Chapter 71
SPedia:
A Central Hub for the Linked Open Data of Scientific Publications

Muhammad Ahtisham Aslam
King Abdulaziz University, Saudi Arabia

Naif Radi Aljohani
King Abdulaziz University, Saudi Arabia

ABSTRACT

Producing the Linked Open Data (LOD) is getting potential to publish high-quality interlinked data. Publishing such data facilitates intelligent searching from the Web of data. In the context of scientific publications, data about millions of scientific documents published by hundreds and thousands of publishers is in silence as it is not published as open data and ultimately is not linked to other datasets. In this paper the authors present SPedia: a semantically enriched knowledge base of data about scientific documents. SPedia knowledge base provides information on more than nine million scientific documents, consisting of more than three hundred million RDF triples. These extracted datasets, allow users to put sophisticated queries by employing semantic Web techniques instead of relying on keyword-based searches. This paper also shows the quality of extracted data by performing sample queries through SPedia SPARQL Endpoint and analyzing results. Finally, the authors describe that how SPedia can serve as central hub for the cloud of LOD of scientific publications.

1. INTRODUCTION

The growth of domains of knowledge in our data intensive age depends particularly on the efficiency and sophistication of the processes of data production, distribution and consumption, among the corresponding community (Andriole, 2010). Specific to scientific domain, there is huge amount of data about vast number of scientific documents such as articles, books, reference works, being produced by academia and industry. Unfortunately, these documents are being published as bounded group of publisher specific resources resulting in lake of collaboration and interconnected resources for knowledge

DOI: 10.4018/978-1-5225-5191-1.ch071

sharing. There is an urgent need to publish and share research publications data. This can enable other researchers to interconnect their data to the one that already published. Ultimately this can be used by researchers and practitioners to share their research (Kauppinen de Espindola, 2011) for better collaboration and future analysis.

The set of best practices for publishing and interconnecting distributed data has termed as Linked Open Data (LOD). These best practices are being used by increasing number of data providers (Bizer, Heath Berners-Lee, 2009; Villazón Terrazas, Vilches, Corcho Gómez-Pérez, 2011) such as government (Lebo et al., 2011), education (Lnenicka, 2015), news (Suárez Jiménez-Guarín, 2014), health (Bukhari Baker, 2013), geography (Correndo, Salvadores, Yang, Gibbins Shadbolt, 2010) and by researchers to extract semantically enriched data from different public resources such as Wikis, as community effort to publish LOD (Erxleben, Gu¨nther, Krötzsch, Mendez Vrandecic, 2014; Vrandečić Krötzsch, 2014; Lehmann et al., 2015). When it comes to the scientific publications data, very little work has been conducted (e.g. Springer., 2015, Hakimpo-ur, Arpinar Sheth, 2007) to publish LOD of scientific documents. It is also acknowledged (Blmel, Dietze, Heller, Jschke Mehlberg, 2014) that in scientific research, structured data is limited and exposed based on proprietary or less-established schemas resulting in unholistic and inconsistent view on research information.

As a step towards publishing linked open data of scientific publications, in this paper we present SPedia: a semantically enriched knowledge base of scientific publications data. SPedia knowledge base adds three hundred million RDF triples to the Web of data which provide information on about nine million scientific documents published in twenty-four disciplines and four different languages. SPedia knowledge base is populated from the scientific publications data of documents published by Springer and we used SpingerLink[1] as source of data. SPedia datasets are available for download from project Web site[2] and can be used to link other open datasets published in the LOD cloud. In SPedia project we have also established a SPARQL Endpoint that can be used to put semantically enriched queries to SPeida for the intelligent query answering purposes rather than to rely on keyword-based searches on unlinked distributed data.

The work presented in this paper makes the following contributions:

- The extraction of structured information from over 9 million documents available on SpringerLink. The resulting datasets contain information about scientific documents from 24 disciplines (e.g., computer science, engineering, social sciences, etc.) and 6 types of documents (e.g., books, chapters, journals, articles, reference works, and reference work entries), written in four different languages;
- Production of semantically enriched datasets as RDF triples which were extracted from the detailed information (e.g., abstract, DOI, ISBN, pdf link, author, organization, etc.) of scientific documents;
- Extraction and triplification of relational information between various document types (e.g., relationships between book and book chapter, journal and journal article, etc.);
- Customized approach and algorithm for crawling, parsing and extraction of useful information and triples generation;
- Development of a SPARQL Endpoint that can be used to put semantically enriched queries against SPedia datasets that consist of more than three hundred million RDF triples.