Squirrel – Crawling RDF Knowledge Graphs on the Web

Michael Röder[1,2] [0000−0002−8609−8277], Geraldo de Souza Jr[1] [0000−0003−3180−0177], and Axel-Cyrille Ngonga Ngomo[1,2] [0000−0001−7112−3516]

[1] DICE group, Department of Computer Science, Paderborn University
michael.roeder|gsjunior|axel.ngonga@upb.de
[2] Institute for Applied Informatics, Leipzig, Germany

Abstract. The increasing number of applications relying on knowledge graphs from the Web leads to a heightened need for crawlers to gather such data. Only a limited number of these frameworks are available, and they often come with severe limitations on the type of data they are able to crawl. Hence, they are not suited to certain scenarios of practical relevance. We address this drawback by presenting Squirrel, an open-source distributed crawler for the RDF knowledge graphs on the Web, which supports a wide range of RDF serializations and additional structured and semi-structured data formats. Squirrel is being used in the extension of national data portals in Germany and is available at https://github.com/dice-group/squirrel under a permissive open license.

Keywords: Linked Data· Crawler · Open Data

1 Reproducibility

Please find the necessary information for the submission in the following table.

Type	Value
Abstract	Printed above
Public repository	https://github.com/dice-group/squirrel
README file	Can be found in the supplementary material
Data	The data is automatically generated. The generation is described in [1]
License	LGPL 3.0 (see https://github.com/dice-group/Squirrel/blob/master/LICENSE)

The resource paper, which describes the crawler in detail and which has been accepted at ISWC, can be found at https://papers.dice-research.org/2020/ISWC_Squirrel/public.pdf.

2 Acknowledgments

This work has been supported by the BMVI (Bundesministerium für Verkehr und digitale Infrastruktur) projects LIMBO (GA no. 19F2029C) and OPAL (GA no. 19F2028A).

References

1. Röder, M., de Souza, G., Kuchelev, D., Desouki, A.A., Ngomo, A.C.N.: Orca: a benchmark for data web crawlers (2019), https://arxiv.org/abs/1912.08026