Services for Large Scale Semantic Integration of Data

by Michalis Mountantonakis and Yannis Tzitzikas (ICS-FORTH)

LODsyndesis is a new suite of services that helps the user to exploit the linked data cloud.

In recent years, there has been an international trend towards publishing open data and an attempt to comply with standards and good practices that make it easier to find, reuse and exploit open data. Linked Data is one such way of publishing structured data and thousands of such datasets have already been published from various domains.

However, the semantic integration of data from these datasets at a large (global) scale has not yet been achieved, and this is perhaps one of the biggest challenges of computing today. As an example, suppose we would like to find and examine all digitally available data about Aristotle in the world of Linked Data. Even if one starts from DBpedia (the database derived by analysing Wikipedia), specifically from the URI "http://dbpedia.org/resource/Aristotle" it is not possible to retrieve all the available data because we should first find ALL equivalent URIs that are used to refer to Aristotle. In the world of Linked Data, equivalence is expressed with "owl:sameAs" relationships. However, since this relation is transitive, one should be aware of the contents of all LOD datasets (of which there are currently thousands) in order to compute the transitive closure of "owl:sameAs", otherwise we would fail to find all equivalent URIs. Consequently, in order to find all URIs about Aristotle, which in turn would be the lever for retrieving all data about Aristotle, we have to

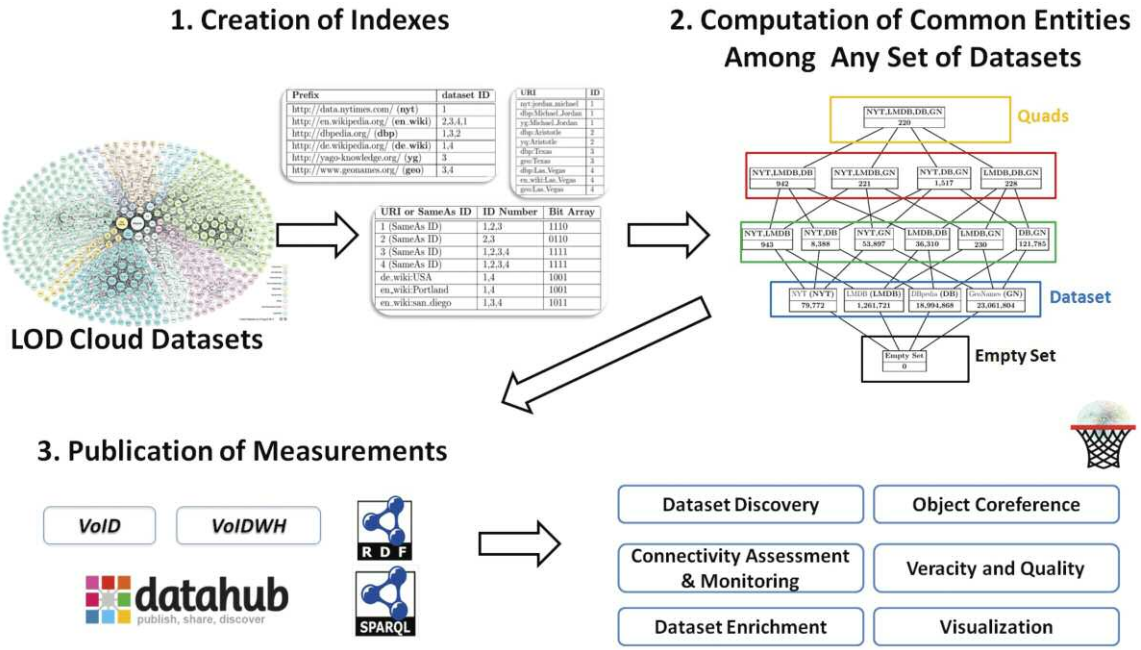

Figure 1: The whole process of LODsyndesis

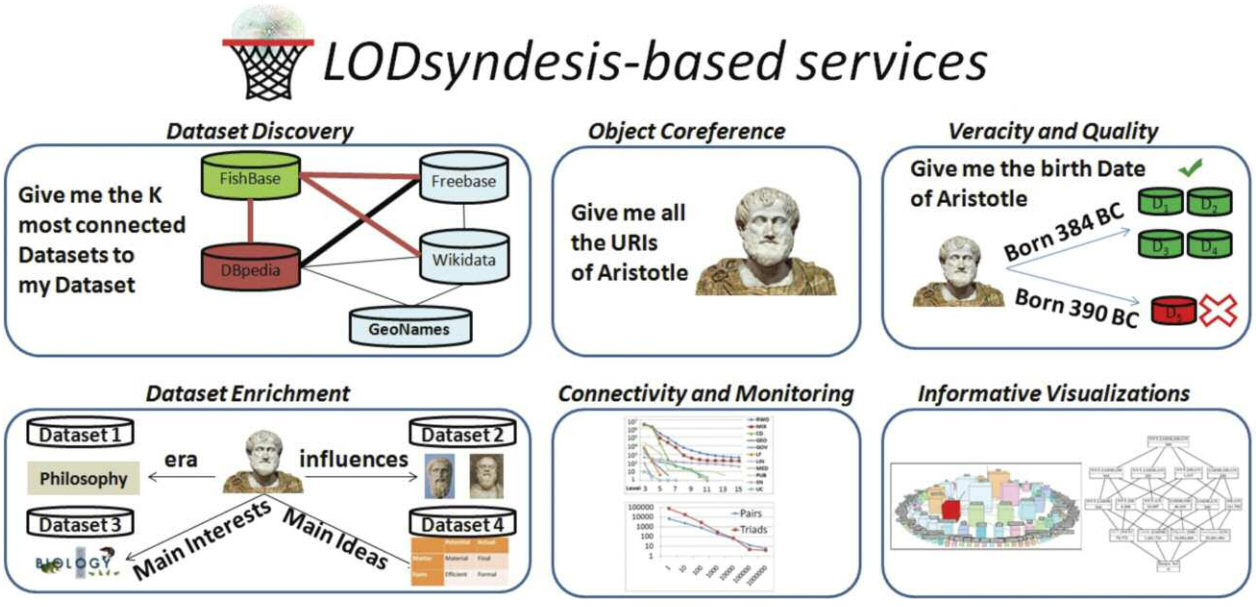

Figure 2: Services offered by LODsyndesis.

index and enrich numerous datasets, through cross-dataset inference.

.

The Information Systems Laboratory of the Institute of Computer Science of FORTH, designs and develops innovative indexes, algorithms and tools to assist the process of semantic integration of data at a large scale. This endeavour started two years ago, and the current suite of services and tools that have been developed are known as "LODsyndesis" [L1]. As shown in Figure 1, the distinctive characteristic of LODsyndesis is that it indexes the contents of all datasets in the Linked Open Data cloud [1]. It then exploits the contents of datasets to measure the commonalities among the datasets. The results of measurements are published on the web and can be used to perform several tasks. Indeed, "global scale" indexing can be used to achieve various outcomes, and an overview of the available LODsyndesis-based services for these tasks can be seen in Figure 2.

First, it is useful for dataset discovery, since it enables content-based dataset discovery services, e.g., it allows answering queries of the form: "find the K datasets that are more connected to a particular dataset". Another task is object co-reference, i.e., how to obtain complete information about one particular entity (identified by a URI) or set of entities, including provenance information. LODSyndesis can also help with assessing the quality and veracity of data, since the collection of all information about an entity, and the cross-dataset inference that can be achieved, allows contradictions to be identified and provides information for data cleaning or for estimating and suggesting which data are probably correct or most accurate. Last but not least, these services can help to enrich datasets with more features that can obtain better predictions in machine learning tasks [2] and the related measurements can be exploited to provide more informative visualisation and monitoring services.

The realisation of the services of LODSyndesis is challenging because they presuppose knowledge of all datasets. Moreover, computing the transitive closure of all "owl:sameAs" relationships is challenging since most of the algorithms for doing this require a lot of memory. To tackle these problems LODsyndesis is based on innovative indexes and algorithms [1] appropriate for the needs of the desired services. The current version of LODsyndesis indexes 1.8 billion triples from 302 datasets, it contains measurements of the number of common entities among any combination of datasets (e.g., how many common entities exist in a triad of datasets), while the performed measurements are also available in DataHub [L2] for direct use. It is worth noting that currently only 38.2% of pairs of datasets are connected (i.e., they share at least one common entity) and only 2% of entities occur in three or more datasets. The aforementioned measurements reveal the sparsity of the current datasets of the LOD Cloud and justify the need for services for assisting integration at large scale. This method is efficient, the time for constructing the indexes and performing all these measurements being only 22 minutes using a cluster of 64 computers.

Over the next two years we plan to improve and extend this suite of services. Specifically, we plan to advance the data discovery services and to design services for estimating the veracity and trust of data.

This project is funded by FORTH and it has been supported by the European Union Horizon 2020 research and innovation programme under the BlueBRIDGE project (Grant agreement No 675680).

Links:

[L1] http://www.ics.forth.gr/isl/LODsyndesis/

[L2] https://datahub.io/dataset/connectivity-of-lod-datasets

References:

[1] M. Mountantonakis and Y. Tzitzikas, "On measuring the lattice of commonalities among several linked datasets," Proceedings of the VLDB Endowment, vol. 9, no. 12, pp. 1101-1112, 2016.

[2] M. Mountantonakis and Y. Tzitzikas, "How Linked Data can aid Machine Learning-based Tasks," in International Conference on Theory and Practice of Digital Libraries, 2017.

Please contact:

Yannis Tzitzikas
FORTH-ICS and University of Crete
+30 2810 391621
tzitzik@ics.forth.gr

On the Interplay between Physical and Digital World Accessibility

by Christophe Ponsard, Jean Vanderdonckt and Lyse Saintjean (CETIC)

Many means of navigating our physical world are now available electronically: maps, streets, schedules, points of interest, etc. While this "digital clone" is continually expanding both in completeness and accessibility, an interesting interplay scenario arises between physical and digital worlds that can benefit all of us, especially the mobility-impaired.

Physical world accessibility is about ensuring that features of physical places (e.g., shops, tourist attractions, offices) are designed to accommodate the (dis)abilities of people visiting them in order to ensure optimal access for all, including the estimated 15% of the population living with some kind of impairment. Our world is undergoing digitisation at an ever-increasing rate. Online maps are becoming so precise that the inner structure of buildings is often captured, tours are available, while locations are ranked by users, etc. As a consequence, many new opportunities have arisen for breaking accessibility barriers and better combining the physical and digital aspects of accessibility. For example, the use of the available online information can be complemented with physical measures to help assess the accessibility of build-