The European Data Portal: Scalable Harvesting and Management of Linked Open Data

Fabian Kirstein[1,2], Simon Dutkowski[1], Benjamin Dittwald[1], and Manfred Hauswirth[1,2,3]

[1] Fraunhofer FOKUS, Berlin, Germany
[2] Weizenbaum Institute for the Networked Society, Berlin, Germany
[3] TU Berlin, Open Distributed Systems, Berlin, Germany
{firstname.lastname}@fokus.fraunhofer.de

Abstract. The European Data Portal fosters the adoption and distribution of Linked Open Data by offering more than 130 million RDF triples. It applies DCAT-AP, the RDF vocabulary for public sector datasets in Europe. Many Open Data solutions do not satisfactorily support this metadata specification. To address this problem, we designed and implemented a novel platform for harvesting and managing native DCAT-AP-compliant RDF. Our approach uses a triplestore as main database and applies state-of-the-art development and deployment patterns to ensure performance and scalability.

Keywords: DCAT-AP · Open Data · Scalability .

1 The European Data Portal

The European Data Portal (EDP)[4] is the one-stop-shop for all metadata of Open Data published by public authorities of the EU. It leverages the RDF vocabulary DCAT Application profile for data portals in Europe (DCAT-AP) to improve interoperability and enable a cross-portal search [2]. The EDP was launched in Nov. 2015, based on the Open Source solution CKAN[5] for building the central data catalog component. CKAN relies on a relational database, requiring expensive transformation concepts for RDF [1]. Therefore, we developed a novel native DCAT-AP platform for harvesting and managing the metadata of the EDP. It was launched successfully in May 2019 for production use and uses the Virtuoso triplestore[6] as primary database, the Elasticsearch search server[7], the reactive Java framework Vert.x[8] and employs a scalable microservice approach.
Our solution has three main components: *The Harvester*, *the Registry* and *the Quality Service*. The Harvester periodically retrieves the metadata from 80 source portals, supporting a variety of data formats (CKAN-API, OAI-PMH, uData,

[4] https://www.europeandataportal.eu/
[5] https://ckan.org/
[6] https://virtuoso.openlinksw.com/
[7] https://www.elastic.co/products/elasticsearch
[8] https://vertx.io/

RDF and SPARQL). If necessary, the acquired metadata is transformed into DCAT-AP-compliant RDF. The Registry supports the resource-based management of the major DCAT-AP entities via a RESTful API (Catalogs, Datasets and Distributions). It receives data from the Harvester, preprocesses it and stores the RDF in the triplestore. The preprocessing includes generation of consistent URIs, validation based on SHACL and creation of a full-text search index. The results of the validation are represented by the Data Quality Vocabulary (DQV)[9]. The Quality Service fetches the metadata and validation results, performs availability checks for external links and creates detailed quality reports. Fig. 1 illustrates the highlevel architecture, including the most essential services.

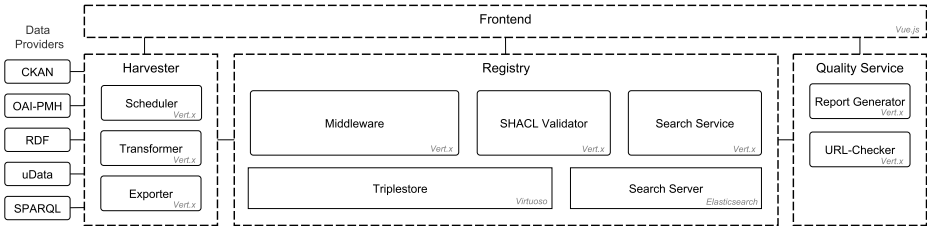

Fig. 1. High-Level Overview of the EDP Architecture

The system makes extensive use of an asynchronous and reactive programming paradigm, applies the container technology Docker[10] for deployment and supports container-orchestration, like Kubernetes[11]. Therefore, we were able to decrease the hardware workload and increase the overall performance, in comparison to the CKAN-based solution. A full harvesting run was accelerated by a factor 30, where the major blocking factor is the limited response time of the data sources. We have shown that our solution and Linked Data can be applied successfully in a real world scenario with significant advantages over established Open Data solutions, e.g., a higher scalability, a more flexible data schema and rich queries with SPARQL. Since the European Commission (EC) establishes DCAT-AP as a standard, our work acts as an enabler for its broad adoption. Open Data from public administrations is an essential source for Linked Open Data. Hence, our solution will have a direct impact on its dissemination. The entire platform is available as Open Source.[12]

References

1. Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., Hauswirth, M.: Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP. In: EGOV2019 – Joint Conference EGOV-CeDEM-EPART 2019 (2019)
2. Dragan, A.: DCAT Application Profile for data portals in Europe (2019), https://joinup.ec.europa.eu/sites/default/files/distribution/access_url/2019-05/e3f7bcdf-eaad-4741-9bf6-dc61327f4eea/DCAT_AP_1.2.1.pdf

[9] https://www.w3.org/TR/vocab-dqv/
[10] https://www.docker.com/
[11] https://kubernetes.io/
[12] https://gitlab.com/european-data-portal