BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–3

SPARQL-enabled identifier conversion with Identifiers.org

Sarala M Wimalaratne [1,*†], Jerven Bolleman [2,†], Nick Juty [1], Toshiaki Katayama [3], Michel Dumontier [4], Nicole Redaschi [2], Nicolas Le Novère [1,5], 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, 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. The lack of correspondence between identifiers introduces a barrier when executing federated SPARQL queries across life science data.

Results: We introduce a novel SPARQL-based service to enable on-the-fly integration of life science data. This service uses the identifiers patterns defined in the Identifiers.org registry to generate a plurality of identifier variants, which can then be used to match source identifiers with target identifiers. We demonstrate the utility of this identifier integration approach by answering queries across major producers of life science Linked Data.

Availability: The SPARQL-based identifier conversion service is available without restriction at http://identifiers.org/services/sparql.

Contact: Sarala M Wimalaratne

1 INTRODUCTION

Semantic Web technologies such as the Resource Description Framework (RDF; http://www.w3.org/TR/rdf-primer/) offer a powerful paradigm for publishing and exploring life science data through standardization of format and data access. For example, the open source Bio2RDF (Callahan *et al.* (2013)) projects converts dozens of public biological databases and datasets from legacy formats into RDF, and provides a mechanism to explore these as Linked Data. Recently, established bioinformatic organizations such as DBCLS (Kawano *et al.* (2014)), NCBI (https://pubchem.ncbi.nlm.nih.gov/rdf/), NexProt (Chichester *et al.* (2014)), and the EMBL-EBI in collaboration with the UniProt consortium have made some datasets available in RDF (Jupp *et al.* (2014)), thereby significantly extending the network of the Linked Open Data .

Each network uses HTTP-based International Resource Identifiers (IRIs) to identify and link data items. This facilitates querying across network-linked resources, but the lack of a universal identifier system requires mappings across all the different identifiers in use. Identifiers.org (Juty *et al.* (2012)) provides resolvable persistent IRIs used to identify individual records (based on the existing entity identifiers assigned directly by the data providers). Although some linked data providers such as Bio2RDF and the EBI now make their data available with identifiers.org URIs (or mappings to them), this practice is not widely implemented. Therefore, the identifier mismatch makes it difficult to query multiple datasets simultaneously. 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.

To address the issue of identifier heterogeneity, we have developed a SPARQL-based service that generates on-the-fly identifier mappings for registered IRI patterns. Here, we describe our novel method and demonstrate its functionality through service-enabled federated SPARQL queries. This system offers an automatic way to link and query over a rapidly growing number of semantic web friendly life science datasets.

2 METHODS

We implemented a SPARQL-based service that generates a set of variant identifiers based on a provided identifier. This service, implemented using the OpenRDF Sesame SPARQL engine (http://www.openrdf.org/), translates an incoming query pattern of the form ¡subjectIRI¿ `owl:sameAs` ?targetIRI and generates a set of triples with the specific subject, predicate, and the generated target IRI. The service queries the curated Identifiers.org Registry to determine the originating data collection, then obtains alternative IRIs patterns, and finally generates and returns alternative IRIs. While

*To whom correspondence should be addressed

†The first two authors should be regarded as joint first authors

this implementation queries the relational database that underpins the Identifiers.org Registry, others could implement a similar service using the Registry's public web services.

3 RESULTS

The Identifiers.org SPARQL service can be used to list all alternative IRI schemes available for an IRI pattern that is available in the Identifiers.org Registry. Currently, the Identifiers.org Registry contains 531 data collections and 1332 IRI patterns. For supported data collections, this service eliminates the need to know the set of valid IRI patterns in advance and the need to devise elaborate string manipulation operations in a federated SPARQL query (figure 1).

The query below illustrates how the service can be used to query across datasets with different IRI schemes. In this example, we obtain the definitions for Gene Ontology terms that are used to annotate the species in a model from the BioModels database. The service bridges the gap between the Identifiers.org-specified and the Bio2RDF-specified identifier for the Gene Ontology term. This query can be executed from the BioModels' SPARQL endpoint.

```
SELECT DISTINCT ?species ?go_term ?go_description WHERE {
   ?model sbmlrdf:species ?s .
   ?s sbmlrdf:name ?species .
   ?s bqbio:isVersionOf ?go_term
   FILTER (?model = <http://identifiers.org/biomodels.db/BIOMD0000000001>)
   FILTER regex(?go_term,"go","i")
   SERVICE <http://identifiers.org/services/sparql>{
      ?go_term owl:sameAs ?go .
   }
   SERVICE <http://bioportal.bio2rdf.org/sparql>{
      ?go dcterms:description ?go_description .
   }
}LIMIT 10
```

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

Similarly we could write a SPARQL query at the uniprot.org SPARQL endpoint that combines information in UniProt and Wikipathways. SERVICE clause using Identifiers.org SPARQL endpoint to transform the purl.uniprot.org form of IRIs to Identifiers.org which are used in UniProtKB and Wikipathways, respectively (figure 2). Further examples including the queries discussed here are available at http://identifiers.org/documentation.

```
SELECT ?protein ?diseaseComment ?pathway ?pathwayName WHERE {
   ?protein up:annotation/up:disease/rdfs:comment ?diseaseComment
   SERVICE <http://identifiers.org/services/sparql>{
      ?protein owl:sameAs ?otherIRIs .
   }
   SERVICE <http://sparql.wikipathways.org/> {
      ?x dcterms:isPartOf ?pathway; ?z ?otherIRIs .
      ?pathway dc:title ?pathwayName
   }
}
LIMIT 10
```

Fig. 2. Using Identifiers.org SPARQL service to transform UniProt IRIs into Identifiers.org IRIs.

4 DISCUSSION

Leveraging the wealth of biomedical big data for discovery requires simple and effective approaches to tame the challenge of working with heterogeneous, overlapping, and diverse data. Of particular concern is assignment of different identifiers for identical resources as well as for conceptually identical resources. Identifier integration is the subject of much research that focuses either on integrating conceptually identical objects or their relations (van Iersel *et al.* (2010), Wein *et al.* (2012), Smith *et al.* (2007), Chambers *et al.* (2013)). In contrast, our work focuses on the problem of having multiple identifiers for the same database object, which is an emerging issue among semantic web data providers. Our solution is rapid, scalable, and will grow to provide new identifier-based mappings as additional IRI patterns are added to the Identifiers.org Registry.

5 CONCLUSION

This IRI conversion service, provided by Identifiers.org as a SPARQL service, will enable users to focus on asking meaningful questions across biological datasets of interest rather than figuring out how to generate the right identifiers.

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

Callahan, A., Cruz-Toledo, J., Ansell, P., and Dumontier, M. (2013). Bio2rdf release 2: Improved coverage, interoperability and provenance of life science linked data. **7882**, 200–212.

Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S., and Overington, J. P. (2013). UniChem: a unified chemical structure cross-referencing and identifier tracking system. *Journal of cheminformatics*, **5**(1), 3.

Chichester, C., Gaudet, P., Karch, O., Groth, P., Lane, L., Bairoch, A., Mons, B., and Loizou, A. (2014). Querying neXtProt nanopublications and their value for insights on sequence variants and tissue expression. *Web Semantics: Science, Services and Agents on the World Wide Web*.

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.

Kawano, S., Watanabe, T., Mizuguchi, S., Araki, N., Katayama, T., and Yamaguchi, A. (2014). TogoTable: cross-database annotation system using the Resource Description Framework (RDF) data model. *Nucleic acids research*, pages gku403–.

Smith, A. K., Cheung, K.-H., Yip, K. Y., Schultz, M., and Gerstein, M. K. (2007). LinkHub: a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics. *BMC bioinformatics*, **8 Suppl 3**, S5.

van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B. R., and Evelo, C. T. (2010). The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC bioinformatics*, **11**, 5.

Wein, S. P., Ct, R. G., Dumousseau, M., Reisinger, F., Hermjakob, H., and Vizca??no, J. A. (2012). Improvements in the protein identifier cross-reference service. *Nucleic Acids Research*, **40**(W1).