Linking US Patent Data with DBpedia

Mayank Singhi, Ying Ding
School of Library and Information Science, Indiana University
Bloomington, Indiana 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 users to ask expressive queries against Wikipedia and to interlink other datasets on 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.

Categories and Subject Descriptors
I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks

General Terms
Measurement, Design,

Keywords
semantic web, citation analysis, Linked Open Data (LOD)

1. INTRODUCTION

DBpeida 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 etc. 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 to query 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. This paper reports the preliminary effort to convert the SDB datasets into RDF Linked Open Data. Users can register for free to cross-search these databases and to download result sets as dumps 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 full text (keywords and other text) fields. Search results retrieved from different databases can be downloaded as data dump in csv file format.

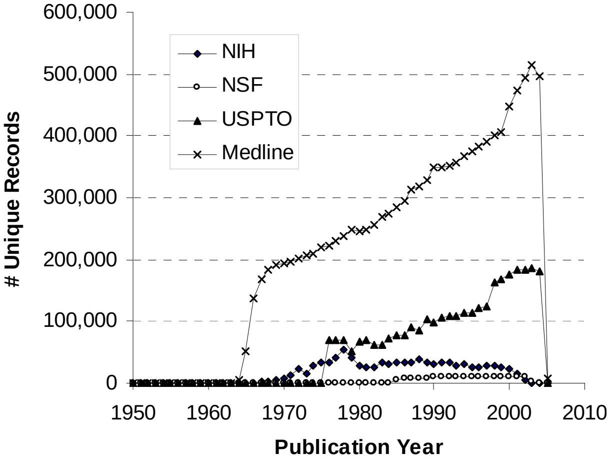


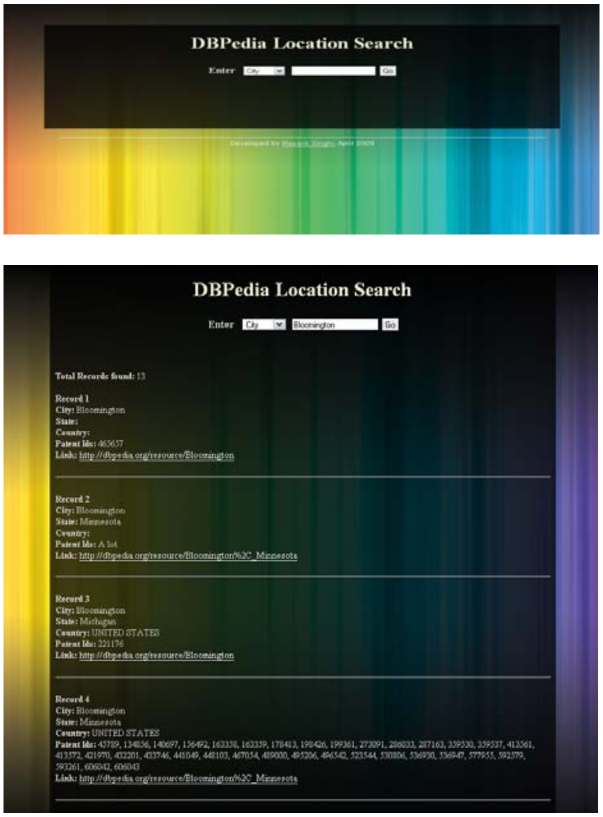
Figure 1: Number of records in SDB per year between 1950 and 2005.

2. IMPLEMENTATION

This application has been implemented using Sparql end point provided by DBpedia (http://dbpedia.org/sparql) which allows us to run sparql queries against DBpedia dataset. We collected the city, state and country information of all the records in US patent database along with their patent ID. The goal was to find the link that describes the location from DBpedia. In order to do that, we first construct a sparql query to look for the city and state. If no link is found, then we look for only the state and if no link is found for the state, we look for the country. Here is an example of a query to look for the city Bloomington and state Indiana -

```
SELECT ?Property ?Value WHERE { { <http://dbpedia.org/resource/Bloomington%2C_Indiana> ?Property ?Value } }
```

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.



Total Records found: 13

Record 1
City: Bloomington
State:
Country:
Patent Ids: 465657
Link: http://dbpedia.org/resource/Bloomington

Record 2
City: Bloomington
State: Minnesota
Country:
Patent Ids: A lot
Link: http://dbpedia.org/resource/Bloomington%2C_Minnesota

Record 3
City: Bloomington
State: Michigan
Country: UNITED STATES
Patent Ids: 221176
Link: http://dbpedia.org/resource/Bloomington

Record 4
City: Bloomington
State: Minnesota
Country: UNITED STATES
Patent Ids: 43789, 134856, 140697, 156492, 163338, 163339, 178413, 198426, 199361, 273091, 286033, 287163, 359530, 359537, 413561, 413572, 421970, 432201, 433746, 441049, 441103, 467054, 489000, 495206, 496542, 523544, 530806, 536930, 536947, 577955, 592579, 593261, 606042, 606043
Link: http://dbpedia.org/resource/Bloomington%2C_Minnesota

Figure 2. Screenshot of the search of linked US Patent data and DBpedia

The US patent data that we used had 66,153 unique location records. This data was not very clean. It had errors like spelling mistakes and missing state/country. The data was not rectified for this application. However, we have presented an approach to clean this data in the improvements section. Table 1 shows the number of records for which what link was found.

Total Records	66, 153
City & State	37,995
Only State	6,063
Only Country	20,570

Table 1. The number of established links between US Patent and DBpedia

A search interface was created to demonstrate the linking between the US patent data and DBpedia (see Figure 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. IMPROVENMENT

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 be 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. In *Proceedings of the 11th International Conference on Scientometrics and Informetrics* (pp. 457-462), Spain, June, 2007.
2. http://sdb.slis.indiana.edu/
3. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann: DBpedia - A Crystallization Point for the Web of Data. To appear in: Journal of Web Semantics (JWS), Special Issue on the Web of Data.
4. Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data - The Story So Far. To appear in: Journal on Semantic Web and Information Systems (IJSWIS), Special Issue on Linked Data.