Literature review

Accompanying with the rapid development of Internet, information explosion brings both challenges and opportunities for traditional web structure. Although individuals and organizations publish numerous amount of data on web via protocols such as TCP/IP and HTML, very few of them could be understood by computers, since that the main content of traditional web technologies is designed for human to read rather than machines. On the current web the scale of content for computer to read and understand is very limited (A Semantic Web Primer). As a consequence, it is difficult to integrate and reuse data on the web if they are difficult to understand by computers (Semantic Web Technology Systems). Besides, information explosion not only brings great amount of data, but also highly increases the complexity of data. With the requirement for more agile approaches to handle links and mash-up data in complex processes, conventional approaches of data integration would collapse in most situations (Challenges and Opportunities). In order to deal with these problems, the concept of Semantic Web is proposed and added to the current web layer to make the machine to understand web document (Berners-Lee, Hendler, & Lassila, 2001)

In order to support Semantic Web technology, the term “Linked Data” is introduced, referring to the method by which data could be described via Uniform Resource Identifier and Resource Description Framework, so that these data could be exchanged, displayed, linked and published (From Bibliographic Records to Data). In order to structure and standardize data so that it could be understood and utilized by machines, rules and standards are proposed such as Resource Description Framework (RDF) and Uniform Resource Identifier (URI). According to the rules raised by Berners-Lee in 2006, HTTP (Hypertext Transfer Protocol) URI strings is used to represent data so that people can locate data on internet wherever they are and different data could be linked together even if they are on the different servers, in this way data could be “Linked”. Then, standards such as RDF are utilized to structure data, to describe the relationship and other metadata, by means of these methods data could be understood by machines. Differs from conventional web structure whose primary units are HTML (HyperText Markup Language) docu­ments connected by hyperlinks, Linked Data depends on documents containing data in RDF format (Klyne and Carroll, 2004---Linked data, the story so far). Once data is machine-readable, which means it is explicitly defined, the link between local data set and external data set could be established (Linked Data-The Story So Far).

Although Linked Data is still relatively new compared to the history of web, the practises of Linked Data have extended Web in a global scale, connecting different data from diverse fields such as people, companies, books, scientific publications, films, music, television and radio programmes, genes, proteins, drugs and clinical trials, online communities, statistical and scientific data, and reviews (Linked Open Data, the story so far). One of the first Linked Data project which start to realize the web of Linked Data was the Linked Open Data Project (LOD-the-essentials). Depedia is a project aiming to extract data from Wikipedia and turning these data into the form of Linked Data so that enormous amount of data in Wikipedia could be readable and available for machines. It is described as one of the most famous components of the decentralized Linked Data effort (["Sir Tim Berners-Lee Talks with Talis about the Semantic Web"](https://web.archive.org/web/20130510134842/http:/talis-podcasts.s3.amazonaws.com/twt20080207_TimBL.html). Talis. 7 February 2008. Archived from [the original](http://talis-podcasts.s3.amazonaws.com/twt20080207_TimBL.html) on 10 May 2013). DBpedia offers tunnels for different application to get access to the data in Wikipedia, and many applications of Linked Data are based on it. For example, the BBC Music Beta project integrates not only database on its own servers, but also the information contained on Wikipedia, and it accomplished this aim by capturing data from DBpedia and develop its own data schema with data. Also, millions of geographical locations worldwide are extracted from Wikipedia and “geogame” by DBpedia (LOD-the-essential); in medical area, current Linked Open Data and technologies are used to automate the process of biomedical ontology generation (Linked open data-based framework for automatic biomedical ontology generation); search engine which crawls data from web via traversing RDF links performs more efficiently over large amount of data than traditional one (Linked Data: Evolving the Web into a Global Data Space). Examples of applications of Linked Data are now distribute widely in many areas such as smart cities, academic communities, social media and healthcare.

According to the description of Tom Heath and Christian Bizer (2011), to develop an application using Linked Data technology should follow 6 steps: accessing the web of data, vocabulary of mapping, identity resolution, provenance tracking, data quality assessment and finally, using data in the application context. In short, these 6 steps could be concluded as two parts: extract raw data from the web and integrate data according to its own data schema. Under current state of web, not most of the publishers would publish their web pages with Linked Data, and it would be more efficient to utilize a dataset of Linked Data than integrate data from all over the Internet. Websites such as Wikipedia, government official website, contain enormous amount of unstructured data, which is highly valuable for academic purpose or commercial purpose.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), 28–37. doi:10.1038/scientificamerican0501-34 PMID:11341160

Abstract:

The purpose of the Semantic Web is to enhance the comprehension of machines towards data within web pages, so that enormous valuable could be fully utilized and reused. Linked Data was introduced aiming to realize the Semantic Web and applications based on Linked Open Data has been adopted in many areas. However, there is still great potential of Linked Data technology. In this article, the author present background of Linked Data and applications based on Linked Data. Background information, importance of Linked Data and current Linked Data application development are introduced in the literature review. Afterwards, the author will propose an idea of the combination of DBpedia and World Bank Open Data from two aspects, which are justification and feasibility.