This is a preprint of the article: Michalis Mountantonakis, Yannis Tzitzikas, "LODsyndesis: The biggest knowledge graph of the Linked Open Data cloud that includes all inferred equivalence relationships", ERCIM News 2018 (114), July 2018

LODsyndesis: The biggest knowledge graph of the Linked Open Data cloud that includes all inferred equivalence relationships

by Michalis Mountantonakis and Yannis Tzitzikas (ICS-FORTH)

LODsyndesis is the biggest knowledge graph that includes all inferred equivalence relationships, which occur between entities and schemas of hundreds of Linked Open Data cloud datasets. LODsyndesis webpage offers several services for exploiting the aforementioned knowledge graph, e.g., a service for collecting fast all the available information (and their provenance) for 400 million of entities, an advanced Dataset Discovery service for finding the most connected datasets to a given dataset, and others.

The internet's enormous volume of digital open data can be a valuable asset for scientists and other users, but this is only possible if it is easily findable, reusable and exploitable. One challenge is to link and integrate these data so that users can find all data about an entity, and to help estimate the veracity and correctness of the data. One way to achieve this is to publish the data in a structured way using Linked Data techniques.

However, the semantic integration of data at a large scale is not a straightforward task, since publishers tend to use different URIs, names, schemas and techniques for creating their data. For instance, to represent a fact, say "Stagira is the birth place of Aristotle", different datasets can use different URIs to represent the entities "Aristotle" and "Stagira", and the schema element "birth place". Figure 1 depicts an example of four datasets that contain information about ``Aristotle". With Linked Data one can partially overcome this difficulty by creating cross-dataset relationships between entities and schemas, i.e., by exploiting some predefined properties, such as owl:sameAs, owl:equivalentProperty and owl:equivalentClass (the equivalence relationships of our example are shown in the upper right side of Figure 1). However, all these relations are transitive and symmetric, which implies that in order to collect all the available information for an entity and to not miss facts that are common to two or more datasets, it is necessary to compute the transitive closure of these relations, a task that presupposes knowledge from all the datasets.

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. The suite of services and tools that have been developed are referred to as "LODsyndesis" [L1]. Comparing to [1], the current version allows the full contents of datasets to be indexed in a parallel way [2,3]. To enable fast access to all the available information about an entity, we have created global scale entity-centric indexes, where we store together all the available information for any entity, by taking into consideration the equivalence relationships among datasets. An example about the entity ``Aristotle" is shown in Figure 1. By collecting all facts about an entity, we can easily spot those that are common in two or more datasets (e.g. we can see that all datasets agree that Stagira is the birth place of Aristotle), the conflicting ones (birthYear), and the complementary ones (Philosopher).

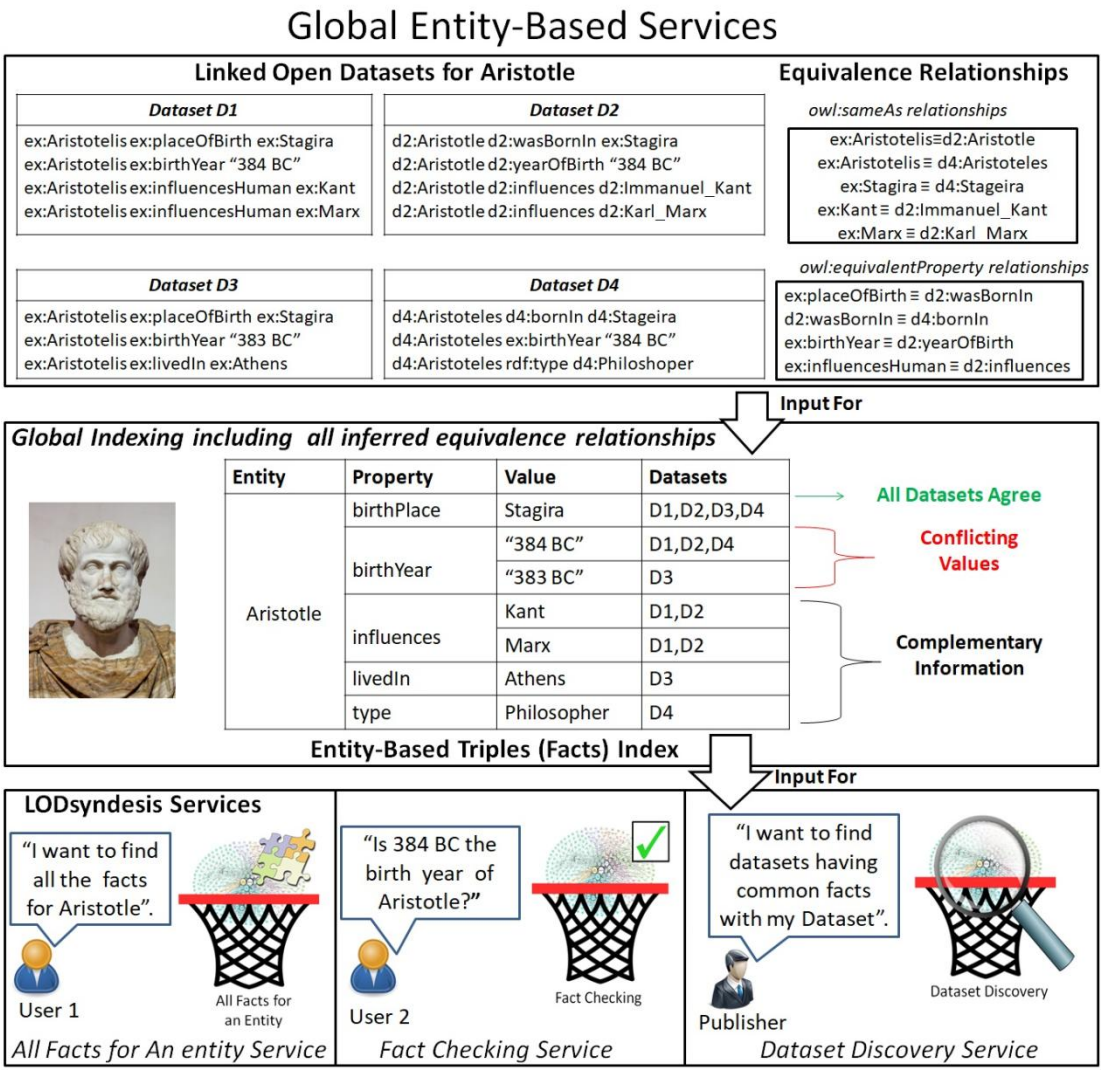

Figure 1: The process of global indexing and the offered LODsyndesis Services

The current version of LODsyndesis indexes two billion triples, which contain information for about 400 million of entities from 400 datasets. Apart from the services introduced in [1], it offers two additional state-of-the-art services: (i) a service for finding all the available information (and its provenance) about an entity, and (ii) a fact checking service where one can check which datasets agree that a fact holds for a specific entity (e.g., check whether the birth date of Aristotle is 384 BC) and which are the contradicting values.

In addition, the current version of LODsyndesis contains measurements about the commonalities among all these (or any combination of) datasets, namely: number of common entities, common schema elements, common literals and common facts (all these measurements have been published also to DataHub [L2] for direct exploitation). These measurements are leveraged by the offered Dataset Discovery Service to enable users to find the datasets that are connected to a given dataset.

The measurements provide some interesting insights about the connectivity of the LOD cloud. As reported in [2,3], only 11 % of the possible datasets' pairs share common entities, and 5.2 % of them share common facts, which means that most datasets contain complementary information, even for the same entities. We have observed that many

publishers do not create equivalence relationships with other datasets; consequently their datasets cannot be easily integrated with other datasets. When it comes to efficiency, the creation of all required indexes and the calculation of the aforementioned measurements takes only 81 minutes using 96 machines. Based on these indexes, the provision of services offered by LODsyndesis are very fast, i.e., on average less than five seconds are needed to find the most connected datasets for a given dataset, whereas, on average it takes less than 10 seconds to show (or export) all the available information of an entity, or to check whether a fact holds (for an entity). As future work, we plan to provide more advanced data discovery and veracity estimation services.

This work has received funding from: a) FORTH and b) the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) .

References:

[1] Mountantonakis, M., & Tzitzikas, Y. (2017). Services for Large Scale Semantic Integration of Data. ERCIM News, 2017.

[2] Mountantonakis, M., & Tzitzikas, Y. (2018). Scalable Methods for Measuring the Connectivity and Quality of Large Numbers of Linked Datasets. Journal of Data and Information Quality (JDIQ), 9(3), 15.

[3] M. Mountantonakis and Y. Tzitzikas, (2018). High Performance Methods for Linked Open Data Connectivity Analytics. Information, 9(6),134, MDPI. (Open Access URL: http://www.mdpi.com/2078-2489/9/6/134)

Links:

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

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

Please contact:

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