On Linking Heterogeneous Dataset Collections
Mayank Kejriwal1 and Daniel P. Miranker1

1 {kejriwal,miranker}@cs.utexas.edu / The University of Texas at Austin

Heterogeneous Entity Resolution (ER)
Logically equivalent entities may occur in several syntactically different forms [1]

![Entity Resolution Table]

Dataset 1
Emergency Contact
<table>
<thead>
<tr>
<th>Name</th>
<th>Contact</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mickey Beats</td>
<td>Joan Beats</td>
<td>Spouse</td>
</tr>
<tr>
<td>Susan Palermo</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>Samuel Crae</td>
<td>Joan Beats</td>
<td>Sister</td>
</tr>
</tbody>
</table>

Dataset 2

Dataset 3

Standard two-step ER workflow
Accepts (at most) a pair of datasets as input, typically of the same data model [3], [5]

Does this approach suffice for a heterogeneous dataset collection?

We propose...
A system that enables scalable linking of entities in a pair of RDF/Relational dataset collections

Dataset Mapping
- High-level algorithm:
  - Devise inexpensive way of scoring two datasets
  - Build matrix of scores, run Hungarian algorithm
  - Use confidence strategy to prune mappings

Example: Dominating Strategy
\[ O(N^2) \]

Preliminary Experimental Results: Blocking

Datasets

COVID/Constitue (Columbia)
- Dataset A: 1.2 million, 2.76 million
- Dataset B: 1.58 million, 2.76 million

COVID/Constitue (Venezuela)
- Dataset A: 2.3 million, 4.55 million

COVID/Constitue (Brazil)
- Dataset A: 2.4 million, 4.55 million

COVID/Constitue (Canada)
- Dataset A: 2.4 million, 4.55 million

Conclusion
Dataset mapping is a viable preprocessing step and experimentally, savings increase as the collection sizes increase. With more testing, and in conjunction with the property table, it can be used to support linkage of heterogeneous collections across Linked Open Data and the Deep Web.

Acknowledgements
We would like to thank the National Science Foundation (NSF) for supporting the research and this trip. We also acknowledge Juan Sequeda for the Constitute datasets.

References

Updated October 9, 2014