Title: Pushing the Limits of Instance Matching Systems: A Semantics-Aware Benchmark for Linked Data

Abstract: With the continuously increasing number of datasets published in the Web of Data and form part of the Linked Open Data Cloud, it becomes more and more essential to identify resources that correspond to the same real world object in order to interlink them and set the basis for large-scale data integration. This requirement becomes apparent in a multitude of domains ranging from science (marine research, biology, astronomy, pharmacology) to semantic publishing and cultural domains. In this context, instance matching is of crucial importance. Choosing the right framework for this purpose remains tedious, as current instance matching benchmarks fail to provide end users and developers with the necessary insights pertaining to how current frameworks behave when dealing with real data. It is though essential at this point to develop, along with data integration systems, benchmarks to determine the weak and strong points of those systems, as well as their overall quality in order to support users in deciding the system to use for their needs. Hence, well defined, and good quality benchmarks are important for comparing the performance of the developed instance matching systems.

In this talk, we will present SPIMBench, the Semantic Publishing Instance Matching Benchmark (SPIMBench) which allows the benchmarking of instance matching systems against not only structural and value heterogeneities, but also against semantics-aware ones involving complex OWL axioms. SPIMBench features a scalable data generator and a weighted gold standard that can be used for debugging instance matching systems and for reporting how well the systems perform for various matching tasks.

SPIMBench\(^1\) is one of the benchmarks developed in the context of the LDBC Project, a Collaborative FP7-STREP project that ended on March 2015. The main scientific and technical objective of LDBC was the development of new benchmarks that will spur research & industry progress in large-scale graph and RDF data management together with the establishment, by a broad community of researchers and, RDF and graph database vendors, of the LDBC Council. The Linked Data Benchmark Council (LDBC)\(^2\) is established as an independent authority responsible for specifying benchmarks, benchmarking procedures and verifying/publishing results for software systems designed to manage graph and RDF data.

Speaker's Bio: Dr. Irini Fundulaki is Principal Researcher of the Information Systems Laboratory of the Institute of Computer Science, FORTH and the Head of the W3C Greece Office. She is also the chair of the ERCIM Open Data Working and represents FORTH in the Board of Directors of LDBC Council. Her research interests are in the area of Linked Data Management and more specifically, on Models, Storage and Indexing Schemes for RDF data provenance and access control, Scalable RDF Query Processing, and on RDF Benchmarks with a focus on instance matching and advanced reasoning. She has published a number of scientific articles that have been widely cited and she has served on the Program Committees of numerous international conferences, journals and workshops. She was one of the organizers of the Benchmarking RDF Systems workshops and one of the tutors of the European Semantic Web Summer School (ESWC).
