intervention of a domain expert. The use of a single extensible format enables organisations to insert their own, potentially private, information into documents without requiring a redesign of the file format or any external data disclosure.

Recently, there have been a number of datasources, representing science, medicine, and other areas (Auer, Bizer, Kobilarov, Lehmann, Cyganiak & Ives 2007, Belleau, Nolin, Tourigny, Rigault & Morissette 2008, Ruttenberg, Rees, Samwald & Marshall 2009), which have been republished using RDF (Resource Description Format). RDF is a domain-neutral information format that can be easily extended by organisations to include references to their own data where necessary without requiring them to customise the file structure to their particular needs. Queries across RDF datasources can be performed without a knowledge of the particular properties used by any of the relevant datasources. Knowledge of the properties used by particular datasources may, however, be used to integrate knowledge from multiple datasources into homogeneous documents. The ease of querying provided by RDF query languages such as SPARQL (SPARQL Query Language for RDF)[1], enables users to share and customise queries easily, and in some cases perform the same queries across datasources from different disciplines.

This paper presents a novel design for a cross-database, customisable query system based on RDF. It includes a description of the model and a prototype implementation, before attempting to show how they could be used and adapted to fit health informatics requirements. Section 2 provides some background into both RDF and non-RDF projects that attempt to integrate or describe large datasources. A brief description of the elements that make up the distributed query model, including ways to integrate other sources of information, is given in Section 3. The applicability of the distributed RDF query model to health informatics and clinical biologists is described in Section 4, along with a case study that starts with a drug and explores the corresponding links to genomics, cheminformatics, and clinical trial datasources. The interlinked datasources provide use cases and future research paths for both the drug and patients who may benefit from the drug. Section 5 discusses issues that relate to the model and its use in the context of health informatics.

2 Background

RDF versions of scientific datasources, have been created by projects such as Bio2RDF (Belleau et al. 2008), Neurocommons (Ruttenberg et al. 2009), Flyweb (Zhao, Klyne & Shotton 2008), and Linked Open

[1] http://www.w3.org/TR/rdf-sparql-query/

Drug Data (LODD)[2]. Where possible, the RDF documents produced by these organisations utilise HTTP URI's to link to RDF documents produced by the other organisations. This is useful, as it matches the basic Linked Data[3] goals which are designed to ensure that data represented in RDF is accessible and contextually linked to other data as required.

SPARQL queries that are required for complex investigations are in most cases limited to execution on a single RDF database. This project aims to simplify the distribution of queries across multiple datasources, given that it is impractical to expect every datasource to be copied to a single local database for complex queries. The majority of systems which attempt to distribute SPARQL queries across a number of endpoints, convert single SPARQL queries into multiple SPARQL queries, before joining and filtering the results to match the original query. These systems generally require that users configure the system with specific knowledge of properties used by each datasource, and in some cases the join may require a large amount of information to be moved, rendering the method relatively inefficient (Adamku & Stuckenschmidt 2005, Langegger, Blochl & Woss 2007, Quilitz & Leser 2008). Other similar systems also require query designers to insert the URL's of each of the datasource endpoints into their queries by redefining the meaning of a SPARQL keyword, making complex datasources inaccessible (Zemánek, Schenk & Svátek 2008). Systems that focus on RDF query performance improvements from a parallelism point of view do not generally require users to use specific predicates, but they may require a suitable distribution of information across a local set of RDF endpoints in order to facilitate fairly random access to specific RDF statements (Harth, Umbrich, Hogan & Decker 2007). The number of requirements given by these query systems highlight the need for a simple method of customising and extending results, particularly in the case where private datasources must be accessed using a unique, secure, method that cannot be disclosed to other users.

The BioGUID project (Page 2009) provides a single point of entry for users to obtain RDF descriptions from a range of datasets. However, it does not provide a generalised mechanism for resolution, as the RDF resolvers are implemented as a set of tools rather than a single configurable implementation. Other projects such as the Distributed Annotation System (DAS) (Prlić, Birney, Cox, Down, Finn, Grf, Jackson, Khri, Kulesha, Pettett, Smith, Stalker & Hubbard 2006) allow distributed customised querying but do not use RDF, so discipline specific file formats must be understood by any software utilising the resulting documents. In comparison to an RDF based query solution, the current DAS implementation suffers in that it requires software updates to support any new classes of information being added to the system. In comparison, RDF based systems can be extended without having multiple data models implemented in software. RDF based solutions enable users to extend an official implementation in a completely valid way using their own predicates, enabling customised configuration-based additions as well as the up to date alternative annotation data that DAS was designed to provide.

The SRS system (Etzold, Harris & Beulah 2003) provides an integrated set of biological datasources, with a custom query language and internal addressing scheme. Although the internal identifiers are unambiguous in the context of the SRS system, they do not have a clear meaning when used in other contexts. In comparison to the many formats offered by SRS and the native document formats of particular scientific datasources, the use of RDF for both documents and query results provides a single method, URI's, to reference items from any of the involved datasources. The locally integrated model that SRS relies on for its queries is not sustainable as the size and number of datasources grows. The approximate number of RDF statements–similar in nature to SQL database records–that are required to represent each of the largest 14 databases in the Bio2RDF project as shown in Table 1, illustrating the scale of the information provided currently in distributed RDF datasources. SRS provides a generic query language, that makes use of the localised database, giving it performance advantages over the distributed RDF query system described in this paper.

Database	Approximate RDF statements
PDB	10,000,000,000
Genbank	5,000,000,000
Refseq	2,600,000,000
Pubmed	1,000,000,000
Uniprot Uniref	800,000,000
Uniprot Uniparc	710,000,000
Uniprot Uniprot	220,000,000
IProClass	182,000,000
NCBI Entrez Geneid	156,000,000
Kegg Pathway	52,000,000
Biocyc	34,000,000
Gene Ontology (GO)	7,400,000
Chebi	5,000,000
NCBI Homologene	4,500,000

Table 1: Bio2RDF datasource sizes

3 Model

In order to allow a simple method of performing and customising queries across the many potential datasources, an easily extensible model was designed and implemented. The model consists of a chain of elements starting with a user query, typically given as part of a URL, and ending with a set of RDF statements. The chain is started by matching the user query against a set of Query Types, any of which could be used in parallel to respond to the query. Each Query Type can be configured to identify the relevant namespace, as required, and utilise these namespaces as the basis of distributing the query across any applicable providers. For each provider, normalisation rules may be configured to be applied to the parts of the query that have been defined to be specific to each endpoint.

The resulting RDF information from each provider is then transformed back using the output part of the normalisation rules which were configured for the provider. The normalised information is then merged into an overall pool of RDF statements that will be returned in a single document to the user. The model focuses on pooling information, as this method provides the simplest way of retrieving information from datasources that may not all use the same query structure or interface. In order to provide for different users of a widely distributed set of configuration information, profiles have been included in order to allow varying levels of flexibility with respect to the inclusion or exclusion of Query Types, Providers and Normalisation Rules based on a user's goals. Profiles make it simple for users to customise the local configuration by adding their own RDF configuration

[2] http://esw.w3.org/topic/HCLSIG/LODD

[3] http://www.w3.org/DesignIssues/LinkedData.html

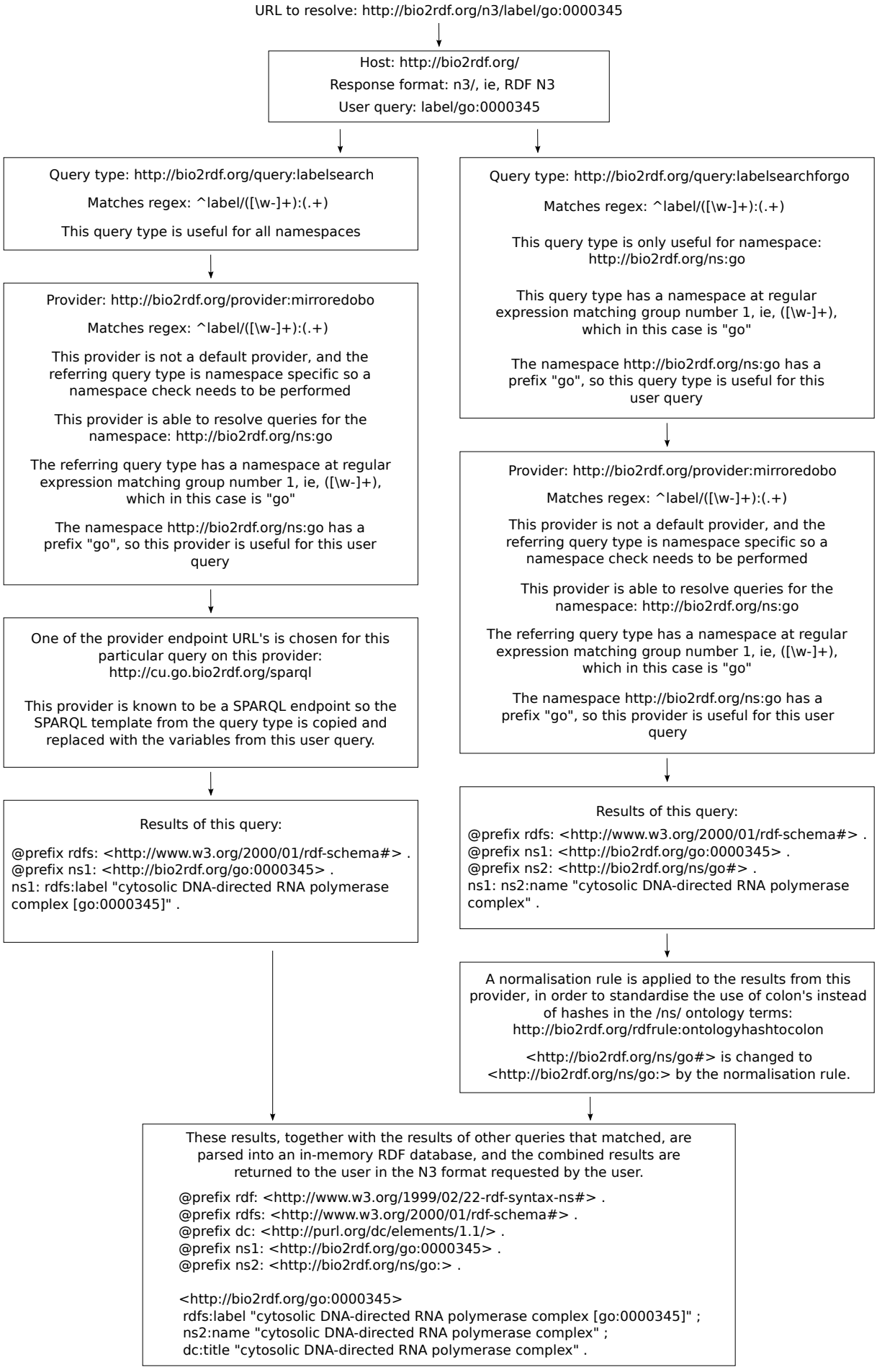

Figure 1: URL resolution using model

snippets.

An example illustrating the steps required by the model, to retrieve a list of labels for the Gene Ontology (GO) item with identifier "0000345", known as "cytosolic DNA-directed RNA polymerase complex", is shown in Figure 1. It illustrates the combination of a generic query, along with a query that is customised for the GO datasource. The queries are designed so that the generic query will be used on any information provider, while the custom GO query will be restricted to providers that contain GO information. If another datasource was available to retrieve labels for GO terms using RDF, then a custom query definition could be added in parallel to these two queries without any side effects.

A prototype server was implemented using Java and JSP and is currently in use by resolvers of the `http://bio2rdf.org/` website. The implementation allows users to select between the different RDF file formats, including an HTML page displaying a list of resources in the corresponding RDF document. The implementation was used to query across all of the Bio2RDF datasets, all of the LODD datasets, most of the Neurocommons (Ruttenberg et al. 2009) datasets, and the DBpedia (Auer et al. 2007) dataset (including the pagelinks set), using namespaces created using the `http://bio2rdf.org/` authority. The model defines the RDF statements required for implementations to use when creating configurations, meaning other implementations can utilise configurations created using the same version of the model vocabulary without reference to the implementation they were originally created or used on.

The simplest possible configuration consists of a single Query Type and a single Provider, as shown in Figure 2. The Query Type needs to be configured with a regular expression that matches user queries. The provider needs to be configured with both a reference to the Query Type, and an endpoint URL that can be used to resolve queries matching Query Type. Although the example trivial, in that the user's query is directly passed to another location, it provides an overview of the features that make up a configuration. One particular feature to be noted is the use of the profile directive to process profile exclude instructions first, and then include in all other cases. In this example, there are no profiles defined, resulting in the items being included in the processing of queries that match the definitions.

```
@prefix query: <http://purl.org/queryall/query:> .
@prefix provider: <http://purl.org/queryall/provider:> .
@prefix profile: <http://purl.org/queryall/profile:> .
@prefix : <http://example.org/> .

:myquery a query:Query ;
   query:inputRegex "(.*)" ;
   profile:profileIncludeExcludeOrder
      profile:excludeThenInclude .

:myprovider a provider:Provider ;
   provider:resolutionStrategy provider:proxy ;
   provider:resolutionMethod provider:httpgeturl ;
   provider:isDefaultSource "true"^^<http://www.w3.org/2001/
XMLSchema#boolean> ;
   provider:endpointUrl "http://myhost.org/${input_1}" ;
   provider:includedInQuery :myquery ;
   profile:profileIncludeExcludeOrder
      profile:excludeThenIncludae .
```

Figure 2: Simple system configuration in Turtle RDF file format

4 Applicability to Health Informatics

The model is designed so that it can be easily customised by users. Extensions can range from additional sources and queries to removal of sources or queries for efficiency or other reasons. This functionality provides a simple way for users to select which sources of information they want to use without having to make choices about every published information source.

In the context of Health Informatics, a hospital may want to utilise information from a drug information site, such as DrugBank and DailyMed, together with their private medical files. In order to do this, the hospital could map references in their medical files to DrugBank and/or DailyMed identifiers and publish the resulting information into RDF. The RDF statements could then be integrated with the DrugBank information without any further changes. They hospital may then create a mapping between the terminologies stored in their internal database and those used by DrugBank and related datasources. These mappings could be made using one of the available SQL to SPARQL converters such as the Virtuoso RDF Views mechanism (Erling & Mikhailov 2007) or the D2RQ server (Bizer & Seaborne 2004).

To distinguish private records from external public datasources, hospitals should create a namespace for their internal records, along with providers matching the internal addresses used for queries about their records. This novel private information, is able to be safely included in the model through the use of private provider configurations indicating the source and the particular RDF formats in which the information is available.

If the hospital then wanted to map a list of diseases into their files, they could find a source for disease descriptions, such as Diseaseome, and either find existing links through DrugBank and DailyMed, or they could attempt to use text mining to discover common disease names between their records and Diseaseome. For scientific research, resources like Diseasome are linked to bioinformatics databases such as the NCBI Entrez GeneID, PDB, PFAM, and OMIM databases. This linkage is defined explicitly in RDF and enables links to be discovered, for instance there may be links from patients and clinical trials to genetic factors. Patients could be directly linked to genes using RDF syntax without reference to diseases, and new diseases could be described internally without requiring outside publication.

Potential side effects are accessible through the Sider database, enabling patients and doctors to both have access to equal information about the potential side effects of a particular course of medication. Side effects that are discovered by the hospital could be recorded using references to the public Sider record, reducing the possibility that the effect would be missed in future cases.

4.1 Case Study

The continuous use of RDF enables users to transition between databases using URI's without having to register a new file format for each database. For ease of reading, the URI's, which can be resolved to retrieve the relevant information used in the following case study, are footnoted. The links between the datasets were observed by resolving the URI's and finding RDF statements inside the document that link to other URI's. This case study utilises a range of datasets that are shown in Figure 3. These datasets are sourced from Bio2RDF, LODD, Neurocommons, and DBpedia. The original image can be found at `http://www4.wiwiss.fu-berlin.`

de/lodd/lodd-datasets_2009-08-06.png. A portion of the case study, highlighting the links between items in the relevant datasources, can be seen in Figure 4.

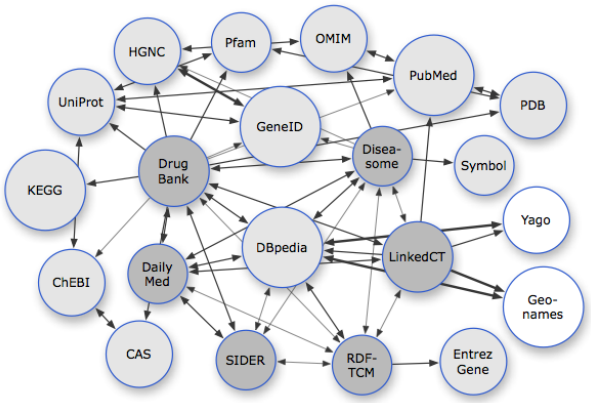

Figure 3: Medicine related RDF datasets

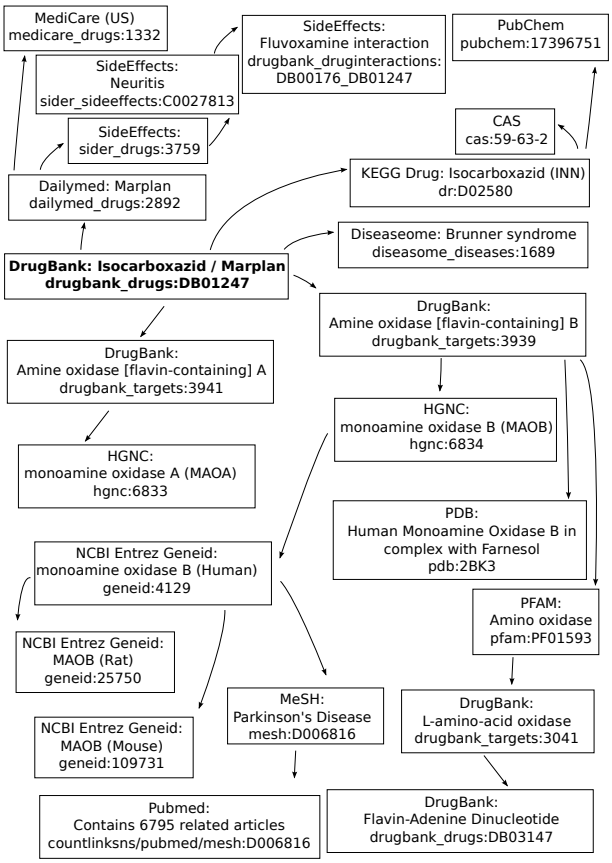

Figure 4: Inter-datasource links in Isocarboxazid case study

This case study is founded around a drug known generically as "Isocarboxazid". It is also known by the brand name "Marplan". The aims of this case study are to discover potential relationships between this drug and patients with reference to publications, genes and proteins that may affect the course of their treatment. In cases where patients are known to have adverse reactions or they do not respond positively to treatment, alternatives may be found by examining the usefulness of drugs that are designed for similar purposes. For the purposes of this case study, the relevant entry in DrugBank is known[4]. According to this DrugBank record, Isocarboxazid is "[a]n MAO inhibitor that is effective in the treatment of major depression, dysthymic disorder, and atypical depression".

The DrugBank entry for Isocarboxazid contains links to the CAS (Chemical Abstracts Service) registry[5], which in turn contains links to the KEGG (Kyoto Encyclopedia of Genes and Genomes) Drug database[6]. The link back to DrugBank from the KEGG Drug database and others in this case could also have been discovered using only the original DrugBank namespace and identifier[7]. The brand name drug database, Dailymed, also contains a description for Marplan[8], which is linked from Sider[9] and the US MediCare database[10]. These alternative URI's could be used to identify more datasources with information about the drug.

The record for Isocarboxazid in the Sider database has a number of typical depression side-effects to watch for, but it also has a potential link to Neuritis[11] a symptom which is different to most of the other 39 side effects that are more clearly depression related. Along with side effects, there are also known drug interactions available using the DrugBank database. An example of these is an indication of a possible adverse reaction between Isocarboxazid and Fluvoxamine[12]. If Fluvoxamine was already being given to the patient, other drugs may need to be investigated, as alternatives to prevent the possibility of a more serious Neuritis side effect. DrugBank contains a simple categorisation system that might reveal useful alternative Antidepressants[13], in this case, such as Nortriptyline[14].

Dailymed contains a list of typically inactive ingredients in each brand-name drug, such as Lactose[15], which may factor into a decision to use one version of a drug over others. The Drugbank entry for Isocarboxazid also contains links to Diseasome, for example, Brunner syndrome[16], which are linked to the OMIM (Online Mendelian Inheritance in Man) entry for Monoamine oxidase A (MAOA)[17].

DrugBank also contains a list of biological targets that Isocarboxazid is known to effect[18]. If Isocarboxazid was not suitable, drugs which also affect this gene; Monoamine oxidase B (MOAB)[19][20][21][22], the protein[23], or the protein family[24], might also cause a similar reaction. The negative link (in this case derived using text mining techniques) between the target gene, monoamine oxidase B,[25], and Huntington's Disease[26][27], might cause a doctor to decide not to give the drug to a patient with a history of Huntington's.

[4] http://bio2rdf.org/drugbank_drugs:DB01247

[5] http://bio2rdf.org/cas:59-63-2
[6] http://bio2rdf.org/dr:D02580
[7] http://bio2rdf.org/links/drugbank_drugs:DB01247
[8] http://bio2rdf.org/dailymed_drugs:2892
[9] http://bio2rdf.org/sider_drugs:3759
[10] http://bio2rdf.org/medicare_drugs:13323
[11] http://bio2rdf.org/sider_sideeffects:C0027813
[12] http://bio2rdf.org/drugbank_druginteractions:DB00176_DB01247
[13] http://bio2rdf.org/drugbank_drugcategory:antidepressants
[14] http://bio2rdf.org/drugbank_drugs:DB00540
[15] http://bio2rdf.org/dailymed_ingredient:lactose
[16] http://bio2rdf.org/diseasome_diseases:1689
[17] http://bio2rdf.org/omim:309850
[18] http://bio2rdf.org/drugbank_targets:3939
[19] http://bio2rdf.org/symbol:MAOB
[20] http://bio2rdf.org/hgnc:6834
[21] http://bio2rdf.org/geneid:4129
[22] http://bio2rdf.org/mgi:96916
[23] http://bio2rdf.org/uniprot:P27338
[24] http://bio2rdf.org/pfam:PF01593
[25] http://bio2rdf.org/geneid:4129
[26] http://bio2rdf.org/mesh:Huntington_Disease
[27] http://bio2rdf.org/mesh:D006816

The location of the MAOB gene on the X chromosome in Humans might warrant an investigation into gender related issues related to the original drug, Isocarboxazid. The homologous MAOB genes in Mice[28] and Rats[29], are also located on chromosome X, indicating that they might be useful targets for non-Human trials studying gender related differences in the effects of the drug.

The Human gene MAOA , can be found in the Traditional Chinese Medicine (TCM) database,[30], as can MAOB[31], although there were no direct links from the Entrez Geneid database to the TCM database. TCM has a range of herbal remedies listed as being relevant to the MAOB gene[32] including Psoralea corylifolia[33]. Psoralea corylifolia is also listed as being relevant to another gene, Superoxide dismutase 1 (SOD1)[34][35]. SOD1 is known to be related to Amyotrophic Lateral Sclerosis[36][37], although the relationship back to the Brunner Syndrome and Isocarboxazid, if any, may only be exploratory given the range of datasources in between.

LinkedCT is an RDF version of the ClinicalTrials.gov website that was setup to register basic information about clinical trials. It provides access to clinical information, and consequently is a rough guide to the level of testing that various treatments have had. The drug and disease databases mentioned above link to individual clinical interventions in LinkedCT, enabling a path between the drugs, affected genes and trials relating to the drugs. Although there are no direct links from LinkedCT to Marplan at the time of publication, a namespace based text search returns a list of potentially interesting items[38]. An example of a result from this search is a trial[39] conducted by John S. March, MD, MPH[40] of Duke University School of Medicine and overseen by the US Government[41]. The trial references published articles, including one titled "The case for practical clinical trials in psychiatry"[42][43]. These articles are linked to textual MeSH (Medical Subject Headings) terms such as "Psychiatry - methods"[44], indicating an area that the study may be related to. The trial is linked to specific primary outcomes and the frequency with which the outcomes were tested, giving information about the scientific methods in use[45].

Although LinkedCT is a useful resource, as with any other resource, there are difficulties with the data being complete and correct. An example of this are recent studies about the use of ClinicalTrials.gov which indicate that a reasonable percentage of clinical trials either do not publish results, register with ClinicalTrials.gov, or reference the ClinicalTrials record in publications resulting from the research (Ross, Mulvey, Hines, Nissen & Krumholz 2009, Mathieu, Boutron, Moher, Altman & Ravaud 2009). These issues may be reduced if people were required to register all drug trials and reference the entry in any publications.

Doctors and patients do not have to know what the URI for a particular resource is, as there is a search functionality available. This searching can either be focused on particular namespaces or it can be performed over the entire known set of RDF datasources, although the latter will inevitably be slower than a focused search as some datasources are up to hundreds of gigabytes in size, representing billions of RDF statements. An example of this may be a search for "MAOB"[46], which reveals resources that were not included in this brief case study.

5 Discussion

Given that the system uses URI's for all of the universal identifiers internally, and they are designed to all be resolvable, it is possible to show labels to humans, and have URI's for computers to use without prejudicing the system to either party. The RDF datasets have been designed with this in mind and the majority should include triples that indicate what the best label for a given resource is after it has been resolved. If an application recognises a URI as fitting the Bio2RDF system it can get labels for a URI using "label/namespace:identifier"[47]. In order to reduce the latency involved with resolving a set of URI's, a list of labels can be resoled using the format "multiplelabel/namespace1:identifier1 / namespace2:identifier2.."[48].

Health Informatics requires that multiple systems be integrated in order to answer questions such as, "what observations were made for patients with heart disease, in the past 2 months, who were not on drugs that have current clinical trials". In order to answer this question multiple datasources may be required, including medical terminology repositories, patient databases, drug databases, and clinical trial databases. The mapping process may be an imprecise operation, as doctors may use ambiguous shorthand notations for their observations, and medical terminology repositories may not contain an exact term for a particular condition. This integration process may still benefit from the use of RDF, as a mapped URI, even if it is incorrect, is unambiguous, and can be identified immediately by resolving the URI in order to verify its suitability.

The ability to easily mix arbitrary sources of information provides both benefits and complications. The major benefits come from the ability to traverse the mixed dataset using a single file structure, RDF, and from being able to publish novel and mixed datasets in the same form for others to use. In the context of science and medicine this provides the ability to annotate studies and factual databases with extra information and provide that information to other users without having to republish the entire database or extend the file format originally used for the database. These benefits, however, also bring complications relating to provenance, privacy and reliability if any of the datasources or users are not trusted. The risk of these reliability complications can be reduced by only including queries and providers that are reasonably trusted.

The query model described here is designed to easily provide access to public datasources using configurations published by organisations such as Bio2RDF, while simultaneously allowing access to private internal datasources using unpublished configurations. Queries are executed using whatever permissions the

[28] http://bio2rdf.org/geneid:109731
[29] http://bio2rdf.org/geneid:25750
[30] http://bio2rdf.org/linksns/tcm_gene/geneid:4128
[31] http://bio2rdf.org/linksns/tcm_gene/geneid:4129
[32] http://bio2rdf.org/linksns/tcm_medicine/tcm_gene:MAOB
[33] http://bio2rdf.org/tcm_medicine:Psoralea_corylifolia
[34] http://bio2rdf.org/tcm_gene:SOD1
[35] http://bio2rdf.org/geneid:6647
[36] http://bio2rdf.org/mesh:Amyotrophic_Lateral_Sclerosis
[37] http://bio2rdf.org/mesh:D000690
[38] http://bio2rdf.org/searchns/linkedct_trials/marplan
[39] http://bio2rdf.org/linkedct_trials:NCT00395213
[40] http://bio2rdf.org/linkedct_overall_official:12333
[41] http://bio2rdf.org/linkedct_oversight:2283
[42] http://bio2rdf.org/linkedct_reference:22113
[43] http://bio2rdf.org/pubmed:15863782
[44] http://bio2rdf.org/mesh:D011570Q000379
[45] http://bio2rdf.org/linkedct_primary_outcomes:55439
[46] http://bio2rdf.org/search/MAOB
[47] http://bio2rdf.org/label/pubmed:15863782
[48] http://bio2rdf.org/multiplelabel/pubmed:15863782/omim:309860

resolving server has, although authentication systems appropriate to each site could be used by modifying the code used for the resolving server to perform the authentication prior to performing the query.

Although some researchers deny that the use of RDF with URI's for entities such as humans and closely related records will cause new privacy issues, the ease and effectiveness with which different RDF documents can be merged has to be considered as a potential privacy issue (Feigenbaum, Herman, Hongsermeier, Neumann & Stephens 2007). They argue that the potential privacy issues are related to the applications that use the data, and that the issues are countered by the benefits that are obtained through the use of the integrated information. In the case of RDF though, the ability to merge documents is a feature, meaning that the merge of a patients record with their credit history, for instance, may be far simpler than would be acceptable for patients. The URI's assigned to humans inside of the system should however be opaque and not contain identifying information such as age, location, or gender, which are particularly easy to map using RDF based reasoning tools in order to re-identify things. The overall RDF scheme has no recognised standard for stating what the rights and restrictions attached to a piece of information are, although there are a few schemes that attempt to perform this operation in limited circumstances. The model and implementation described in this work do not introduce new privacy issues, as the most important privacy issue in both cases is to preventing unauthorised bulk access to patient records, something which must be restricted using authentication and authorisation policies.

6 Conclusion

Knowledge management and informatics scenarios require that contextual information be provided to give evidence and background for particular sets of information. The aim of the distributed query system described here is to easily provide links into each of the relevant sets of information. The links are designed to make it easy for applications to cross traditional knowledge boundaries, such as the boundary between chemical theory and clinical pathology observations. The information required to convert a link into a document is all contained within the link reference, so applications only need to know how to resolve HTTP URL's in order to retrieve the related information. The document returned will always be in a known model, RDF, with a choice of file formats that represent the model, so applications can easily interpret the content without having to implement each known data structure in program code.

The steps required to integrate new sources of information into the query model include mapping current datasources to RDF; mapping relevant references to other datasources using URI's; and finally creation of configuration definitions describing the queries and locations that can be used to access information from the datasource.

The case study showed the diversity of links between the different health and biology related datasources currently available in RDF. It showed that the datasources were transparently accessible using the distributed query model and that URI based links between datasources can provide new insights that may not be easy to discover in current systems. Public datasources currently provide information about drugs, public drug trials, diseases, genetic information, and publications; while private datasources may provide linked information about patients, treatments, and private drug trials, in the future. These can all be integrated by customising the distributed query model for each particular need through the use of simple RDF configuration files.

7 Acknowledgements

This research was funded through a Smart State National & International Research Alliance Scholarship by the Queensland State Government and Microsoft Research. It was supported by the Microsoft Queensland University of Technology eResearch Centre.

References

Adamku, G. & Stuckenschmidt, H. (2005), Implementation and evaluation of a distributed rdf storage and retrieval system, *in* 'Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)', IEEE Computer Society, Los Alamitos, CA, USA, pp. 393–396.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. (2007), 'Dbpedia: A nucleus for a web of open data', *Lecture Notes in Computer Science* **4825**, 722.

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. & Morissette, J. (2008), 'Bio2rdf: Towards a mashup to build bioinformatics knowledge systems', *Journal of Biomedical Informatics* **41**(5), 706–716.

Bizer, C. & Seaborne, A. (2004), D2rq-treating non-rdf databases as virtual rdf graphs, *in* 'Proceedings of the 3rd International Semantic Web Conference (ISWC2004)', Citeseer.

Erling, O. & Mikhailov, I. (2007), Rdf support in the virtuoso dbms, *in* 'Proceedings of the 1st Conference on Social Semantic Web (CSSW)', Springer, pp. 7–24.

Etzold, T., Harris, H. & Beulah, S. (2003), 'SRS: An integration platform for databanks and analysis tools in bioinformatics', *Bioinformatics Managing Scientific Data* pp. 35–74.

Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E. & Stephens, S. (2007), 'The semantic web in action', *Scientific American* **297**, 90–97.

Harth, A., Umbrich, J., Hogan, A. & Decker, S. (2007), 'Yars2: A federated repository for querying graph structured data from the webs', *Lecture Notes in Computer Science* **4825**, 211.

Langegger, A., Blochl, M. & Woss, W. (2007), Sharing data on the grid using ontologies and distributed sparql queries, *in* '18th International Conference on Database and Expert Systems Applications, 2007. DEXA '07', pp. 450–454.

Mathieu, S., Boutron, I., Moher, D., Altman, D. G. & Ravaud, P. (2009), 'Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials', *JAMA* **302**(9), 977–984.

Page, R. (2009), Bioguid: resolving, discovering, and minting identifiers for biodiversity informatics. Available from Nature Precedings `http://hdl.handle.net/10101/npre.2009.3079.11`.

Prlić, A., Birney, E., Cox, T., Down, T., Finn, R., Grf, S., Jackson, D., Khri, A., Kulesha, E., Pettett, R., Smith, J., Stalker, J. & Hubbard, T. (2006), *The Distributed Annotation System for Integration of Biological Data*, pp. 195–203.

Quilitz, B. & Leser, U. (2008), 'Querying distributed rdf data sources with sparql', *Lecture Notes in Computer Science* **5021**, 524.

Ross, J. S., Mulvey, G. K., Hines, E. M., Nissen, S. E. & Krumholz, H. M. (2009), 'Trial publication after registration in clinicaltrials.gov: A cross-sectional analysis', *PLoS Med* **6**(9), e1000144.

Ruttenberg, A., Rees, J., Samwald, M. & Marshall, M. (2009), 'Life sciences on the semantic web: the neurocommons and beyond', *Briefings in Bioinformatics* **10**(2), 193.

Zemánek, J., Schenk, S. & Svátek, V. (2008), Optimizing sparql queries over disparate rdf data sources through distributed semi-joins.

Zhao, J., Klyne, G. & Shotton, D. (2008), Provenance and linked data in biological data webs, *in* 'Linked Open Data Workshop at The 17th International World Wide Web Conference'.