Paper: S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *DBpedia: A Nucleus for a Web of Open Data*, In Proc. The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC 2007 + ASWC 2007), Nov. 2007.

Reviewer Name & ID: Nabiha Asghar, 20448761
Email: nasghar@uwaterloo.ca **Date Reviewed:** 21st October 2014

The DBpedia project is the cornerstone of a global endeavor to create a well-linked online web of structured information, to facilitate sophisticated querying over multi-domain data. DBpedia extracts information from Wikipedia, which is one of the most popular online encyclopedias but has very limited query capabilities due to its collaborative and inconsistency-prone nature. This extracted content is republished and made widely available by DBpedia, through a variety of interfaces, in a form that is semantically friendly and can be integrated with other data sources on the Web. This paper is an excellent introduction to this ground breaking project. It describes three integral components of DBpedia: information extraction and access mechanisms, data inter-linkage with other datasets and user interfaces.

The information extraction framework is the technical core of DBpedia. Auer *et al.* explain with a nice example how pattern matching techniques are used to extract content from the in-built syntactic structures on Wikipedia pages, such as infobox templates, lists of objects and web links, and how this content is converted to RDF (Resource Development Framework) triples. Additional RDF triples are created by extracting relationships directly from Wikipedia's relational database dumps. The resulting dataset consists of over a 100 million RDF triples and contains information about almost 2 million resources, each identified by a unique URI. The authors have altruistically made this dataset available for free download, and it is also accessible to public as Linked Data, that is, URIs accessible through HTTP on conventional web browsers as well as Semantic Web agents. A SPARQL interface is also provided to support manual and automated querying. The dataset is linked to other datasets on the web via RDF links, which can be navigated through Semantic Web browsers. This linkage allows users to make sophisticated queries and retrieve information from beyond DBpedia's own knowledge base. The web based user interface of DBpedia allows SPARQL tables to be easily added to websites. A neat feature of these tables is that they are automatically updated if the corresponding information on Wikipedia changes. Keyword based search of the data is supported, which explores deep relations between entities and returns ranked results. A query based user interface is also available, which provides query guidance to users unfamiliar with a desired item's identifiers or properties in DBpedia's knowledge base.

Overall, this paper's presentation of DBpedia's important building blocks is quite coherent. It highlights the significance of this research very effectively, and rightly so, because DBpedia has expanded tremendously since this publication; it is central to a fast-growing Linked Open Data cloud, and many tools and applications have been built around it. The paper is mostly self-contained, albeit it assumes basic knowledge of RDF and Semantic Web. The discourse uses easy terminology and has a natural flow. The main contributions are listed succinctly in the beginning, and each of them is subsequently explained in its own dedicated section. Also, important ideas have been explained through very helpful examples. Due to these examples, I was able to experiment with the accessibility features of DBpedia and found them to be comprehensive and user friendly. However, the paper fails to answer some important questions. For example, the authors imply that the free text in the articles is also parsed for information extraction, but they do not state this clearly, nor do they explain exactly how any semantic relationships are extracted from it. This is slightly unsettling, given that the free text of an article can contain crucial information about different entities. Moreover, nothing has been said about the completeness of DBpedia's dataset. Is any information lost when full text articles are represented as a set of RDF triples? If so, how can this loss be quantified and minimized? If not, is there a way to prove its completeness? Furthermore, Wikipedia has images that may contain important information about entities. How is this information extracted and captured in RDF?

As well, the authors have not discussed the performance evaluation of DBpedia's user interfaces in detail, which is another crucial omission because the underlying dataset is very huge. The paper could also benefit from some concrete comparison between DBpedia and Freebase, another Wikipedia based database. Though the illustrations and figures are beneficial, they have taken up too much space, some of which could be put to better use by providing answers to the questions mentioned above.