Pushing the limits of Instance Matching Systems: A Semantics-Aware Benchmark

Irini Fundulaki  
*Institute of Computer Science – FORTH*

Joint work with  
G. Flouris, E. Daskalaki, T. Saveta  
*Institute of Computer Science – FORTH, Greece*  
and  
Axel Ngonga-Ngomo  
*University of Leipzig, Germany*  
and  
Melanie Herschel, Germany  
*University of Stuttgart*
Linked Data - The LOD Cloud
Linked Data – The LOD Cloud

Same entity can be found in different sources with different names

*Adapted from Suchanek & Weikum tutorial@SIGMOD 2013
Different Descriptions of the same Entity in Different Sources

"Athens, Greece description in GeoNames"

"Athens, Greece description in DBpedia"
Instance Matching: the cornerstone for Linked Data

How can we automatically recognize multiple appearances of the same entity across or within sources?

Social and Open Data

Data Evolution

Data Acquisition

Data Integration
Instance Matching for Linked Data

- RDF triples
- Sparse data
- Many sources to match
- Rich semantics
- Value, Structure Semantic heterogeneities

*Adapted from Suchanek & Weikum tutorial@SIGMOD 2013*
Instance Matching

• People **interconnect** their dataset with existing ones with links that are often **manually curated** (or semi-automatically generated).

• Size and number of data sets is **huge**, so it is vital to automatically detect additional links: **making the LOD graph more dense**.

• *Instance matching research* has led to the development of various systems.
  – How to compare these?
  – How can we assess their performance?
  – How can we push the systems to get better?

**Instance Matching Benchmarks**
Instance Matching Benchmarks

• Benchmarks are generally a set of tests to assess the performance of systems
• Make competing products comparable
• Accelerate progress both for academia and vendors and make technology viable
• Instance Matching Benchmark Ingredients
  • Datasets that will be matched by the systems under test
  • Gold Standard/Ground Truth/Reference Alignment is the “correct answer sheet” used to judge the completeness and soundness of the instance matching algorithms
  • Test Cases that address the ability of the systems to address different heterogeneities in data and structure
  • Metrics that assess the performance of the systems.
Semantic Publishing Instance Matching Benchmark (SPIMBench)

• Developed in the context of the European FP7 STREP Project LDBC (October 2012 – March 2015)
• Builds on LDBC’s Semantic Publishing Benchmark (SPB)
• Supports scalable data generation in the order of billion of triples
• Follows the distribution parameters of SPB for producing the datasets to be matched

➢ Supports value and structural transformations of the state of the art SWING platform
➢ Introduces and implements semantics-aware transformations for expressive OWL constructs
➢ Provides a Weighted Gold Standard that can be used to debug instance matching systems
➢ Proposes average similarity score & standard deviation to quantify the difficulty of the matches discovered by an Instance Matching system in addition to the standard metrics
SPIMBench: Transformations

- **Lexical:** modifications of data type property values (instance level)
- **Structural:** modifications of the structure of properties (instance level)
- **Semantics-aware:** focus on complex OWL schema axioms
  - class & property equivalence
  - instance (in)equality (same, different from)
  - class & property disjointness
  - class & property hierarchies
  - property constraints (functional, inverse functional)
  - complex class definitions (union, intersection)
- **Combinations of Transformations**
  - **Simple:** applied on different triples of the same resource
  - **Complex:** applied on the same triple of the same resource
SPIMBench Architecture

Download from: https://github.com/jsaveta/SPIMBench
SPIMBench Weights

- Follow an information-theoretical approach to compute the weight of \((u_i, u'_i)\) by measuring the information loss that results from applying transformations from \(u_i\) to \(u'_i\)

- apply multi-relational learning (MRL) \(\mathcal{L}\) to the input knowledge base \(\mathcal{K}\) and the transformed knowledge base \(\mathcal{K}'\) using RESCAL

\[
\text{final weight} = 1 - \sum (T_0 \ldots T_i)
\]
Conclusions

- Domain specific benchmark aims at testing Instance Matching Systems for Linked Data
- Produces large datasets, can be used for testing the ability of Instance Matching Systems to scale
- Highly configurable and produces test cases of varying difficulty
- First benchmark to address complex OWL constructs
- Will be used as a benchmark for the Ontology Alignment Evaluation Initiative 2015.
- Can be used to check the whether systems address in a satisfactory manner challenges introduced by RDA initiatives.
- Available at: https://github.com/jsaveta/SPIMBench
Linked Data Benchmark Council (LDBC)

• An independent authority responsible for specifying benchmarks, benchmarking procedures and verifying/publishing results for software systems designed to manage graph and RDF data.

• Outcome of the LDBC EU project (http://ldbc.eu) and was founded by Graph and RDF Database Vendors (OpenLink Software, Ontotext AD, Neo Technology, Sparsity technologies), Research Institutes (STI International, FORTH) and Individuals (Peter Boncz)

• LDBC aims to make insightful the critical properties of graph and RDF data management technology

• Information at http://ldbcouncil.org