Learningtree
A Generic Benchmark Generator for Linked Data
T. Saveta¹, E. Daskalaki¹, G. Flouris¹, J. Fundulaki¹, A.-C. Ngonga Ngomo²
¹ FORTH-ICS, ² University of Leipzig
(jsaveta, eva, fgeo, fundul)@ics.forth.gr, ngonga@informatik.uni-leipzig.de

Motivation
The widespread adoption of Semantic Web Technologies and the publication of large interrelated RDF datasets and ontologies in the Web has made the integration of data a crucial task. Data linking in this context is essential in order to provide a view of the underlying information; this is achieved by instance and schema matching techniques. To aid the users to choose among the systems that perform such tasks, a number of benchmarks have been developed.

Demonstration
LANCE
A novel instance matching benchmark generator for assessing instance matching techniques for RDF data with an associated schema.

LANCE approach
LANCE⁵ is a flexible, generic and domain-independent benchmark generator which takes into consideration RDFS and OWL constructs in order to evaluate instance matching systems. LANCE supports:
• Semantics-aware transformations
• Standard value and structure based transformations⁶,⁷
• Weighted gold standard based on tensor factorization
• Varying degrees of difficulty and fine-grained evaluation metrics

LANCE++ will be the next version of LANCE to support streaming data as well as spatiotemporal data.

Transformations-based Test Cases

Value-based
• Blank Character Addition/Deletion
• Random Character Addition/Deletion/Modification
• Token Addition/Deletion/Shuffle
• Country & Simple Abbreviation
• Date Format
• Synonym/Antonym
• Stem of a Word
• Multilingual

Structure-based
• Property Addition/Deletion • Property Aggregation • Property Extraction

Combination of transformations
More than one transformation types per instance.

Simple (SC): One transformation per triple.
Complex (CC): Combination of two transformations per triple (value-based and structure-based or value-based and semantics-aware).

Applicability
• Evaluated LogMap⁹, LIMES¹⁰ running the EAGLE¹¹ algorithm and OT⁰¹
• Entire source dataset transformed

Comments:
LogMap responds well to the value-based test cases but fails to find matches when the instance is involved in semantics-aware test cases. EAGLE and OT: give very good precision results for the value-based test cases but are not able to find matches in the remaining categories.

References