Demonstration of `LODChain`: How to Tackle the Problem of Low Connectivity for your RDF Dataset

Michalis Mountantonakis[1,2,*], Yannis Tzitzikas[1,2]

[1]*Institute of Computer Science, FORTH, Heraklion, Greece*
[2]*Department of Computer Science, University of Crete, Heraklion, Greece*

Abstract

This paper demonstrates `LODChain`, which is an online service for tackling the problem of low connectivity for a given RDF dataset, by strengthening its connections to the rest of LOD Cloud. The demo focuses on presenting the connectivity analytics, visualizations and enrichment services for the user/publisher, which are offered for several parts of the dataset, e.g., for owl:sameAs mappings, entities, schema elements and triples. Moreover, we show how these analytics can be exploited for improving the connectivity of a dataset, and thus its discoverability and reusability, by using a scenario with a real RDF dataset.

Keywords

Connectivity, Data Integration, Visualizations, Data Discovery, Data Enrichment, owl:sameAs

1. Introduction

Current approaches for dataset search and discovery are mainly metadata-based and ignore the content of the datasets [1], and existing RDF approaches do not favor their discoverability and reusability [2]. As a result, “Linked Data Cloud consists of loosely inter-linked individual subgraphs” [3]. For tackling the problem of low connectivity, we present `LODChain`; an online web service, which is available at https://demos.isl.ics.forth.gr/LODChain, that strengthens the connectivity of an RDF dataset, by computing the transitive and symmetric closure of equivalence relationships, such as owl:sameAs, between the given dataset and hundreds of RDF datasets indexed by LODsyndesis [4]. Moreover, it finds common entities, schema elements and triples, and produces connectivity analytics, visualizations and enrichment services. The target is to aid the publishers a) to evaluate the connectivity of their dataset, e.g., before publishing it to the LOD Cloud, b) to strengthen its connectivity, i.e, by discovering new connections, and c) to verify and enrich its content by finding common and complementary data.

ISWC 2022: The 21st International Semantic Web Conference, 23-27 October 2022

mountant@ics.forth.gr (M. Mountantonakis); tzitzik@ics.forth.gr (Y. Tzitzikas)
https://users.ics.forth.gr/~mountant/ (M. Mountantonakis); https://users.ics.forth.gr/~tzitzik/ (Y. Tzitzikas)
0000-0002-1951-0241 (M. Mountantonakis); 0000-0001-8847-2130 (Y. Tzitzikas)

CEUR Workshop Proceedings (CEUR-WS.org)

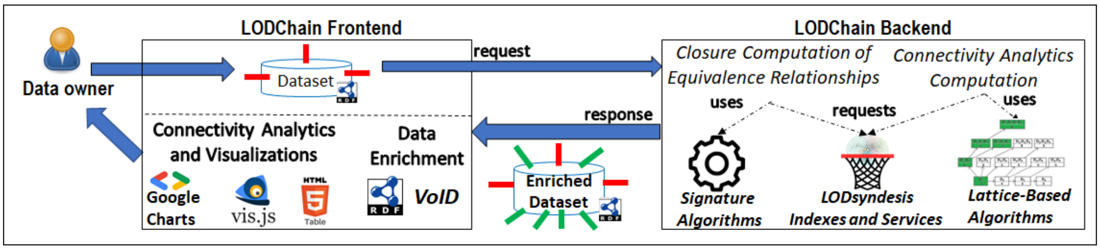


Figure 1: The Process and Architecture of `LODChain` Web Application

This demo paper is a complement to an accepted paper of ISWC'22 Resource Track [5]. The accepted resource paper presents all the methods and algorithms for computing the connectivity analytics of `LODChain`, and an evaluation with real datasets. On the contrary, this demo paper focuses on presenting in more details the offered analytics, visualizations, and data enrichment services, by providing a scenario of how the services can be used for a real dataset.

The rest of this demo paper is organized as follows: §2 introduces the related work, §3 describes in brief the process of `LODChain`, §4 presents all its services, and §5 concludes the paper.

2. Related Work

There are several approaches for aiding the interlinking, discoverability and reusability of RDF datasets. Concerning interlinking, applications like WIMU [6], MetaLink [7] and LODsyndesis [4] can be used for finding all the datasets and equivalent URIs of a given entity. Regarding discoverability and reusability, services such as LOD Cloud (http://lod-cloud.net), Google Dataset Search [1], LODatio [8] and LODAtlas [9] offer metadata-based search for discovering and reusing RDF datasets. Comparing to these approaches, `LODChain` is the first online service for strengthening the connectivity of an RDF dataset (of any domain), even before its actual publishing, by exploiting the content of hundreds of RDF datasets.

3. The Process and Architecture of `LODChain`

Fig. 1 shows the architecture of `LODChain` web application. First, the data owner uploads a dataset in RDF format (e.g., N-Triples, RDF/XML), which can include connections to other datasets. `LODChain` analyzes the content of the dataset, for computing the transitive and symmetric closure of equivalence relationships, i.e., owl:sameAs, owl:equivalentProperty, and owl:equivalentClass relationships, between the given dataset and hundreds of LOD datasets, for discovering new connections. Then it produces connectivity analytics, by accessing the indexes and services of LODsyndesis, and by using signature algorithms and lattice-based maximization algorithms [4]. Afterwards, connectivity analytics for the enriched dataset are presented to the user through HTML tables, charts and visualizations (produced by popular Javascript libraries), whereas additional RDF files are offered, which can be used for enriching the content of the given dataset.

4. Connectivity Analytics, Visualizations & Enrichment

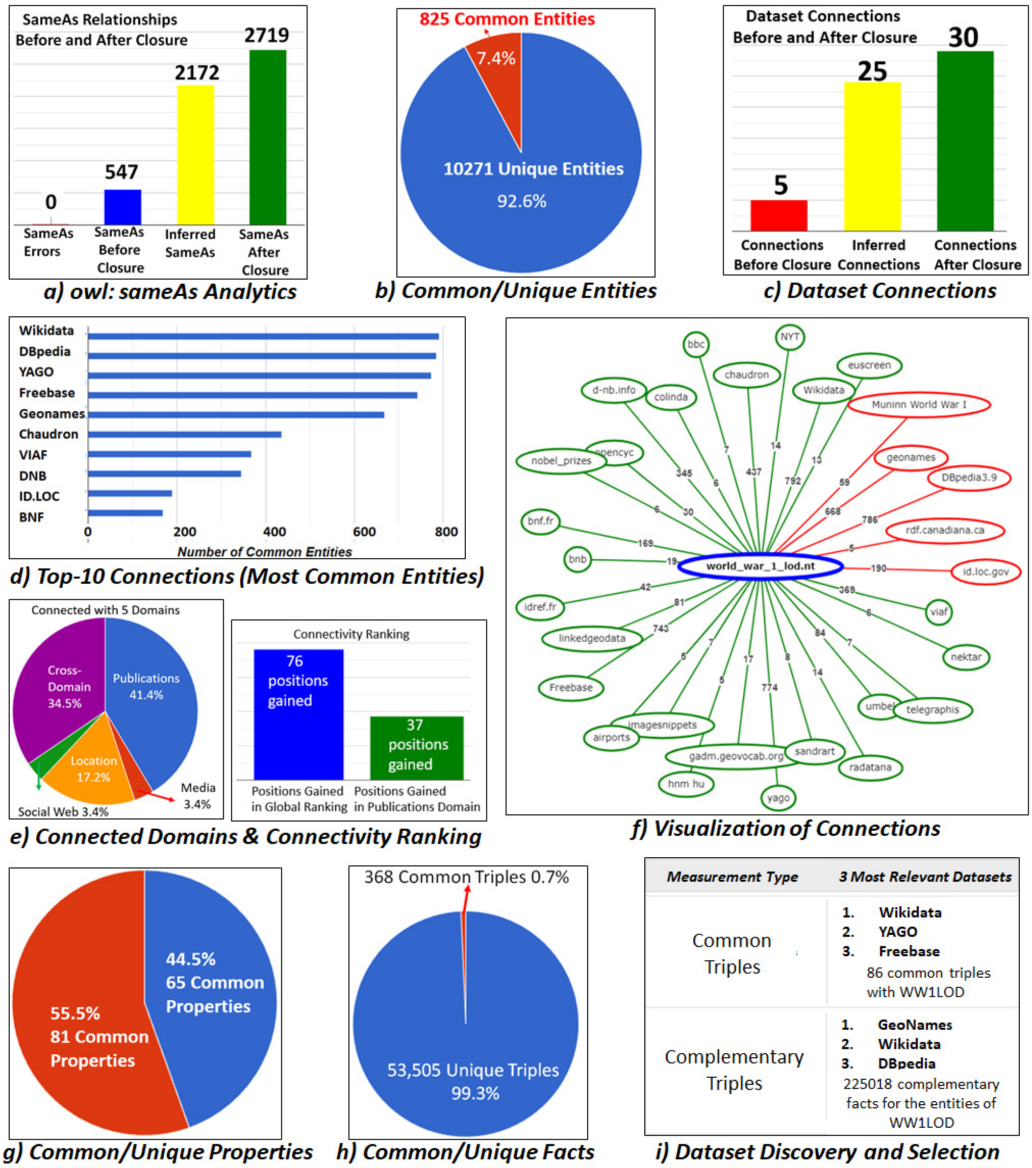
We present the offered services, divided in six categories by using a real dataset from publications domain, i.e., *WW1LOD* [10], including 47,616 triples, 547 owl:sameAs mappings and connections with 5 RDF datasets. The target is to explain how the results of each category can be used for evaluating/improving the connectivity of a given dataset. For more details, a video is available in https://youtu.be/Kh9751p32tM, and screenshots including analytics for more datasets are presented in https://zenodo.org/record/6467419. In the demonstration, we will showcase the offered analytics and visualizations, by showing results over such real datasets.

Category 1. Analytics over owl:sameAs Mappings. The objective is to evaluate for the given dataset if `LODChain` a) inferred new mappings through the computation of owl:sameAs closure, and b) detected possible owl:sameAs errors, i.e., `LODChain` checks if an entity is equivalent with two or more real entities in LODsyndesis. `LODChain` offers a chart analyzing the results of the computation of closure, including the number of a) possible errors, and b) owl:sameAs relationships before and after the computation of closure. In case of detecting errors, they can be downloaded from the user for aiding the process of fixing them. As Fig. 2(a) shows, we inferred thousands of owl:sameAs mappings for *WW1LOD*, without detecting errors, which is very positive for its connectivity, i.e., we expect that more connections have been discovered.

Category 2. Analytics over Entities. The objective is to evaluate how connected is the given dataset to the rest of LOD Cloud, and to quantify the gain of the computation of owl:sameAs closure. `LODChain` provides charts for the number of unique/common entities, the connections before and after the closure, the top-10 connected datasets, and its connectivity compared to datasets of several domains. For instance, Fig. 2(c-d) shows that *WW1LOD* shares 825 entities with other datasets (such as DBpedia and Wikidata), whereas 25 new inferred connections were discovered through `LODChain`. Moreover, a graph visualizes all the connections of the dataset (Fig. 2(f)), where the red nodes depict the old connections and the green nodes the new ones (due to closure), and the labels of edges the number of common entities. From these charts, we can see that the connectivity of *WW1LOD* was highly increased. On the contrary, in a scenario where a dataset has either zero or few connections, even after the computation of closure, it could be an indication of bad/low connectivity, e.g., few or/and incorrect links to other datasets. Finally, charts like Fig. 2(e) can be used for quantifying the connectivity of a dataset comparing to other domains, e.g., *WW1LOD* is connected with datasets from 5 different domains.

Category 3. Analytics over Schema Elements. The goal is to evaluate if existing ontologies are reused, since it is important for a) comparing the values of the same fact for an entity [4], e.g., for data verification and/or for detecting conflicts, and b) for enabling schema-based integration, e.g., for creating a mediator or a data warehouse [11]. `LODChain` offers charts showing the number of common/unique properties or classes, and the top-10 most connected datasets for these URIs. For example, *WW1LOD* shares many common properties (Fig. 2(g)), which can be important for schema-based integration. On the contrary, if a dataset has a low number of common properties/classes, existing schemas are not reused and the creation of further mappings will be probably needed, e.g., for achieving schema-based integration.

Category 4. Analytics over Triples. The objective is to evaluate whether a dataset shares common triples (or facts) with others, and if there are available complementary facts for its entities. `LODChain` offers charts showing the number of common/unique facts, and the top-10



a) owl: sameAs Analytics b) Common/Unique Entities c) Dataset Connections

d) Top-10 Connections (Most Common Entities) f) Visualization of Connections

e) Connected Domains & Connectivity Ranking

g) Common/Unique Properties h) Common/Unique Facts i) Dataset Discovery and Selection

Figure 2: Visualizations of the Connectivity Analytics of LODChain for the dataset *WW1LOD*

datasets offering the most common and complementary facts for the entities of the given dataset. Fig. 2(i) shows that *WW1LOD* contains 368 common facts with at least one RDF dataset. Certainly, it is not required for a dataset to contain common facts with other ones for being included in the LOD Cloud, however, they can be of primary importance for data verification. On the contrary, the existence of complementary facts can be useful for enriching the content of the given dataset, e.g., we found 362,339 complementary facts for the entities of *WW1LOD*.

Category 5. Dataset Discovery and Selection. In many cases, the publishers desire to select and integrate their dataset with only K datasets, i.e., since it can be time-consuming to export and to integrate with any available relevant dataset [11]. LODChain offers measurements for finding the K most relevant datasets for a given dataset according to the number of common entities, properties, classes and facts, and of complementary facts. Fig. 2(i) shows the top-3 datasets according to the number of a) common and b) complementary facts for our running example. The most relevant datasets differ in these two cases, which shows that for different needs (verification versus data enrichment), different combinations of datasets can be used.

Category 6. Data Enrichment. `LODChain` gives the option to the publisher to export the results of the connectivity analytics and to enrich the given dataset with additional data for creating an enriched version of their dataset. In particular, one can export in RDF format all the common entities, schema elements and facts (and their provenance), the inferred owl:sameAs relationships, complementary facts, and rich metadata including the results of the analytics. All these data (or any subset of them) can be used from a publisher for improving his/her dataset.

5. Concluding Remarks

We presented the connectivity analytics, visualizations and enrichment services, offered by `LODChain`; a web application for strengthening the connectivity of an RDF dataset over hundreds of datasets. As a future work, we plan to extend `LODChain` for offering more services, e.g., finding connections for almost or totally disconnected datasets (by performing instance matching), and services for further aiding the process of fixing connectivity errors, e.g., owl:sameAs errors.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 coordination and support action 4CH (Grant agreement No 101004468).

References

[1] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, P. Groth, Dataset search: a survey, The VLDB Journal 29 (2020) 251–272.

[2] J. Debattista, J. Attard, R. Brennan, D. O'Sullivan, Is the LOD Cloud at risk of becoming a museum for datasets? looking ahead towards a fully collaborative and sustainable lod cloud, in: Proceedings of WWW Conference, 2019, pp. 850–858.

[3] P. Hitzler, A review of the semantic web field, Communications of the ACM 64 (2021) 76–83.

[4] M. Mountantonakis, Y. Tzitzikas, Content-based union and complement metrics for dataset search over RDF knowledge graphs, ACM JDIQ 12 (2020) 1–31.

[5] M. Mountantonakis, Y. Tzitzikas, LODChain: Strengthen the connectivity of your RDF dataset to the rest LOD Cloud, in: ISWC 2022 (Accepted in Resource Track), 2022.

[6] A. Valdestilhas, T. Soru, M. Nentwig, E. Marx, M. Saleem, A.-C. N. Ngomo, Where is my URI?, in: European Semantic Web Conference, Springer, 2018, pp. 671–681.

[7] W. Beek, J. Raad, E. Acar, F. v. Harmelen, Metalink: A travel guide to the LOD cloud, in: European Semantic Web Conference, Springer, 2020, pp. 481–496.

[8] T. Gottron, A. Scherp, B. Krayer, A. Peters, LODatio: A schema-based retrieval system for linked open data at web-scale, in: ESWC, Springer, 2013, pp. 142–146.

[9] E. Pietriga, H. Gözükan, C. Appert, M. Destandau, Š. Čebirić, F. Goasdoué, I. Manolescu, Browsing linked data catalogs with LODAtlas, in: ISWC, Springer, 2018, pp. 137–153.

[10] E. Mäkelä, J. Törnroos, T. Lindquist, E. Hyvönen, WW1LOD: an application of CIDOC-CRM to World War 1 linked data, IJDL 18 (2017) 333–343.

[11] M. Mountantonakis, Y. Tzitzikas, Large-scale semantic integration of linked data: A survey, ACM CSUR 52 (2019) 1–40.