LOD for All: Unlocking Infinite Opportunities

Aisha Naseer[1], Terunobu Kume[2], Tetsuya Izu[1], Nobuyuki Igata[2]

[1]Fujitsu Laboratories of Europe Limited
aisha.naseer@uk.fujitsu.com
tetsuya.izu@uk.fujitsu.com

[2]Fujitsu Laboratories Limited
t-kume@jp.fujitsu.com
igata@jp.fujitsu.com

Abstract. LOD for all platform is among the world's first repositories enabling unified access to the Linked Open Data (LOD) through a single query to the entire LOD datasets. It realises two types of searches: data catalogue search and instance data search. LOD for all offers full-text search based on the heterogeneous storage and using the REST APIs. Majority of LOD sites mainly provide SPARQL endpoints, while LOD for all platform provides more easy-to-use REST APIs, thus saving the developer's effort for preparing the development environments. Also, the LOD for all platform aims to provide an easy-to-implement environment for various LOD applications. Moreover, the LOD for all platform offers capability to display links to other datasets, Classes, Properties, or instances. The LOD for all also offers extensibility in terms of when new datasets are added the relationships or links are automatically displayed.

1 Introduction

The prevalence of terms such as data deluge [1] and information explosion [2] has convinced the data scientists and technologists to consider new ways of handling and manipulating vast amounts of data. In addition to the multitude of private or enterprise data, a significant proportion of data is openly available in the public domain for use as Linked Open Data (LOD)[1]. Now-a-days, the open datasets are becoming more common and available for others to use and study [3]. However, open data when linked yields more benefits and value [4]. The LOD initiative holds the potential to connect various publically available datasets for meaningful use. Hence, we offer the LOD for all[2] platform, which we claim as among the world's first repositories enabling unified access to the Linked Open Data (LOD). It provides a single-stop entry point for the utilisation of LOD, thus allowing searching, browsing, and filtering of the LOD datasets and helping to push the boundaries of our knowledge. The LOD for all highlights links to other datasets and other data, offering multi-level search at the global and dataset levels. Another novel aspect of the LOD for all platform is the provision of unified search capabilities through a single query to the entire LOD datasets. Moreover, the platform hosts instance data for several selected LOD datasets, as mentioned in detail in later sections. Users can search the LOD for all platform to

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://lod4all.net/

adfa, p. 1, 2011.

query: structure of datasets (such as Classes, Properties), and instance data contained within those datasets. LOD for all also hosts the metadata for LOD datasets, thus facilitating the provision of detailed information about a particular dataset including its description, licence, download link, and other useful statistics such as number of Classes, Properties, in-links, and out-links. Moreover, the LOD for all platform is based on the REST API to achieve the ultimate goal of our service, which is to enable and promote the utilization of open data through building an ecosystem around LOD and enable building applications on the top of it.

The LOD for all has been open to the public since January 2014 and is available as a free service at the link: http://lod4all.net/. In future, LOD for all will provide an application development platform utilising LOD that will enable hosting of data and applications, thus encourage developing LOD applications without integrating data on user's side; this capability will be made available later.

2 Description

An exponential growth in the size of data on a daily basis is changing the ways that were previously unobvious, in which we conduct our day-to-day business and approach the world around us. Governments and communities worldwide are supporting growth of the LOD world by publishing their own data on the web and providing a knowhow to the public about using such data [5]. Considering, data as the new currency for science, education, government and commerce [6], it demands increasing abilities and advanced semantic web technologies to not only gather, process, and visualise large datasets but also to study hidden insights from them for meaningful use, such as exploiting LOD in financial reporting [9] and healthcare [10]. It requires not only the real-time access to billions of database elements, such as triples, but also the computational power to efficiently process such data. Our platform 'LOD for all' offers large-scale data storage, enhanced search, and processing for Linked Open Data with unified access, through a Web interface which enables to understand the overall outlook of LOD data (and datasets), as shown in Fig. 1.

Major LOD sites provide SPARQL endpoints, however, broadcast is required if we are to proceed (for example, ?s owl:sameAs dbpedia:hoge); moreover, only 50% of the SPARQL endpoints are currently available [8], thus retrieving incomplete query results. The LOD for all platform offers crawl and centre-processing model to ensure the availability of the majority of SPARQL endpoints. One of the most effective ways of using the LOD data is by exploiting its metadata [7]. The LOD for all platform provides various metadata for the LOD datasets, which is collected from three publically available sources of metadata including Datahub[3], Linked Open Vocabularies (LOV[4]), and LODStats[5]. This enable the provision of detailed information about datasets such as dataset description, licence, SPARQL endpoint, and other useful statistics such as no. of triples, Classes, Properties, in-links, and out-links to other datasets.

[3] http://datahub.io

[4] http://lov.okfn.org/dataset/lov

[5] http://stats.lod2.eu

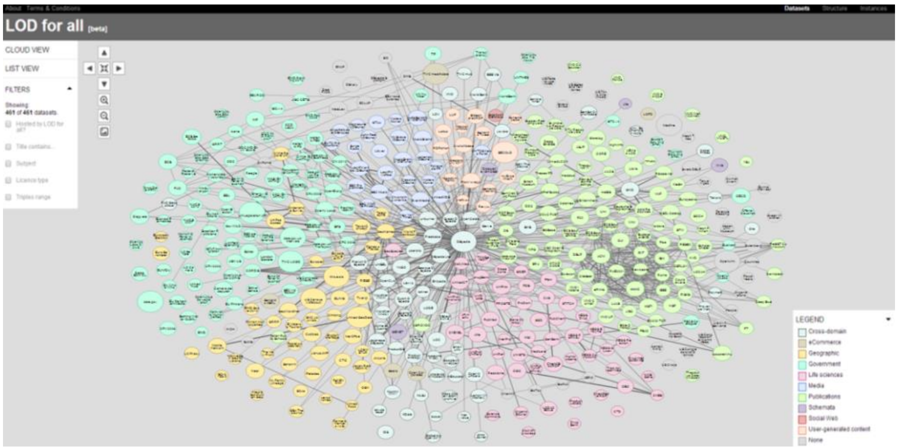

Fig. 1. LOD for all Interface

The LOD for all realises two types of searches: data catalogue search (such as datahub.io in CKAN) and data instance search (such as LOD cache by OpenLinkVirtuoso[6]). The LOD for all platform aims to provide an easy-to-implement environment for various LOD applications.

The following sub-sections describe the novelty, the datasets included, and the features/functionalities of the LOD for all platform.

2.1 Why LOD for all is Innovative?

LOD for all platform is among the world's first[7] repositories enabling unified access to the Linked Open Data (LOD) through a single query to the entire LOD datasets, which currently exists on different sites and in different formats. LOD for all platform is capable of selecting up to 10 times faster than was previously possible by Berlin SPARQL Benchmark (BSBM[8]). The key innovation of the new technology is the ability to easily find and use publically available datasets in order to combine these with other public or private datasets and gain new insights such as displaying links to other datasets and other data. This global repository for Linked Open Data enables unified access and a single-stop entry point for the utilisation of LOD. Another novel aspect of the LOD for all platform is the provision of unified search capabilities through a single query to the entire LOD datasets by using full-text search on a heterogeneous storage. Moreover, the LOD for all platform hosts instance data for several selected LOD datasets.

[6] http://virtuoso.openlinksw.com/

[7] http://www.fujitsu.com/global/about/resources/news/press-releases/2013/0403-02.html

[8] http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark

2.2 LOD for all Datasets

The LOD for all platform hosts metadata and instance data for a selection of LOD datasets that are published on the Datahub. These data are processed in interesting ways to derive meaningful information, for example how the properties and concepts are linked across various datasets. LOD for all uses the Datahub's CKAN APIs to obtain the VoID[9] descriptions of datasets based on the following criteria:

- include all datasets from the “lodcloud” organisation (refer to ‘Linking Open Data Cloud’[10] organisation)
- include a subset of datasets with “lod” tag, for which licences to host instance data are obtained (refer to ‘lod’[11] tag)

The various dataset selections are shown in Fig. 2.

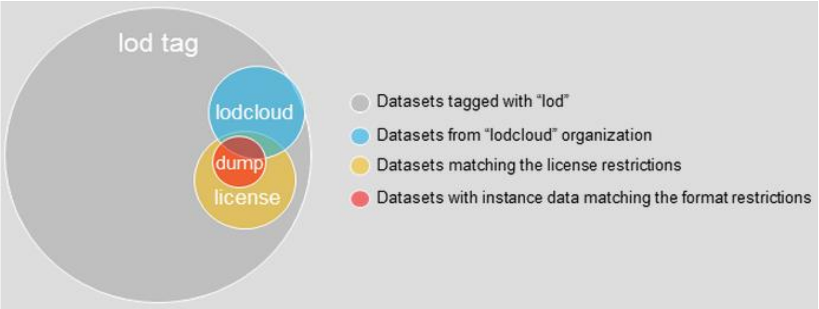
lod tag
lodcloud
dump
license
Datasets tagged with “lod”
Datasets from “lodcloud” organization
Datasets matching the license restrictions
Datasets with instance data matching the format restrictions

Fig. 2. LOD for all Dataset Selection

2.3 LOD for all Features or Functions

The LOD for all platform possesses the following key features:

Datasets Search, Browsing, Filtering and Download.
LOD for all offers unified access point for the linked open dataset, which users can easily search, browse, filter, and download. Our linking feature displays links in the datasets at the triple level thus enabling browsing of LOD datasets to the lowest level of granularity in terms of the structure and the instances. Moreover, LOD for all platform contains custom crawlers that are developed to obtain dumps of the instance data for selected LOD datasets considering their license restrictions. These dumps are made available for the users to be downloaded in the RDF format.

Moreover, the platform offers filtering options to visualise only a subset of the LOD datasets based on the filtering criteria such as the dataset title, license, subject, or number of triples.

[9] http://semanticweb.org/wiki/VoID
[10] http://datahub.io/dataset?organization=lodcloud
[11] http://datahub.io/dataset?tags=lod

Periodic Update with Active Service.

LOD for all deals with data dynamism and huge backlog by providing periodic updates to the LOD datasets, without stopping the server, while new triples are periodically added. Since the LOD datasets are subject to frequent updates, the LOD for all platform's custom crawlers crawl the web for fetching open data to conduct bulk update of the datasets considering their license restrictions. These disparate data are obtained from multiple sources and then added to the LOD for all platform.

Structure Search.

LOD for all offers structure search feature that enables discovery of Classes and Properties and allowing linking. The structure search feature uses LOD vocabularies and lightweight semantics. It enables the users to explore RDF vocabularies across various LOD datasets, and allows searches on two basic concepts in RDF schema: Class and Property, where a Class is anything of type owl:Class, and a Property is either a type of owl:DatatypeProperty or owl:ObjectProperty.

- Class search – LOD for all platform returns all Classes matching the user keyword, plus, it returns all other Classes having structural relationships with the matching classes such as "subClassOf", "disjointWith", "equivalentClass", "sameAs", etc.
- Property search – LOD for all platform returns all Properties matching the user keyword, plus, it returns all other Properties and Classes having structural relationships with the matching Property such as "domain", "range", "subPropertyOf", etc.

Instance Search.

Another unique feature of the LOD for all platform is the multi-level instance search at the Global level and the Dataset level, while offering a single query to the entire LOD datasets. LOD for all offers full-text search based on the heterogeneous storage. Since the LOD for all hosts instance data, it becomes feasible to display links to other datasets and other LOD data that are hosted by the LOD for all platform. Considering the massive amounts of data, the LOD for all platform offers reasonable velocity in terms of the instance search. Moreover, we provide simple user interface like a Web search, so the user does not need to write SPARQL queries for the instance search.

Noticeably, both the structure search and the instance search features offer the capabilities of linking that displays links to other datasets, Classes, Properties, or instances. LOD for all provides support to grasp the overall LOD by combining datasets, vocabulary (by structure search), and data (by instance search). The LOD for all platform also offers extensibility when the new datasets are added and new relationships or links are automatically displayed.

Metadata Generation.

The LOD for all platform also hosts metadata for several LOD datasets thus enabling interoperability and facilitating integration with private enterprise data. These metadata are integrated or collected from three publically available sources [7] of metadata:

- Datahub – it provides free access to many open datasets and other publically available metadata. Types of metadata it provides include: dataset name, label, description, license, tags, etc. These metadata can be accessed through JSON-based CKAN API[12], or downloadable from https://github.com/lodcloud/datahub2void.
- LOV – it provides easy access methods to the ecosystem of vocabularies, provides explicit links, and metrics on how they are used in LOD cloud. It contains descriptions of RDFS vocabularies used by LOD datasets, which are metadata either formally declared by the vocabulary publishers or added by the LOV curators.
- LODStats – it provides statistics about various datasets registered with DataHub. These statistics are computed based on declarative description of statistical dataset characteristics, offering 32 statistics defined by the VoID descriptions including: class usage, property usage, data types used, average length of string literals.

Using these open metadata sources and performing statistical graph analytics, the metadata generation feature enables provision of detailed information about datasets such as dataset descriptions, licences, SPARQL endpoints, and other useful statistics such as such as number of Triples, Classes, Properties, and Entities in a particular LOD dataset; also providing in-links, and out-links to other datasets.

3 Acknowledgement

The LOD for all platform and public service results from a collaboration between Fujitsu Laboratories (Japan and Europe) and the INSIGHT@ NUI Galway, Ireland (formerly known as DERI) in the scope of the KI2NA project. We also acknowledge the enormous efforts of our team in the design and development of the lod4all.net service. The key contributors to the lod4all.net service are (in alphabetical order of surnames):

Neil Benn, Nuno Carvalho, Bo Hu, Hiromichi Kobashi, Terunobu Kume, Vivian Lee, Takahide Matsutsuka, Roger Menday, Seiji Okajima, Eduarda Mendes Rodrigues.

References

1. Casacuberta, D., and Vallverdú, J. "E-Science and the data deluge". Philosophical Psychology, 27(1):126-140, 2014.
2. Nambiar, R., Chitor, R., and Joshi, A. "Data Management - A Look Back and a Look Ahead". Lecture Notes in Computer Science, 8163:11-19, 2014.
3. Dalton, L. "On the Reporting of New Information from Open Data Sets". American Journal of Surgical Pathology. 38(3):433-434, 2014.

[12] http://ckan.readthedocs.org/en/latest/

4. Berners-Lee, T. “Linked Data, 2009”. Retrieved: August 20, 2014, from http://www.w3.org/DesignIssues/LinkedData.html
5. Nobuyuki, I., Nishino, F., Kume, T. and Matsutsuka, T. “Information Integration and Utilization Technology using Linked Data”. Fujitsu Science and Technology Journal (FSTJ), vol. 50, no. 1, pp. 3-8, January 2014.
6. Pire, C. M., Guedon, J-C., and Blatecky, A. “Scientific Data Infrastructures: Transforming Science, Education, and Society”. Zeitschrift für Bibliothekswesen und Bibliographie, 60(6):325-331, 2013.
7. Lee, V., Naseer, A., and Kume, T. “LOD Metadata Harvester: Enabling Better Decision Making”. Proceedings of the IEEE/ACM 29th International Conference on Automated Software Engineering (ASE 2014), Stanford, USA, May 2014.
8. Aranda, C.B., Hogan, A., Umbrich, J., Vandenbussche, P-Y. “SPARQL Web-Querying Infrastructure: Ready for Action?” Proceedings of the International Semantic Web Conference (ISWC 2013), pp. 277-293, 2013.
9. Goto. M., Hu, B., Naseer, A., and Vandenbussche, P-Y. “Linked Data for Financial Reporting”. Proceedings of the Fourth International Workshop on Consuming Linked Data (COLD 2013), Sydney, Australia, October 2013.
10. Novacek, V., and Naseer, A. “Linking the Scientific and Clinical Data with KI2NA-LHC – An Outline”. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS 2013), Porto, Portugal, June 2013.

Appendix

The LOD for all platform meets the various evaluation requirements, including the mandatory or minimal requirements and the additional desirable features. These are explicitly summarised as follows:

LOD for all meets minimal requirements.

1. LOD for all is an end-user platform, providing a practical value to the general Web user and especially to the domain experts that want to build application on top of the LOD for all platform. It show-cases the use of semantic web technologies, such as RDF and SPARQL, through its various features and functionalities, such as structure search and metadata generation features.

2. The information sources or datasets used in LOD for all platform are:

- LOD datasets, which are under diverse ownership or control
- open metadata are heterogeneous syntactically, structurally, and semantically
- all real world data from the Datahub.io and not toy examples

3. In the LOD for all platform, the meaning of data plays a central role in a way that:

- meaning is represented using Semantic Web technologies, such as RDF

- data is manipulated or processed in interesting ways such as retrieving results from a single query to the entire LOD datasets to derive useful information, such as display links among various data elements and LOD datasets
- the semantic information processing plays a central role in achieving things that alternative technologies cannot do; for example, the Instance search feature enables finding resources related to the concept being searched; the search results for "Person" are triples where the concept "Person" is present in labels (e.g. rdfs:label), descriptions and other literals

LOD for all offers additional desirable features.
In addition to the above minimum requirements, the LOD for all platform offers other desirable features, such as:

- attractive and functional Web interface like a Web search for human users, so the user does not need to write SPARQL queries
- scalable in terms of the amount of data used and the distributed components working together; it uses significant amounts of data currently published on the Datahub.io (LOD Cloud organisation) while hosting metadata for 461 datasets out of which 79 datasets are hosted with the instance data
- LOD for all has been up and running as a free public service since January 2014
- novelty in applying semantic technology that has not been considered before by provision of unified search capabilities through a single query to the entire LOD datasets by using REST APIs
- different functionality going beyond pure information retrieval, the LOD for all platform is based on the REST APIs to achieve the ultimate goal of our service, which is to enable and promote the utilization of open data through building an ecosystem around LOD and enable building applications on the top of it
- contextual information is used for rankings in the instance search feature through the use of full text search
- the data contained within the LOD for all platform is multi-language; the 'About' page provides information in Japanese language as well, and it is accessibility on a range of devices including smartphones, ipads, or tablet PCs