BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

IRI translation for SPARQL using identifiers.org

Jerven Bolleman [1], Camille Laibe [2,*], Toshiaki Katayama [3] Nicole Redaschi [1] and Nick Jutty [2*]

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: On the semantic web and in life science data in particular there are many sources of information (documents available via http) about unique resources. Each of these sources is likely to use their own IRI as an idenfier for conceptualy the same resource or database record. For example http://identifiers.org/uniprot/P05067 and http://purl.uniprot.org/uniprot/P05067 are both identifiers for the same UniProtKB database record a unique resource. This multitude of identifiers introduce a barrier when executing federated SPARQL queries between different databases.

Results: We introduce a virtual SPARQL endpoint that uses identifiers.org 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://identifers.org/mapping and is free to use. The underlying prototype code can be found at github.

Contact: Camille Laibe

1 INTRODUCTION

The use of RDF to model and SPARQL to query data is becoming more common in the life sciences, see for examples Jupp *et al.* (2014) and Katayama *et al.* (2014). As anyone can say anything about anything in the semantic web there are more and more places where one can find information about the "same" thing. For example the bio2rdf (Belleau *et al.* (2008)) data mashup contains more information about a record than the primary sources. For this reason bio2rdf makes this mashup available via http IRIs on the bio2rdf domain, allowing direct http access to records as bio2rdf sees them. Unfortunatly this introduces two identifiers for what is conceptually the same resource. A multitude of identifiers or web addresses for the same resouce is an unfortunate side effect of using the very practical decision to support the http protocol to enable access to information. When doing cross database queries using the SPARQL federalisation abilities this IRI mismatch makes it easy to miss links in the data.

Another example is WikiPathways Kelder *et al.* (2011), which uses UniProtKB accessions to identify proteins, in the RDF representation of WikiPathways they use the identifers.org IRIs as identifiers of resources in databases they reference. While the cannonical source at uniprot.org (The UniProt Consortium (2013)) uses purl.uniprot.org IRIs. This means that a cross database query between the WikiPathways and uniprot sparql endpoint needs to rewrite IRIs on the fly as shown in figure 1.

2 METHODS

We use the openrdf sesame SPARQL engine**?** and extend it to translate specific query patterns to allow query translation. In practical terms this can be done by hand see figure (1) but it is much easier when using the identifiers.org sparqlmapping service as shown in figure 2.

The openrdf SPARQL engine is easy to extend, the main integration point is the $getStatements$ method of a $TripleSource$. All in all we required a just under 1300 lines of java code to make the knowledge contained in the identifiers.org available via http to any other SPARQL service.

3 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 is supposed to identify all sources of information about a single database record as identified by one identifier; independantly of the information provider. Identifiers.org could maintain lists of currently valid IRIs used in all SPARQL endpoints and in RDF datasets, unfortunatly this would be extremely time consuming and always out of sync. This is why the SPARQL service described in this papper 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 suprise end users. If the user does not put in the predcate `owl:sameAs` in a basic graph pattern in the SPARQL query, the query will return no results at all. The SPARQL correct solution

*to whom correspondence should be addressed

```
SELECT ?protein ?diseaseComment ?pathway
WHERE {
  ?protein
    up:annotation/up:disease/rdfs:comment
      ?diseaseComment .
  BIND(iri(
       concat("http://identifiers.org/uniprot/",
            substr(str(?protein), 33))
      ) AS ?wikiProtein)
  SERVICE<http://sparql.wikipathways.org/>
  { ?pathway wp:bdbUniprot ?wikiProtein . }
}
```

Fig. 1. A SPARQL query at the uniprot.org sparql endpoint that combines information in UniProtKB and Wikipathways. Note the BIND clause where the IRI is transformed from the purl.uniprot.org form to the identifiers.org one. (PREFIX declarations ommited for space reasons)

```
SELECT ?protein ?diseaseComment ?pathway
WHERE {
  ?protein
    up:annotation/up:disease/rdfs:comment
      ?diseaseComment .
  SERVICE<http://identifiers.org/sparqlmapping>
  { ?protein owl:sameAs ?wikiProtien . }
  SERVICE<http://sparql.wikipathways.org/>
  { ?pathway wp:bdbUniprot ?wikiProtein . }
}
```

Fig. 2. The same query as in figure 1 but now using a SERVICE clause using identifiers.org to transform the purl.uniprot.org form of the IRIs to identifiers.org form.

would be 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/mapping 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 database as if all these queries where matterialized, at a fraction of the cost.

4 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 the software development.

Funding: . Funding: Japanese grants, EMBL and European Union grants XXXXX. National Institutes of Health grants [1U41HG006104-03] and NCBO [U54-HG004028] 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.

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.

Katayama, T., Wilkinson, M. D., Aoki-Kinoshita, K. F., Kawashima, S., Yamamoto, Y., Yamaguchi, A., Okamoto, S., Kawano, S., Kim, J.-D., Wang, Y., Wu, H., Kano, Y., Ono, H., Bono, H., Kocbek, S., Aerts, J., Akune, Y., Antezana, E., Arakawa, K., Aranda, B., Baran, J., Bolleman, J., Bonnal, R. J. P., Buttigieg, P. L., Campbell, M. P., an Chen, Y., Chiba, H., Cock, P. J. A., Cohen, K. B., Constantin, A., Duck, G., Dumontier, M., Fujisawa, T., Fujiwara, T., Goto, N., Hoehndorf, R., Igarashi, Y., Itaya, H., Ito, M., Iwasaki, W., Kalaš, M., Katoda, T., Kim, T., Kokubu, A., Komiyama, Y., Kotera, M., Laibe, C., Lapp, H., Lütteke, T., Marshall, M. S., Mori, T., Mori, H., Morita, M., Murakami, K., Nakao, M., Narimatsu, H., Nishide, H., Nishimura, Y., Nystrom-Persson, J., Ogishima, S., Okamura, Y., Okuda, S., Oshita, K., Packer, N. H., Prins, P., Ranzinger, R., Rocca-Serra, P., Sansone, S., Sawaki, H., Shin, S.-H., Splendiani, A., Strozzi, F., Tadaka, S., Toukach, P., Uchiyama, I., Umezaki, M., Vos, R., Whetzel, P. L., Yamada, I., Yamasaki, C., Yamashita, R., York, W. S., Zmasek, C. M., Kawamoto, S., and Takagi, T. (2014). Biohackathon series in 2011 and 2012: penetration of ontology and Linked Data in life science domains. *Journal of Biomedical Semantics*, **5**, 5.

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*.

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