LOD-a-lot: A Single-File Enabler for Data Science

Wouter Beek w.g.j.beek@vu.nl
Dept. of Computer Science, VU University Amsterdam, NL

Javier D. Fernández jfernand@wu.ac.at
Vienna University of Economics and Business

Ruben Verborgh ruben.verborgh@ugent.be
Ghent University – imec

Keywords

Data Science Data Access Linked Open Data

Abstract

Many data scientists make use of Linked Open Data (LOD) as a huge interconnected knowledge base represented in RDF. However, the distributed nature of the information and the lack of a scalable approach to manage and consume such Big Semantic Data makes it difficult and expensive to conduct large-scale studies. As a consequence, most scientists restrict their analyses to one or two datasets (often DBpedia) that contain -- at most -- hundreds of millions of RDF triples. LOD-a-lot is a dataset that integrates a large portion (over 28 billion triples) of the LOD Cloud into a single ready-to-consume file that can be easily downloaded, shared and queried, locally or online, with a small memory footprint. This paper shows there exists a~wide collection of Data Science use cases that can be performed over such a LOD-a-lot file. For these use cases LOD-a-lot significantly reduces the cost and complexity of conducting Data Science.