Similarity and Delta for Semantic Web Graphs

Krishnamurthy Koduvayur Viswanathan

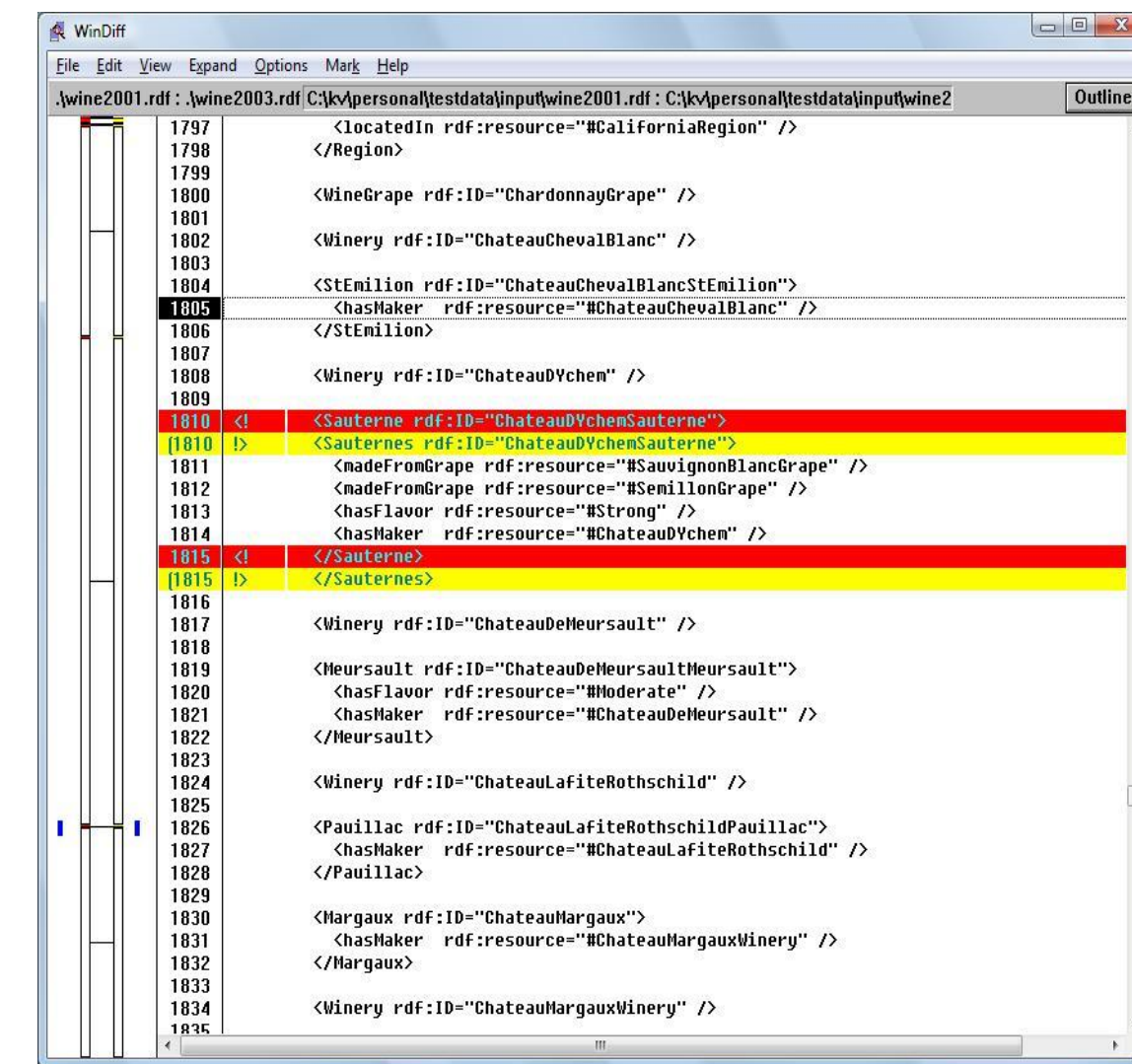**UMBC** — AN HONORS UNIVERSITY IN MARYLAND
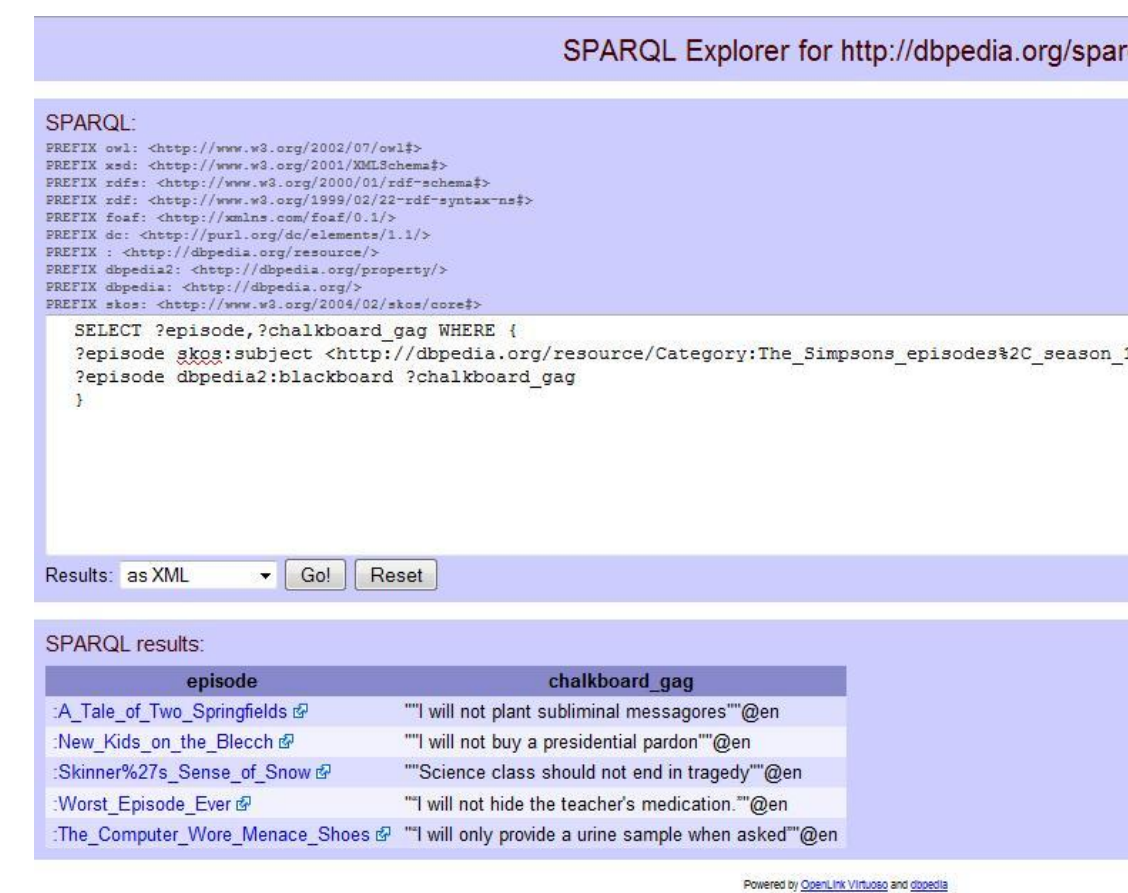
ebiquity group

Contributions

- Detect pairs of similar semantic web graphs and versioning relationships between ontologies
- Generate a delta between successive ontology versions that have been detected

Similarity of SW Graphs

- Identical copies
- Different Base URIs
- Minor textual differences
- Same structure, but different textual content

- Same SWD expressed in different formats

```
# The N-Triples statements below are equivalent to this RDF/XML:
#
# <rdf:RDF xmlns="http://xmlns.com/foaf/0.1/"
#          xmlns:dc="http://purl.org/dc/terms/"
#          xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
#   <Document rdf:about="http://www.w3.org/2001/sw/RDFCore/ntriples/">
#     <dc:title>N-Triples</dc:title>
#     <maker>
#       <Person rdf:nodeID="art">
#         <name>Art Barstow</name>
#       </Person>
#     </maker>
#     <maker>
#       <Person rdf:nodeID="dave">
#         <name>Dave Beckett</name>
#       </Person>
#     </maker>
#   </Document>
# </rdf:RDF>

<http://www.w3.org/2001/sw/RDFCore/ntriples/> <http://www.w3.org/1999/02/22-rdf-syntax-ns#> <http://xmlns.com/foaf/0.1/Document> .
<http://www.w3.org/2001/sw/RDFCore/ntriples/> <http://purl.org/dc/terms/title> "N-Triples" .
<http://www.w3.org/2001/sw/RDFCore/ntriples/> <http://xmlns.com/foaf/0.1/maker> _:art .
<http://www.w3.org/2001/sw/RDFCore/ntriples/> <http://xmlns.com/foaf/0.1/maker> _:dave .
_:art <http://www.w3.org/1999/02/22-rdf-syntax-ns#> <http://xmlns.com/foaf/0.1/Person> .
_:art <http://xmlns.com/foaf/0.1/name> "Art Barstow" .
_:dave <http://www.w3.org/1999/02/22-rdf-syntax-ns#> <http://xmlns.com/foaf/0.1/Person> .
_:dave <http://xmlns.com/foaf/0.1/name> "Dave Beckett" .
```

- Ontology versions

- SW graphs generated by results of a standing SPARQL query executed at different times on a dataset

Work Flow

Input: Corpus of SWDs → Convert to n-triples format → Convert to canonical form → Create Reduced Forms → Compute Text-Based Similarity Metrics → Identify pairs of similar documents → Identify ontology versions → Generate delta between versions

SW Graph Canonicalization

- Assigns uniform identifiers to blank nodes
- Provides a deterministic order to statements
- Empirical method that works for most examples

document1.nt (input)	canonicalized document1.nt (output)
<person:John> <a:livesIn> _:x .	_:g2 <a:hasCapital> _:g1 .
_:x <a:IsPartOf> "USA" .	_:g2 <a:IsPartOf> "USA" .
<person:John> <a:likes> "cheese" .	<person:John> <a:likes> "cheese" .
_:x <a:hasCapital> _:y .	<person:John> <a:livesIn> _:g2 .

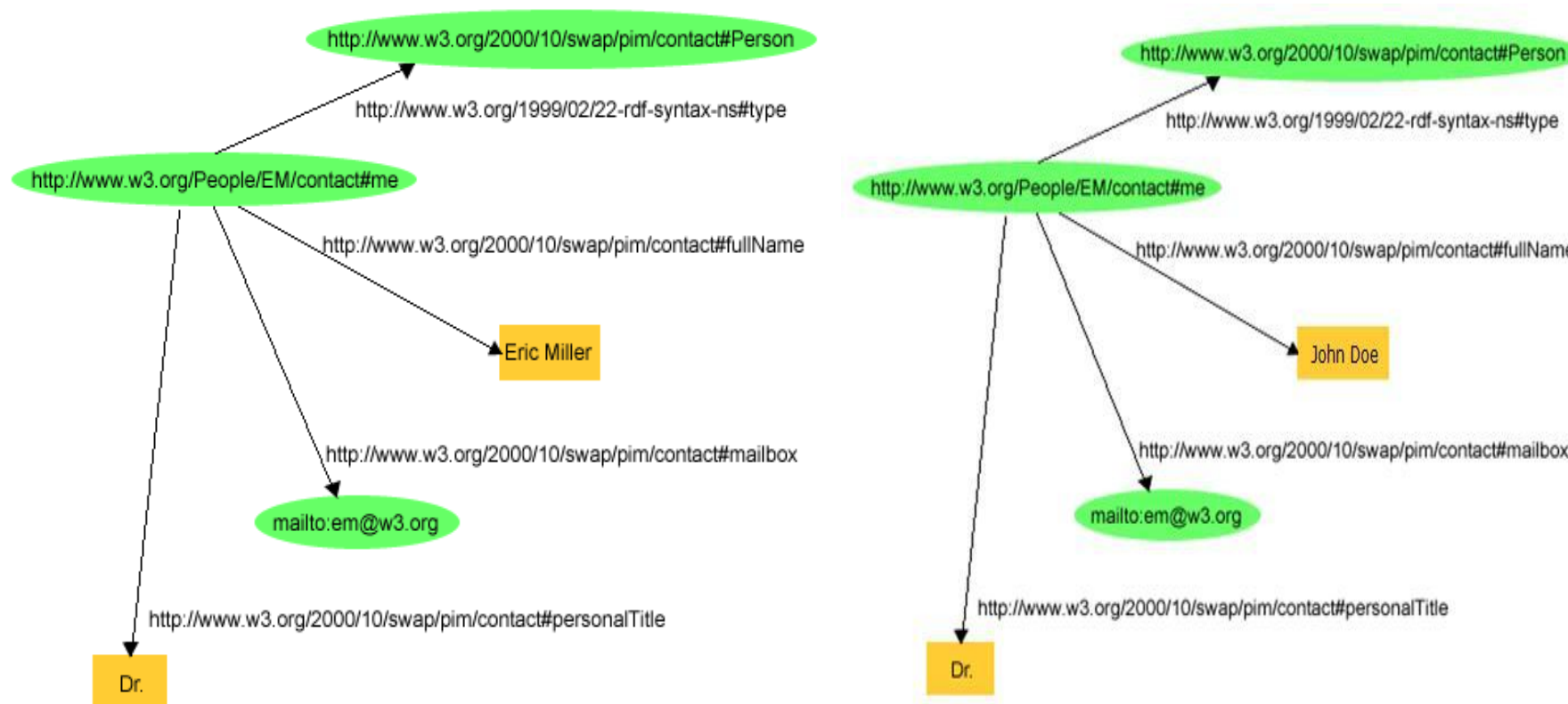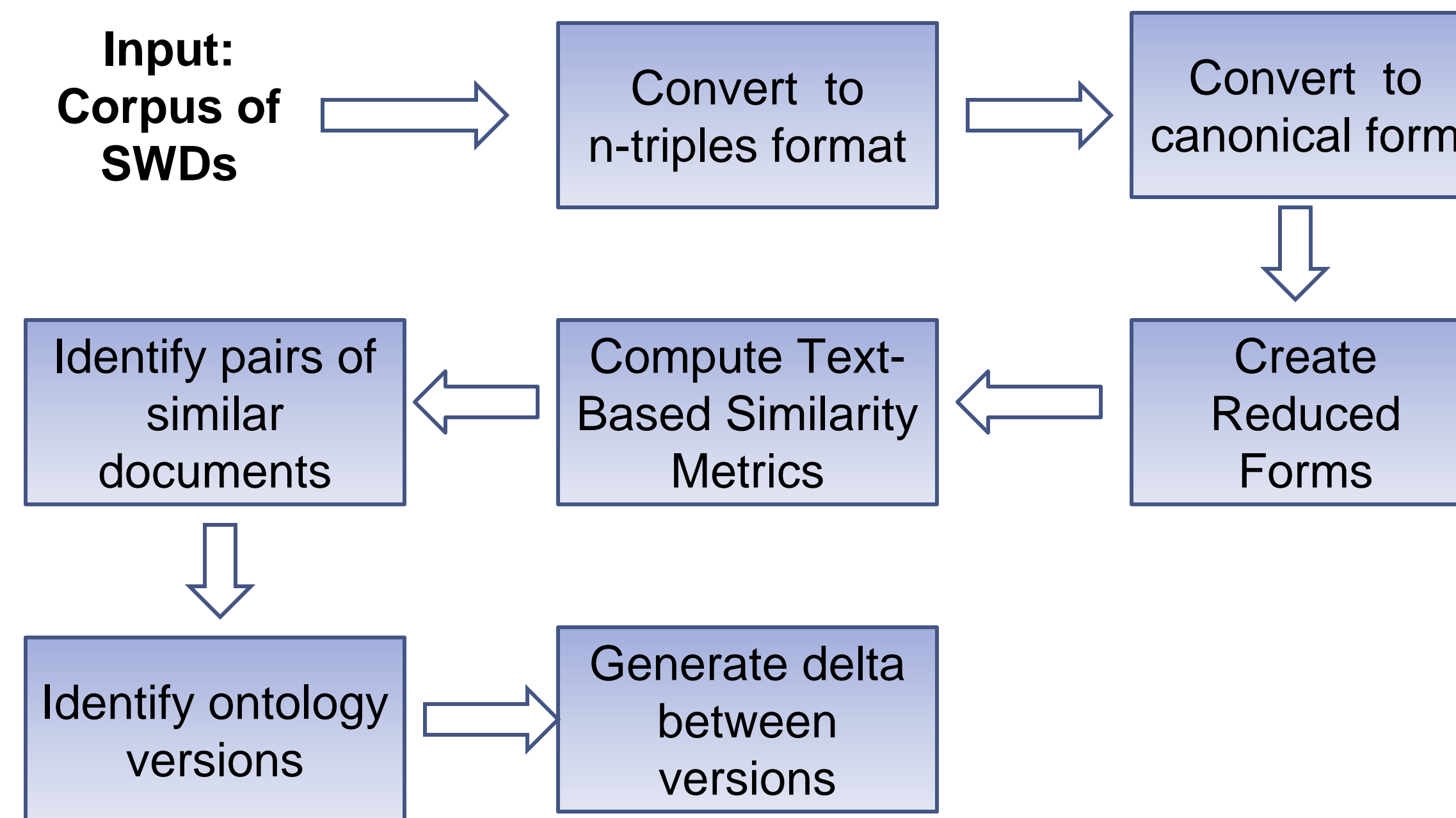| document2.nt (input)	canonicalized document2.nt (output)
_:a <a:hasCapital> _:b .	_:g2 <a:hasCapital> _:g1 .
<person:John> <a:livesIn> _:a .	_:g2 <a:IsPartOf> "USA" .
_:a <a:IsPartOf> "USA" .	<person:John> <a:likes> "cheese" .
<person:John> <a:likes> "cheese" .	<person:John> <a:livesIn> _:g2 .

Reduced Forms

- Only the literals from the original n-triples file
- All content except the literals from the original n-triples file
- The base-URI of every node replaced by the empty string
- All the literals and the base-URI of every node replaced by the empty string

Generating Deltas

- Compute the deductive closure of documents before comparing them
- Describe the deltas as the smallest set of atomic triples
- Canonicalization smoothens most disparities amongst statements in similar graphs
- Compare only local names of entities in the graph, i.e. ignore the global namespaces

Preliminary Results

- Preliminary dataset of 8300 triples, across 23 RDF graphs (ontologies and data)
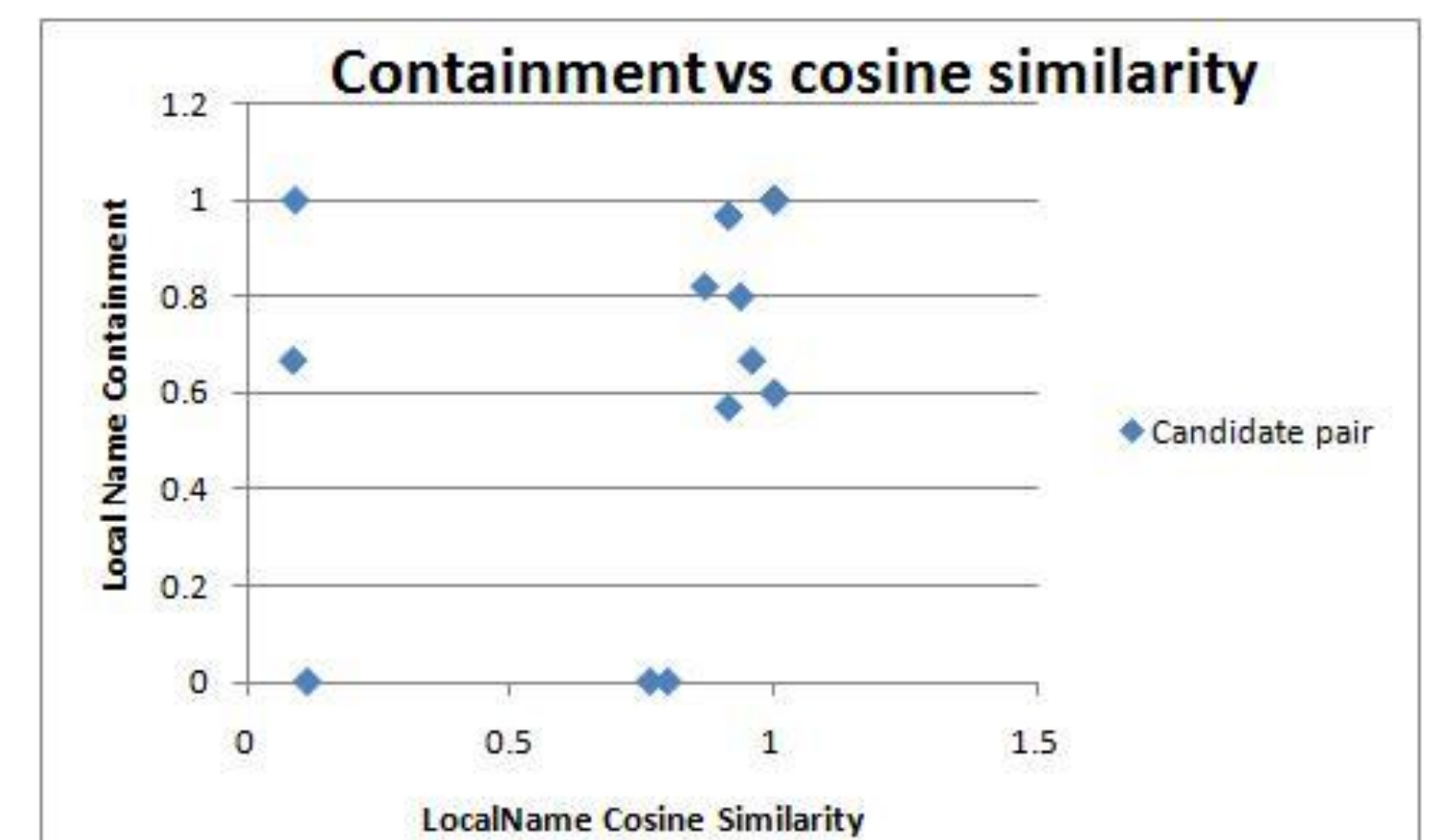
- 17 different combinations of similarity metrics are generated
- The pairs on the top right of the graph identify ontology versions

In collaboration with Dr. Tim Finin