KartoGraphI: Drawing a Map of Linked Data

Pierre Maillot, Olivier Corby, Catherine Faron, Fabien Gandon, Franck Michel

KartoGraphI: Drawing a Map of Linked Data

Pierre Maillot[0000−0002−9814−439X],
Olivier Corby[0000−0001−6610−0969], Catherine Faron[0000−0001−5959−5561],
Fabien Gandon[0000−0003−0543−1232], and Franck Michel[0000−0001−9064−0463]

University Cote d'Azur, Inria, CNRS, I3S, France
first.last@inria.fr

Abstract. A large number of semantic Web knowledge bases have been developed and published on the Web. To help the user identify the knowledge bases relevant for a given problem, and estimate their usability, we propose a declarative indexing framework and an associated visualization Web application, KartoGraphI. It provides an overview of important characteristics for more than 400 knowledge bases including, for instance, dataset location, SPARQL compatibility level, shared vocabularies, etc.

1 Introduction

In recent years, a large number of semantic Web knowledge bases (KB) have been developed and published on the Web. The reuse and joint exploitation of multiple KBs requires them to comply with a number of good practices. To help the user identify a set of relevant and usable KB that answers her need, we propose the *KartoGraphI* Web application[1] that provides visualizations of certain characteristics of KBs. It relies on a (meta)dataset describing KBs available on the Web, automatically generated by the *IndeGx* framework by querying KBs with SPARQL.

This paper presents a demonstration of KartoGraphI together with an illustration of the main features of IndeGx. It is organized as follows: Section 2 presents the IndeGx framework and generated dataset. Section 3 presents the KartoGraphI Web application. Section 4 presents the proposed demonstration and concludes.

2 IndeGx: Description of Semantic Web KB

IndeGx is a framework designed to index public KBs that are available online through a SPARQL endpoint. The indexing process uses SPARQL queries to either extract the available metadata from a KB or to generate as much metadata as the endpoint allows it. The RDF data generated by IndeGx is expressed using a combination of well-established description vocabularies such as SPARQL Service Description, VoID and DCAT. The generated metadata not only describes KBs but also conveys an estimation of certain quality criteria. These metadata can be used either by humans e.g. to select suitable sources, or agents to automate processes such as query optimization or rewriting.

IndeGx processes publicly available KBs whose endpoints are listed on the LOD Cloud, Wikidata, SPARQLES [3], Yummy Data [4] and Linked Wiki. At the time of

[1] KartoGraphI: *http://prod-dekalog.inria.fr/*

writing, IndeGx has processed and indexed 553 KBs. The generation of a KB description depends on the capabilities of its endpoint. In particular, complex queries, such as those using counting and string-based operators on large quantities of triples or disjunctions, may fail due to the limits in time and resources of the endpoint.

IndeGx can be extended with new KB description features as long as they can be extracted with SPARQL queries. Those new features will be used in future indexations. The operations made by IndeGx to extract metadata are described in RDF in a GitHub repository using the EARL and the Manifest vocabularies. This structure makes it possible to add operations used in other similar approaches. IndeGx already uses the operations described in the approaches SPORTAL [2] and SPARQLES.

Asserted Dataset Descriptions. Provenance information concerns, among others, a dataset's authors, contributors, editors, license, source datasets, generation method, date of creation and modification. Those provenance metadata are important for transparency of and trust in the KB. They are generally asserted by the KB creators as part of a VoID description, but are difficult to discover when they are not given. IndeGx extracts the available provenance information of a KB by looking in its dataset for metadata descriptions made using VoID, DCAT or SPARQL-SD vocabularies. We separate the most basic provenance metadata into 4 categories: the authorship, the licensing, the time, and the source. The results show that less than 1% of KBs with a public SPARQL endpoint contain such metadata.

Endpoint Capabilities. The content of a KB is only as reachable as its endpoint is available. Approaches such as SPARQLES [3] estimate the performance of SPARQL endpoints using SPARQL queries. IndeGx reuses SPARQLES queries and extends them to survey the implementation of the features of the SPARQL language by endpoints.

Data Quality Measures. Several criteria have been identified as indicators of the quality of a KBs, such as those listed in [1,4]. IndeGx evaluates KBs according to several of those quality criteria using SPARQL queries.

Traceability of the indexing process. IndeGx generates traceability metadata for every operation done on every dataset. The metadata contain information such as the start and end time of each operation, the query submitted and the error returned if it so happens. The generated traces support the study of the average performance of endpoints and of the responses of each endpoint to determine some of their characteristics.

3 KartoGraphI: Drawing a Map from Semantic Web KBs

KartoGraphI is a Web application visualizing the data generated by IndeGx. It draws various visualizations and statistics grouped along several view points described below.

Used/Shared Vocabularies. The (re)use of well-known vocabularies in a dataset facilitates its re-usability by others and is one of the core principles of Linked Data. IndeGx extracts the list of vocabularies from the namespaces of classes and properties a dataset uses.

KartoGraphI first generates a visualization of the endpoints linked to all the vocabularies, including the vocabularies used by no other KB (Figure 1). Secondly, a refined

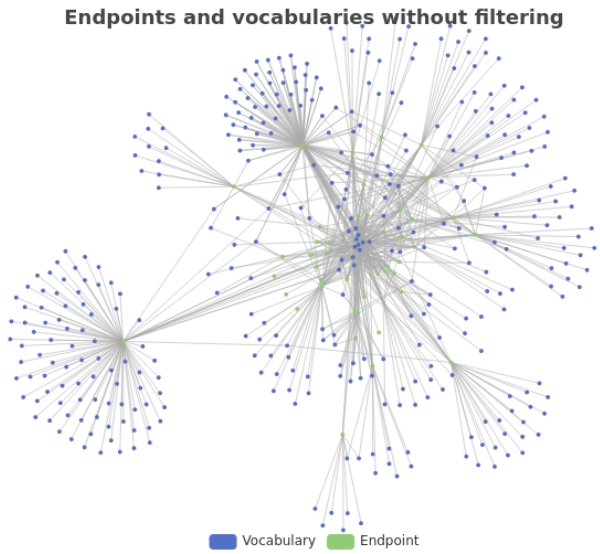

Fig. 1. Graph of the endpoints listed on LOD Cloud website and their vocabularies. The vocabularies are extracted by IndeGx from the namespaces of classes and properties of the datasets.

version of this visualization focuses on well-known vocabularies by showing only a subset of 3̇487 namespaces extracted from Linked Open Vocabularies (LOV) and prefix.cc. A major advantage of this refined visualization is to filter out any unknown vocabulary that may result of a typographical error or the non-respect of naming conventions. However, it can also filter out new or very specialized vocabularies that are not listed on those sites.

KartoGraphI also offers a third visualization that connects the endpoints to the keywords associated to their well known vocabularies. This association of endpoints and keywords can be used to identify the domains a KB pertains to.

Dataset Population. VoID allows to give statistics about the triples and instances present in a dataset. Such metadata give an idea of the scale of a dataset and the most advanced statistics can be used to optimize queries. After trying to retrieve such information from the metadata given in a KB, IndeGx uses SPARQL queries to update the retrieved metadata and as much as possible extract all the possible statistics that can be expressed with VoID. They include statistics such as the number of triples using a certain property with an instance of a certain class as an object. Such statistics are extracted using CPU- and/or memory-intensive queries that only endpoints backed by large hardware resources can answer.

Endpoint Capabilities. The results displayed by KartoGraphI (Figure 2), show that two-thirds of the endpoints tested by IndeGx support at least 80% of all tested SPARQL features. The endpoints from the remaining third either returned errors or did not answer any SPARQL query.

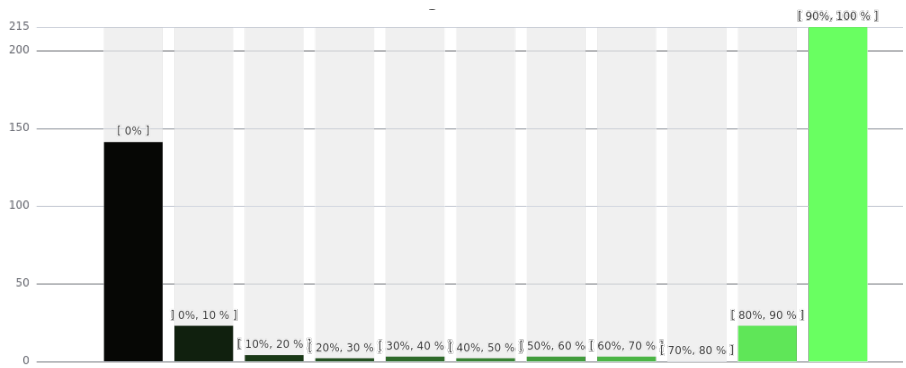

Fig. 2. Bar chart of the distribution of the endpoints according to their coverage of SPARQL features.

Data Quality Measures. The quality of the data of a KB can be measured according to different criteria. Among them, one of the structural quality measures depends on the respect of community best practices such as avoiding the usage of blank nodes and RDF data structures such as lists, sequences, and bags. It is to be noted that vocabularies such as OWL and SHACL use blank nodes and data structures to define vocabularies. Due to that, this measure may not be adapted as it is for datasets such as ontology repositories. This kind of criteria appears to be well complied with among the KBs processed. The proportion of resources that are not blank nodes in a dataset is close to 100%. The proportion of triples that are not part of the definition of RDF data structures is 76.2%. Other quality measures include the interlinking of a KB with others, through the use of shared vocabularies. They also include the readability of the KB by humans, based on the presence of labels and short readable URIs. The interlinking of KBs is generally good among the KBs processed, with on average 75% of the vocabularies used in a KB being listed on vocabulary portals. On the other side, the readability of KBs is not good in general according to our measures: an average 42% of the resources have no label, and an average 49% of them do not have short URIs.

Geolocation. The geolocation of endpoints is often overlooked as metadata of a KB. Using only SPARQL queries, IndeGx could only get endpoint timezones as an indicator of their location. To work around this lack, KartoGraphI uses external APIs to determine the location of endpoints from their URL. The resulting map, shown in Figure 3, illustrates that, although the Linked Data is an international effort, it is over-represented in Europe whereas there is an "empty diagonal" from South-West to North-East, containing no publicly available KBs. It is however possible that KBs hosted by countries in this zone be listed in resources currently not considered by IndeGx.

4 Proposed Demonstration

The demo will be as follow: Attendees will be guided through the different visualizations of the KartoGraphI Web application. A commentary of the notable features observed in the KBs descriptions created by IndeGx from publicly accessible endpoints

Fig. 3. Geolocation of the endpoints in KartoGraphI: green items represent an endpoints' IP locations; orange items represent endpoints that gave a timezone not matching this location.

will be given, as shown in the video available at `http://prod-dekalog.inria.fr/SubmissionESWC2022`.

KartoGraphI and IndeGx offer a systematic evaluation of sets of publicly available endpoints based on SPARQL. They show that among the KBs with an endpoint listed as publicly available, one third of them are not reachable. The remaining two thirds are usable thanks to their reuse of common vocabularies. By contrast, the readability of their content is not high on average, very few contain self-descriptions. Their resources are in majority not labeled to describe their content to humans.

In future works, KartoGraphI and IndeGx will extract and present new features present in KBs such as the language tag present. In the future, KartoGraphI will also offer a query editor using IndeGx metadata to guide federated query writing using IndeGx metadata.

References

1. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the quality of the LOD cloud: An empirical investigation. Semantic Web **9**(6), 859–901 (2018)
2. Hasnain, A., Mehmood, Q., Zainab, S.S.e., Hogan, A.: SPORTAL: Profiling the Content of Public SPARQL Endpoints. International Journal on Semantic Web and Information Systems (IJSWIS) pp. 134–163 (Jul 2016). `https://doi.org/10.4018/IJSWIS.2016070105`
3. Vandenbussche, P.Y., Umbrich, J., Matteis, L., Hogan, A., Buil-Aranda, C.: SPARQLES: Monitoring public SPARQL endpoints. Semantic Web **8**(6), 1049–1065 (Aug 2017). `https://doi.org/10.3233/SW-170254`
4. Yamamoto, Y., Yamaguchi, A., Splendiani, A.: YummyData: providing high-quality open life science data. Database: The Journal of Biological Databases & Curation **2018** (2018)