SUMMARY

The World Wide Web is a huge collection of interlinked information. For us, humans, this information is readily accessible in formats we like, such as news articles rendered as text, audio broadcasts, and videos via services such as YouTube. Where human beings can understand these formats, software agents often cannot.

A complementary approach to the World Wide Web is that of Linked Data, where information is represented in a machine readable format. Linked Data uses the same underpinnings as the World Wide Web: both use the Hypertext Transfer Protocol (HTTP) to access and retrieve information. In Linked Data, URLs and string denote ’things’, called resources. These resources can be anything, such as geographical locations, documents, abstract concepts like ‘Democracy‘, or numbers and strings.

The Linked Data architecture enables consumption without a-priori knowledge of the schema and content, and enables publishing without knowing how a dataset is going to be used. This unknown (re)use is an intrinsic positive quality of Linked Data, but it presents problems as well for both consumers and publishers of Linked Data. Below, we discuss the five problems we identified together with the corresponding contributions of this thesis.

impractical linked data re-use Publishing Linked Data as static files seems straightforward, but even this method can be difficult in practice: many Linked Datasets still do not adhere to standards and best practices.

We developed a centralized service called the LOD Laundromat, that re-publishes a clean version of as many Linked Open Datasets as possible, providing a wealth of uniform clean data that can be used with little effort.

queryable linked data is expensive to host The de-facto standard for hosting queryable Linked Data is the SPARQL endpoint. Its flexibility and rich querying language offers advantages for Linked Data providers, but comes at a cost: these SPARQL triple-stores are expensive to host. We presented two solutions to this problem.

First, we presented a sampling method called SampLD, that reduces the dataset size, and thus reduces the hardware costs for hosting a SPARQL endpoint.

A second, orthogonal, solution is to reduce query language complexity instead of reducing the completeness of query results. We did so by combining the LOD Laundromat with a Triple Pattern Fragment API that only supports simple triple pattern queries. This re-

sulted in a low cost Linked Data API that consumes a fraction in memory and processing power compared to SPARQL triple-stores.

FORMULATING SPARQL QUERIES IS DIFFICULT The complexity and expressiveness of SPARQL makes it an unforgiving and difficult query language. But the state-of-the-art in query editors does not provide the tooling and features that web developers are accustomed to. We improved this state-of-the-art by developing YASGUI, a SPARQL query editor accessible from the browser, that has many features known to web developers such as syntax highlighting, syntax checking, and auto-completion functionality.

PROBLEMATIC ACCESS TO LINKED (META-)DATA The distributed nature of Linked Data and the absence structural descriptions of datasets makes locating and accessing Linked Datasets difficult. A centralized solution such as the LOD Laundromat does not directly solve this issue, as finding datasets according to structural criteria still requires manual processing and exploration.

Therefore we extended the LOD Laundromat with a structural Meta-Dataset that includes an IRI and namespace to dataset index, dataset characteristics, and provenance information about the performed processing steps.

VARIETY OF LINKED DATA RESEARCH Linked Data research is suffering from the unavailability of resources and tools to study Linked Data and its use at large.

The first approach we presented focused on Linked Data *use* –a research area that is strongly restricted by the limited number of available query logs. We enable tracking Linked Data usage from the *client* side using the YASGUI query editor as a measuring device, and show how the queries collected by YASGUI enable us to investigate usage patterns that are difficult to measure otherwise.

The second approach focused on increasing the variety from a *data-centric* perspective. We showed that existing research only evaluates on a handful of datasets, and we presented an alternative approach for running experiments on a much broader scale using the LOD Laundromat, the corresponding Meta-Dataset, and the exposed LOD Laundromat Triple Pattern Fragments API. By re-evaluating three recent publications, we this new evaluation paradigm brings up interesting research questions as to how algorithmic performance relates to (structural) properties of the data.

This thesis presents a number of advancements for building Linked Data based services. The presented solutions are targeted at both consumers and publishers of Linked Data, and are a step towards a web of Linked Data that is more accessible and technically scalable.