A virtual SPARQL endpoint to support IRI translation using Identifiers.org

Sarala M Wimalaratne [1*], Jerven Bolleman [2], Nick Jutty [1], Toshiaki Katayama [3], Michel Dumontier [4], Nicole Redaschi [2], Henning Hermjakob [1], Nicolas Le Novere [1] and Camille Laibe [1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2]Swiss-Prot group, SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1211 Geneve, Switzerland
[3]Database Center for Life Science, Research Organization of Information and Systems, 178-4-4 Wakashiba, Kashiwa, Chiba 277-0871, Japan
[4]Stanford Center for Biomedical Informatics Research, Stanford University, CA 94305-5479, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT
Motivation: On the semantic web, in life science data in particular, the data is often distributed via multiple resources. Each of these sources is likely to use their own IRI (International Resource Identifier) for conceptually the same resource or database record. This multitude of identifiers introduce a barrier when executing federated SPARQL queries between different resources.
Results: We introduce a virtual SPARQL endpoint that uses Identifiers.org Registry knowledge about IRI patterns to automatically translate identifiers in one IRI pattern to another, making cross database querying easier and more robust. This endpoint supports the full SPARQL 1.1.
Availability: This IRI translating endpoint is available at http://identifiers.org/services/sparql and is free to use.
Contact: Sarala M Wimalaratne

1 INTRODUCTION

The use of Semantic Web technologies such as Resource Description Framework (RDF; Manola and Miller (2004)) for publishing data is becoming popular in the life sciences. SPARQL (Prud'hommeaux and Seaborne (2008)) is a query language for RDF, which supports complex queries, and can merge data distributed over multiple RDF resources. Thus provides a platform for linking and querying ever growing life science data resources.

Bio2RDF (Belleau *et al.* (2008)) data mashup is one such resource that contains publicly available databases in RDF linked together. Normalized URIs in Bio2RDF domain is introduced to support federated queries cross datasets. More recently, to meet the needs of the semantic community, EBI has exposed some of their datasets in RDF via the EBI RDF platform (Jupp *et al.* (2014)). To provide a consistent URI scheme and enable federated queries, each dataset follows a set of rules when creating new URIs for their datasets and pointing to external resources (Jupp *et al.* (2014)).

The two platforms provide linked datasets which can be queried easily as individual resources. The use of two different URI schemes for what is conceptually the same resource makes it difficult to perform cross platform queries. Thus a cross database query between resources needs to rewrite IRIs on the fly. This also requires users to know in advance the IRI type being used in the resource. In addition, string manipulation supported by SPARQL language is used to construct relevant IRIs.

To resolve such cross queries issues we have developed a SPARQL endpoint to support IRI conversion between datasets using the content in Identifiers.org Registry.

Identifiers.org (Juty *et al.* (2012)) Registry contains information about datasets that are distributed via multiple resources, using different IRIs to identify individual records. For example, plasma membrane in Gene ontology is available from AmiGO (Carbon *et al.* (2009)), QuickGO (Binns *et al.* (2009)), Bio2RDF etc. Identifiers.org also provides resolvable persistent IRIs used to identify data based on existing local record identifiers already assigned by the data providers. The provision of a resolvable identifiers (URLs) fits well with the Semantic Web vision, and the Linked Data initiative. This has lead to some resources adopting Identifiers.org IRIs to publishing and referencing resources. The knowledge in the Identifiers.org Registry can be used to convert between IRIs used in different datasets.

2 METHODS

OpenRDF Sesame SPARQL engine (Visser *et al.* (2014)) was extended to translate specific query patterns to allow IRI translation. The main integration point is the $getStatements$ method of a class extending $org.openrdf.query.algebra.evaluation.TripleSource$.

[*]to whom correspondence should be addressed

Using the information in the Identifiers.org Registry, IRI conversion statements are generated programatically. The result is exposed as a virtual SPARQL endpoint at http://identifiers.org/services/sparql

LODEStar interface http://www.ebi.ac.uk/fgpt/sw/lodestar/ used by the EBI RDF platform is reused for querying virtual Identifiers.org SPARQL endpoint.

3 RESULTS

The Identifiers.org virtual SPARQL endpoint can be used to list all different URI schemes available for a given URI. The endpoint also could be used to check whether two IRIs describes the same concept. Example queries can be found at the Identifiers.org SPARQL endpoint.

With the new Identifiers.org SPARQL service, it is possible to eliminate string manipulation within SPARQL query (figure 1). This solution also does not require the users to know which IRI patter is being used within a resource.

The example below illustrates how the service could be used to integrate resources between the EBI RDF platform and Bio2RDF. This query lists GO annotation and its description a particulate model. It is executed from BioModels SPARQL endpoint at http://www.ebi.ac.uk/rdf/services/biomodels/sparql and connects to Identifiers.org to retrieve IRI conversions. These are then passed into Bio2RDF SPARQL endpoint to collect descriptions about each GO term.

```
SELECT DISTINCT ?species ?annotation ?description WHERE {
   <http://identifiers.org/biomodels.db/BIOMD0000000001>
      sbmlrdf:species ?species .
   ?species bqbio:isVersionOf ?annotation .

   SERVICE <http://identifiers.org/services/sparql>{
      ?annotation owl:sameAs ?otherURIs .
   }

   SERVICE <http://bioportal.bio2rdf.org/sparql>{
      ?otherURIs dcterms:description ?description .
   }
}LIMIT 10
```

Fig. 1. Using virtual Identifiers.org SPARQL endpoint to transform Identifiers.org IRIs into Bio2RDF IRIs.

4 DISCUSSION

There are a number of mapping services aimed to map identifiers from one databases to an other. These are conceptual mappings e.g. they relate an identifier for a gene to its related protien products. In comparison Identifiers.org has a different purpose, it identifies all sources of information about a single database record as identified by one identifier; independently of the information provider. Identifiers.org could maintain lists of currently valid IRIs used in all SPARQL endpoints and in RDF datasets, unfortunately this would be extremely time consuming and always out of sync. Thus the SPARQL service described in this paper only translates patterns and leaves it up to the real third party SPARQL endpoints to determine if a translated IRI is valid or not.

While the SPARQL service is SPARQL 1.1. compliant it's virtual nature can surprise end users. If the user does not put in the IRI `owl:sameAs` as the predicate in a basic graph pattern in the SPARQL query, the query will not return results at all. The SPARQL correct solution would be able to return the infinite possible IRI translations, however that technically correct solution is not at all usefull for any user. Instead we take the approach that Identifiers.org SPARQL endpoint is a remote service not under direct control by the user, and in our case it is under control of entity that deletes everything in the graph when such a query arrives and restores it after the query is answered. This solution allows the end-user the illusion of a valid SPARQL service that can translate all IRI patterns in the Identifier.org Registry as if all these queries where materialized, at a fraction of the cost.

5 CONCLUSION

Making identifiers.org mappings available on the fly make it easier to effectively use the federated query capacities of SPARQL 1.1 to combine information from multiple sources.

ACKNOWLEDGEMENT

The authors thank the DBCLS RDF Summit organisers for hosting initial software development and S. Jupp for developing the user interface to browse the linked dataset.

Funding: Japanese grants, BBSRC (BB/J019305/1), EMBL and European Union grants XXXXX. National Institutes of Health grants [1U41HG006104] and the Swiss Federal Government through the Federal Office of Education and Science.

REFERENCES

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. Semantic Mashup of Biomedical Data.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics*, **25**(22), 3045–3046.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Lomax, J., Mungall, C., Hitz, B., Balakrishnan, R., Dolan, M., Wood, V., Hong, E., and Gaudet, P. (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics*, **25**(2), 288–289.

Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S. M., Martin, M., Le Novre, N., Parkinson, H., Birney, E., and Jenkinson, A. M. (2014). The ebi rdf platform: linked open data for the life sciences. *Bioinformatics*, **30**(9), 1338–1339.

Juty, N., Le Novre, N., and Laibe, C. (2012). Identifiers.org and miriam registry: community resources to provide persistent identification. *Nucleic Acids Research*, **40**(D1), D580–D586.

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2011). Wikipathways: building research communities on biological pathways. *Nucleic Acids Research*.

Manola, F. and Miller, E. (2004). RDF Primer.

Prud'hommeaux, E. and Seaborne, A. (2008). SPARQL Query Language for RDF.

The UniProt Consortium (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, **41**(D1), D43–D47.

Visser, D., Broekstra, J., and Leigh, J. (2014). System documentation for sesame 2. Technical report.