BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–3

Identifiers.org virtual SPARQL endpoint for IRI schemes conversion

Sarala M Wimalaratne [1*], Jerven Bolleman [2], Nick Juty [1], Toshiaki Katayama [3], Michel Dumontier [4], Nicole Redaschi [2], Nicolas Le Novère [5], Henning Hermjakob [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
[5]Babraham Institute, Babraham Research Campus, Cambridge, CB22 3AT, UK

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: On the Semantic Web, in life sciences 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 datasets.

Results: We introduce a virtual SPARQL endpoint provided by Identifiers.org, using its Registry's knowledge about IRI patterns, to automatically translate identifiers from one IRI pattern to another, making cross resources querying easier and more robust. This endpoint is fully SPARQL 1.1 compliant.

Availability: This IRI translating endpoint is available at http://identifiers.org/services/sparql and is free to use for all.

Contact: Sarala M Wimalaratne

1 INTRODUCTION

The use of Semantic Web technologies such as the Resource Description Framework (RDF; Manola and Miller (2004)) for publishing data is becoming popular in the life sciences. For example, Bio2RDF (Belleau *et al.* (2008)) data mashup is one such resource built on the data from various publicly available biological databases translated into RDF and linked together. And, more recently, EMBL-EBI launched the EBI RDF platform (Jupp *et al.* (2014)) in order to exposed some of the datasets it produces in RDF.

Bio2RDF uses normalized URIs (using in its own domain name) to support federated queries cross datasets. While all datasets part of the EBI RDF platform follow a set of rules regarding the URIs they can use, resulting in a consistent set of heterogeneous URIs used across all involved resources. Therefore, those two platforms provide linked datasets which can be queried easily as individual resources. However, the use of two different URI schemes for what are conceptually the same concepts makes it difficult to perform cross platform queries using SPARQL (Prud'hommeaux and Seaborne (2008)), a query language for RDF, which supports complex queries, and can merge data distributed over multiple RDF resources. Thus, rewriting IRIs on the fly is required when querying between resources. String manipulation, supported by SPARQL, may be used for this purpose, but requires users to know in advance the IRI types being used in each resource, making it a cumbersome and inefficient solution.

Identifiers.org (Juty *et al.* (2012)) provides resolvable persistent IRIs used to identify individual records (based on the existing entity identifiers created by the data providers) which are necessary for the Semantic Web vision, and Linked Data initiative. In order to achieve this, it relies on its Registry which contains information about datasets (or data collections) and lists the resources (or physical locations) which distribute them. The recoded information includes the different IRIs used to identify and access the individual records. For example, "plasma membrane" from the Gene Ontology (Ashburner *et al.* (2000)) is available from AmiGO (Carbon *et al.* (2009)), QuickGO (Binns *et al.* (2009)), Bio2RDF, etc.

To resolve cross queries issues, Identifiers.org has developed a virtual SPARQL endpoint to support IRI conversion between datasets. This relies on its existing infrastructure and Registry. Consequently, this provides an automatic system for linking and querying ever growing life science data resources.

2 METHODS

The OpenRDF Sesame SPARQL engine (Visser *et al.* (2014)) was extended to translate specific query patterns to allow IRI translation. Specially, the service has been designed to answer queries requesting IRIs and making use of the predicate `owl:sameAs`. When such queries are detected, the system determines from which data collection the identified entity comes from, then

*to whom correspondence should be addressed

extracts from the Registry the alternative root URIs for this concept and finally generates the complete IRIs and returns them. The main integration point within the SESAME engine is the $getStatements$ method of a class extending $org.openrdf.query.algebra.evaluation.TripleSource$.

The database underlying Identifiers.org's Registry was extended in order to be able to record alternative IRI schemes and the way to convert between them. Additionally, there is a constant curational effort to make sure this additional information is been recorded for all current and future data collections listed in the Registry.

This SPARQL endpoint is directly driven by the database used by the Registry and not by a triple store because it would not have been possible to provide the on demand conversion services otherwise. This explains the "virtual" in its name.

The resulting service is exposed as a virtual SPARQL endpoint at http://identifiers.org/services/sparql. LODEStar (Jupp (2014)), used by the EBI RDF platform, has been deployed to provide an interface for querying it.

3 RESULTS

The Identifiers.org virtual SPARQL endpoint can be used to list all different URI schemes available for a given URI. It can also be used to check whether two IRIs describes the same concept (although it is not a mapping service between entities provided by different datasets). Example queries can be found at the endpoint page URL.

With the Identifiers.org SPARQL service, it is possible to avoid using string manipulation within SPARQL query (figure 1) and does not require users to know which IRI patterns are being used within a given resource.

The query below illustrates how the service can be used to integrate resources using different IRI schemes. In this example, the resources are the EBI RDF platform (which uses IRIs of the form "http://identifiers.org/go/GO:0006915" for GO terms) and Bio2RDF (which uses IRIs of the form "http://bio2rdf.org/go:0006915" for GO terms). The query lists the cross-references to Gene Ontology terms present in the computational model "Edelstein1996 - EPSP ACh event" (BIOMD0000000001) from BioModels Database (Li *et al.* (2010)). It is executed from the BioModels' SPARQL endpoint (http://www.ebi.ac.uk/rdf/services/biomodels/sparql) and connects to Identifiers.org to retrieve the alternative IRIs for the retrieved GO terms. Those are then passed to the Bio2RDF SPARQL endpoint to retrieve the description of 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 Identifiers.org virtual SPARQL endpoint to transform Identifiers.org IRIs into Bio2RDF IRIs.

4 DISCUSSION

There are a number of mapping services aimed to connect identifiers from one database to another. However, these are conceptual mappings, e.g. they relate an identifier for a gene to its related protein products. In comparison, Identifiers.org records all access URLs and IRI based identification schemes used to identify records from a data collection.

Also, while the produced endpoint is SPARQL 1.1 compliant, its virtual nature (not based on an actual triple store) means the queries available are limited to IRIs using the `owl:sameAs` predicate in a basic graph pattern; otherwise no results will be returned. The correct SPARQL solution would be to return the infinite possible IRI translations, but although a technically correct solution, would not be useful at all. 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 provides users the illusion of a valid SPARQL service that can translate all IRI patterns recorded in the Identifier.org Registry as if all these queries where materialised, at a fraction of the cost.

5 CONCLUSION

This IRI conversion service, provided by Identifiers.org as a virtual SPARQL endpoint, enhances the federated query capacities of SPARQL 1.1 to effectively combine information from multiple heterogeneous datasets.

ACKNOWLEDGEMENT

The authors wish to thank the DBCLS RDF Summit and Biohackathon organisers for fostering the initial discussions and development efforts, and Simon Jupp for developing the generic LODEStar user interface used by this endpoint.

Funding: this work received support from the BBSRC (BB/J019305/1), Japanese grants (?), National Institutes of Health grants (1U41HG006104), the Swiss Federal Government through the Federal Office of Education and Science, and EMBL.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1), 25–29.

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. (2014). Lodestar: Linked data browser and sparql endpoint. Technical report.

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

Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M. I., Snoep, J. L., Hucka, M., Le Novère, N., and Laibe, C. (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, **4**, 92.

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

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

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