

# Linking US Patent Data with DBpedia

Mayank Singhi, Ying Ding  
School of Library and Information Science, Indiana University  
Bloomington, IN 47408, USA  
[msinghi, dingying]@indiana.edu

**Abstract.** DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows user to ask expressive queries against Wikipedia and to interlink other datasets to the Web with DBpedia data. This paper presents a simple approach to link any dataset to DBpedia and thus making the dataset available for anyone's use.

**Keywords:** Semantic Web, Citation Analysis, Linked Open Data (LOD)

## 1 Introduction

DBpedia forms the core of the current Linked Open Data (LOD) bubbles. Interlinking with DBpedia derives further links with other existing LOD datasets [3]. In order to create a link with DBpedia, the first step is to select the parameters through which the desirable DBpedia dataset will be selected. A parameter could be any entity present in the DBpedia space like city, country and name etcetera. Once the parameters have been selected, the second step is to query the DBpedia dataset and to access the result. Since DBpedia dataset is organized as RDF triples, two commonly known options are through Silk and DBpedia SPARQL endpoint.

In this paper, the implementation of linking the US patent database with DBpedia is discussed with location as the common entity along with drawbacks and further improvements. The entity location was selected because of the assumption that the probability of matching location in the DBpedia space will be the highest.

The US Patent data is from the Scholarly Database (SDB) at Indiana University to serve researchers and practitioners interested in the analysis, modeling and visualization of large-scale scholarly datasets [1]. The online interface at <http://sdb.slis.indiana.edu> provides access to four datasets: Medline papers, U.S. Patent and Trademark Office patents (USPTO), National Science Foundation (NSF) funding, and National Institutes of Health (NIH) funding – over 20 million records in total [2], see Figure 1. Users can register for free to cross-search these databases and to download result sets as dump for scientometrics research and science policy practice. SDB supports search across paper, patent, and funding databases: simply select a year range and relevant database(s) and enter search term(s) in creators (author/awardees/inventor), title, abstract, and other keywords. Search results retrieved from different databases can be downloaded as data dump in csv file format.



DBpedia sparql endpoint allows users to get the results in the form of html, spreadsheet, xml, JSON or javascript. We ran the sparql query for every unique location and collected the links. All the locations, patent IDs and the links were then stored in a database table.

The US patent data that we used had 66,153 unique location records. We found 37,995 links for the combination of city and state, 6,063 links for only state and 20,570 links for only country.

A search interface was created to demonstrate the linking between the US patent data and DBpedia (see Fig. 2). The search interface allows the user to search for city, state or country and it retrieves all the records that match the search query. A record includes the city, state, country, patent IDs and the link to DBpedia location page.

### 3 Improvement

The first and foremost improvement should be to clean the data. The major errors in the data are spelling mistakes and missing state/country information. These errors can be fixed with the help of Google maps. To correct the spelling mistakes, a script can be written that queries maps.google.com for every city, state or country. If the spelling is wrong, maps.google.com will display a suggestion in front of “Did you mean” clause. That suggestion can be then appended to the state and/or country and use it to query again. If google maps returns a result, the suggestion can replace the original keyword and stored in the database. Also, using the returned result, we can extract the state or country if it was missing in the original query.

### 4 Conclusion

This paper presents a simple approach and an application to demonstrate linking with DBpedia using Sparql queries. This approach can be extended to create multitude of applications and use all the information available from DBpedia. This paper shows the starting effort to converting and linking SDB datasets to Linked Open data bubbles [4].

### References

1. La Rowe, G., Ambre, S., Burgoon J., Ke, W., Börner, K. (2007). The scholarly database and its utility for scientometrics research. Proceedings of the 11<sup>th</sup> International Conference on Scientometrics on Informetrics (pp 457-462), Spain, June 2007.
2. Scholarly Database on <http://sdb.slis.indiana.edu/>
3. Bizer, C., Heath, T., Berners-Lee, T. Linked Data - The story so far. Journal on Semantic Web and Information Systems. Special Issue on Linked Data.