Monitoring the evolution of LOD

The Linked Open Data cloud offers a vast amount of data of both domain-specific and domain-independent nature. Linked data describes a method for publishing structured data so that it can be interlinked and become more useful. Such interlinking enables data from a wide variety of sources like *geonames.org*, *DBpedia*, etc. to be connected and queried. There is a growing number of published and openly available datasets[1]. These datasets are subject to changes and evolve over time, based on the dynamics of their content.

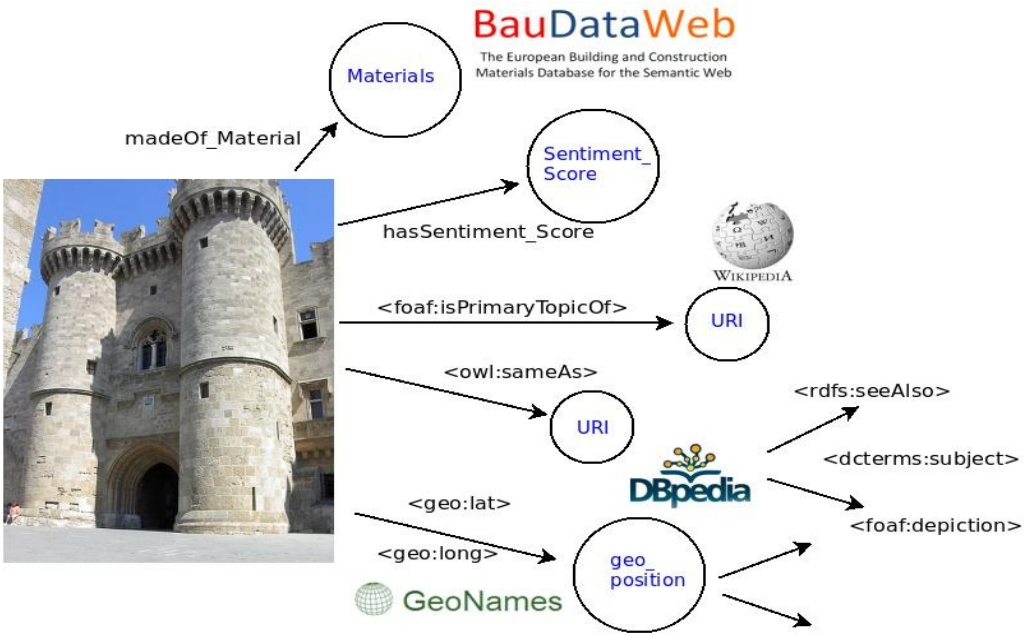

Figure 1

For example, Figure 1 depicts how different datasets can be linked together in order to create a richer knowledge base that can cater to extended queries. Now, let us assume that the building in the figure undergoes renovation and the materials used are changed, or the Wikipedia page link is revised. Such evolutionary changes will need to be handled constantly, in order to maintain persistent correctness of the links.

LOD is a distributed knowledge graph where almost any entity in any dataset can be related to any other entity in another dataset. While this is a huge advantage on the one hand, it gives rise to many challenges with respect to evolutionary changes. Due to the inherent nature of linkage in the LOD cloud, changes with respect to one part of the LOD graph are propagated throughout the graph. Hence, measuring the impact of a change in one dataset (entity) on other datasets (entities) within the LOD graph is crucial, where impact might be affected by, for instance, distance in the graph (eg. path length), the nature of the change and the dynamics of the content of each dataset. Therefore, one of the challenges resulting from the evolution of LOD is to ensure that there is little or no negative affect as a result of evolution with respect to the links

1 http://datahub.io/organization/lodcloud | http://lod-cloud.net/state

entailed. Ensuring persistent correctness of computed links and relations and improving the consistency of LOD as an actual data graph over time, are important aspects of research.

Emerging from the challenges mentioned above, is a need for preservation strategies that can handle the evolutionary aspects of the linked datasets. The recurring link computation and graph archival for dynamic datasets and the frequency of such computations are important to investigate.

M.Sc./Diploma thesis project

L3S Research Center, under supervision of Prof. Dr. Wolfgang Nejdl offers a M.Sc. or Diploma thesis in the scope of the DURAARK project. Taking into account the research context described above, the aim of the thesis is to provide methods that can cater to handling evolution in LOD. Exploring the evolution of entities by leveraging LOD datasets and external vocabularies is within the scope of this work. Measuring the impact and relevance of evolutionary changes for specific entities, factoring in semantic relatedness in the LOD graph, and designing preservation strategies dependent on dataset dynamics will be a significant outcome of this thesis project.

Are you interested in working with information extraction, semantic web and search technologies? Want to be part of an international research team working with a new exciting research topic?
The research tasks would entail the following:

- Research state of the art in LOD evolution
- Entity extraction and monitoring
- Develop scalable methods to measure the impact of evolutionary changes in the LOD graph
- Investigate which datasets are important to preserve (only direct links or also distant neighbours) ?
- Create knowledge about the interdependencies between datasets
- Judge to what extent existing links might have to be recomputed after datasets change
- Simple linking (via archiving) for static datasets (eg. historic statistics in *data.gov.uk*)
- Explore entity evolution by leveraging the LOD cloud

You should be:

- An interested and motivated worker with a keen will to learn
- Familiar with information extraction, data modelling, Semantic Web and Linked Data
- Familiar with graph-based data representation
- Familiar with programming languages (ideally Python, Java)

Are you interested or have questions? Contact us:
Dr. Stefan Dietze, Email: dietze@l3s.de,
Ujwal Gadiraju, Email: gadiraju@l3s.de,
Forschungszentrum L3S, Appelstr. 9a, 30167 Hannover