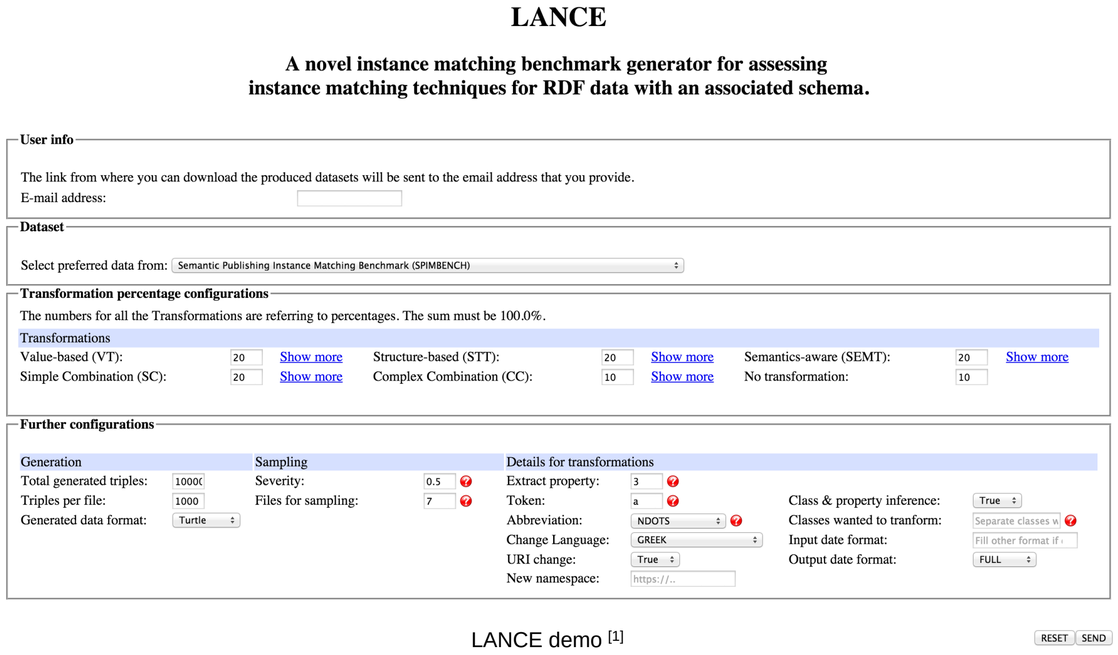
LANCE
A Generic Benchmark Generator for Linked Data

T. Saveta[1], E. Daskalaki[1], G. Flouris[1], I. Fundulaki[1], A.-C. Ngonga Ngomo[2]
[#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 an integrated 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

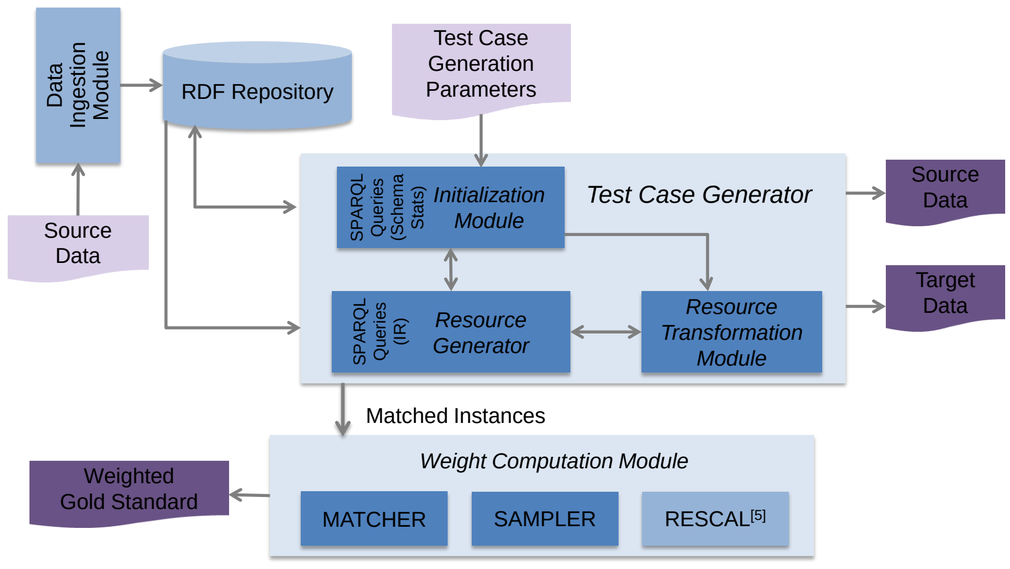


LANCE demo [1]

LANCE Approach

LANCE[2] 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 transformations[3,4]
- 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.



LANCE architecture

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
- Multilinguality

Structure-based

- Property Addition/Deletion
- Property Aggregation
- Property Extraction

Combination of transformations

More than one transformation types per instance.

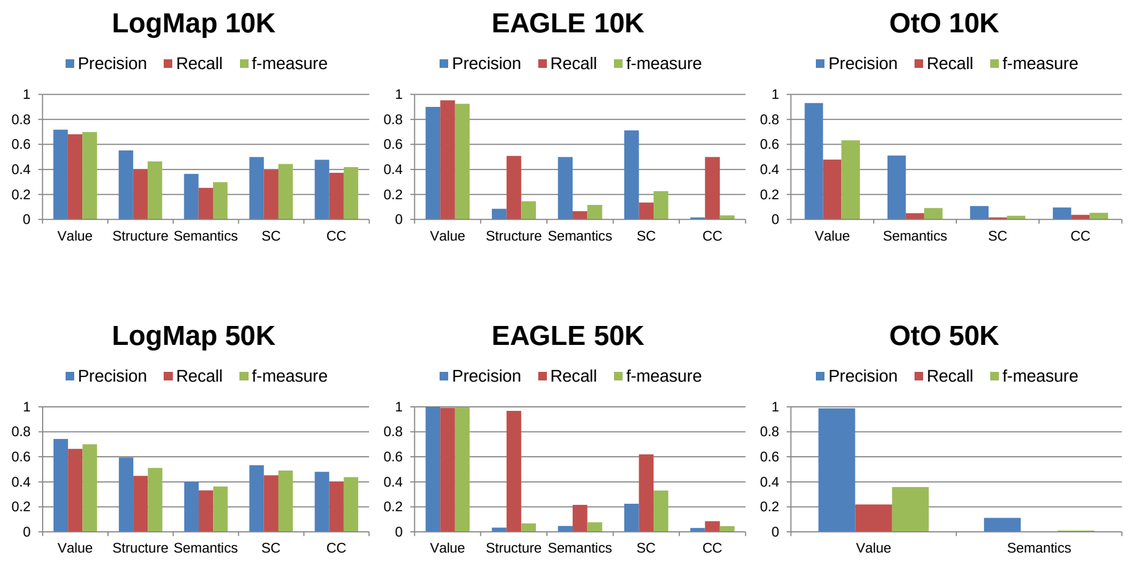
Simple (SC): One transformation per triple.

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

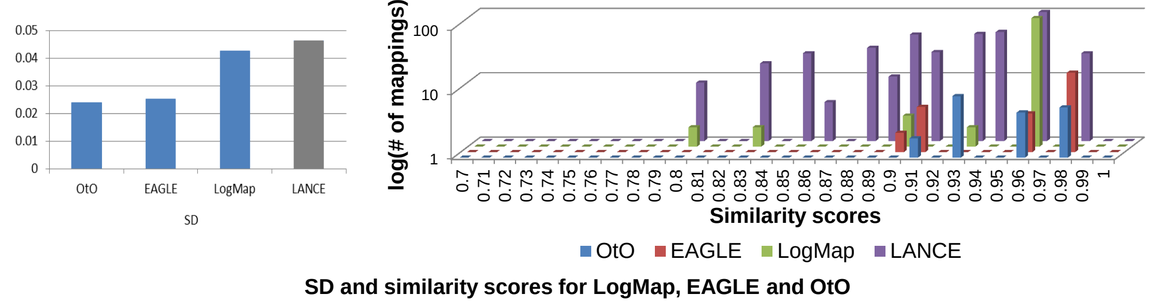
RDFS/OWL	SD	TD	SCHEMA TRIPLES	GS
owl:sameAs	(u_1, rdf:type, C) (u_2, rdf:type, C)	(u_1', rdf:type, C) (u_2', rdf:type, C) (u_1', owl:sameAs, u_2')		$u_1 \sim u_1$' $u_1 \sim u_2$' $u_2 \sim u_2$' $u_2 \sim u_1$'
owl:differentFrom	(u_1, rdf:type, C)	(u_1', rdf:type, C) (u_1'', rdf:type, C) (u_1', owl:differentFrom, u_1'')		$u_1 \sim u_1$'
owl:equivalentClass	(u_1, rdf:type, C)	(u_1', rdf:type, C')	(C, owl:equivalentClass, C')	$u_1 \sim u_1$'
owl:disjointWith	(u_1, rdf:type, C)	(u_1', rdf:type, C')	(C, owl:disjointWith, C')	
owl:FunctionalProperty	(u_1, rdf:type, C) (u_1, p_1, o_1)	(u_1, rdf:type, C) (u_1, p_1, o_2)	(p_1, rdf:type, owl:FunctionalProperty)	$o_1 \sim o_2$
owl:InverseFunctionalProperty	(u_1, rdf:type, C) (u_1, p_1, o_1)	(u_1', rdf:type, C) (o_1, p_1, u_1')	(p_1, rdf:type, owl:InverseFunctionalProperty)	$u_1 \sim u_1$'
owl:unionOf	(u_1, rdf:type, C)	(u_1', rdf:type, C')	(C', owl:unionOf, {C_0, C_1,..})	$u_1 \sim u_1$'
owl:intersectionOf	(u_1, rdf:type, C)	(u_1', rdf:type, C')	C owl: intersectionOf{C,D,E,F} C' owl: intersectionOf {C,D}	$u_1 \sim u_1$'

Applicability

- Evaluated LogMap[6], LIMES[7] running the EAGLE[8] algorithm and OtO[9]
- Entire source dataset transformed



Applicability experiments for LogMap, EAGLE and OtO



SD and similarity scores for LogMap, EAGLE and OtO

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 OtO: give very good precision results for the value-based test cases but are not able to find matches in the remaining categories.

Comments:
LogMap: can address the “difficult” test cases.
EAGLE and OtO: cannot address the challenges imposed by the, harder, semantics-aware test cases.

LANCE is able to determine the capabilities of the IM systems and also reflect the difficulty of the test cases through the weighted gold standard.

This work was partially supported by the projects
HOBBIT (H2020 #688227) and LDBC (FP7-ICT-2011-8 #317548).

This poster is a companion to the ISWC demo [1] and research paper [2].

References

[1] T. Saveta, E. Daskalaki, G. Flouris, I. Fundulaki and A.-C. Ngonga Ngomo. LANCE: A Generic Benchmark Generator for Linked Data . In ISWC, 2015. Find LANCE demo here: http://tinyurl.com/pvex9hu
[2] T. Saveta, E. Daskalaki, G. Flouris, I. Fundulaki, M. Herschel and A.-C. Ngonga Ngomo. LANCE: Piercing to the Heart of Instance Matching Tools. In ISWC, 2015.
[3] A. Ferrara, D. Lorusso, S. Montanelli, and G. Varese. Towards a Benchmark for Instance Matching. In OM, 2008.
[4] A. Ferrara, S. Montanelli, J. Noessner, and H. Stuckenschmidt. Benchmarking Matching Applications on the Semantic Web. In ESWC, 2011.
[5] Maximilian Nickel, and Volker Tresp. Tensor Factorization for Multi-relational Learning. ECML/PKDD 3, volume 8190 of Lecture Notes in Computer Science, page 617-621. Springer, 2013.
[6] E. Jimenez-Ruiz and B. C. Grau. Logmap: Logic-based and scalable ontology matching. In ISWC, 2011.
[7] A.-C. Ngonga Ngomo and Soren Auer. LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In IJCAI, 2011.
[8] A.-C. Ngonga Ngomo and K. Lyko. EAGLE: Efficient Active Learning of Link Specifications using Genetic Programming. In ESWC, 2012.
[9] E. Daskalaki and D. Plexousakis. OtO Matching System: A Multi-strategy Approach to Instance Matching. In CAiSE, 2012.