Populating the Semantic Web – Combining Text and Relational Databases as RDF Graphs

Kate Byrne, ICCS
Supervisors: Prof Ewan Klein, Dr Claire Grover

20th March 2009
Outline

Motivation and Research Questions

Overview of the “Tether” System
RDB2RDF
incorporating domain thesauri

Text to RDF Pipeline
NER and RE
mapping to RDF

Results and Conclusions
Motivation

1. Information retrieval
   - knowledge locked in text documents

2. Populating the Semantic Web
   - accumulated wisdom at risk of being side-lined

3. Data management
   - integrating the silos
   - flexible presentation
Motivation

1. Information retrieval
   - knowledge locked in text documents

2. Populating the Semantic Web
   - accumulated wisdom at risk of being side-lined

3. Data management
   - integrating the silos
   - flexible presentation
Motivation

1. Information retrieval
   • knowledge locked in text documents

2. Populating the Semantic Web
   • accumulated wisdom at risk of being side-lined

3. Data management
   • integrating the silos
   • flexible presentation
Research Questions

1. How should RDB data be converted to RDF?
2. How does query performance over RDB and RDF compare?
3. How well do NER and RE tools perform in combination?
4. Does including text relations improve retrieval?
5. Will the data become contaminated with false statements?
6. Should data curators be investing in this technology?
Overview of Tether

Relational database

Published domain thesauri

Text documents

txt2rdf pipeline

Graph of triples

Relational Text documents of triples

Graph

Graph

Published domain thesauri

Graph

Graph
Data from RCAHMS

- The “memory keeper” for Scotland
- http://www.rcahms.gov.uk/
- One of Scotland’s 6 National Collections

- Recording Scotland’s places, from the Neolithic to Now:
  - Skara Brae
  - Informatics Forum
RDB2RDF – Relational Database to RDF Graph

Converting relational data to RDF is straightforward.
RDB2RDF – Relational Database to RDF Graph

Converting relational data to RDF is straightforward.
Basic RDB2RDF Procedure
“Table as Class; Column as Predicate” Conversion

<table>
<thead>
<tr>
<th>siteNo</th>
<th>name</th>
<th>parish</th>
<th>classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dirleton Castle</td>
<td>Dirleton</td>
<td>defence</td>
</tr>
<tr>
<td>2</td>
<td>Dirleton Cottage</td>
<td>Dirleton</td>
<td>residential</td>
</tr>
<tr>
<td>3</td>
<td>Drem Airfield</td>
<td>Dirleton</td>
<td>military</td>
</tr>
<tr>
<td>4</td>
<td>Jamie’s Neuk</td>
<td>Dirleton</td>
<td>military</td>
</tr>
</tbody>
</table>

@prefix : <http://www.ltg.ed.ac.uk/tether/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

• Each row (instance): central node with cluster of attributes
Problems with Basic Conversion

- Set of 12 guidelines proposed, including:
  - redundant triples at relational joins
  - handling RDB metadata and URI generation
  - bnodes, RDB null fields, coded values
- Example – use of literals
Literals or Resources?

```
rdfs:Class
  rdf:type
  :site
  rdf:type

  SiteNo#1
  :classification
  "defence"
  :parish
  "Dirleton"

  :dirletonCastle
  :location
  "East Lothian"

"Dirleton Castle"
"Dirleton"
"East Lothian"
```
Literals or Resources?
Literals or Resources?

```
rdfs:Class :site
:classification "defence"
:name "Dirleton"
:parish "defence"
:classification  

:dirletonCastle rdf:type rdfs:Class
:name "Dirleton Castle"
:parish "defence"
:classification :SiteNo#1
"Dirleton Castle"

rdfs:label :dirletonCastle "East Lothian"
:name
"Dirleton"
:parish"defence"
:classification
:SiteNo#1
"Dirleton Castle"
```
Incorporating Domain Thesauri

• Plenty available and more coming
  • terminology for site and object classifications
• RDF makes integration easy
• Graph provides natural representation for hierarchy
• Important aid for non-expert retrieval
Grounding Against Thesaurus

A Neolithic burial monument comprising a stone-built chamber within a mound of stones.

"CHAMBERED CAIRN"

"PASSAGE GRAVE"

"HEEL CAIRN"

"STALL CAIRN"

SITEID#SITE123:CLASSN/SITETYPE#CHAMBERED CAIRN

HASCLASSN:CLASSN/SITETYPE#CHAMBERED+CAIRN
Grounding Against Thesaurus

"A Neolithic burial monument comprising a stone–built chamber within a mound of stones."

"CHAMBERED CAIRN"

"Heel Cairn"

"Stalled Cairn"

"Passage Grave"

"CHAMBERED TOMB"

"RING CAIRN"

"SQUARE CAIRN"

"PASSAGE GRAVE"

"STALL CAIRN"
NLP Work – *txt2rdf*

**Text documents**

**Pre-processing**
- Tokenise
- Sentence and para split
- POS tag

**Named Entity Recognition**
- Multi-word tokens and features
- List of NEs and classes
- Trained NER model

**Relation Extraction**
- Set of NE pairs and features
- List of relations and classes
- Trained RE model

**RDF**
- Translation

**Graph of triples**
Finding Binary Relations in Text

- NER as first step
- Special attention paid to NE nesting
- Then look for relations between pairs of NEs:
  - generate all possible pairings per document
  - add features –
    - *NE classes, word separation, POS tags, nesting, in sentence*...
  - syntactic clues limited as relations are inter-sentential
- Sequential tagger labels each pairing
Named Entity Recognition

- 11 categories:
  - ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT

- Unorthodox ones:
  - EVENT - verb phrases not noun phrases: visited, was found
  - SITETYPE, ARTEFACT, ROLE, EVENT – class terms

- Nesting:
  - [[[Edinburgh]^{PLACE} University]^{ORG} Library]^{ORG}
Named Entity Recognition

- 11 categories:
  ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT

- Unorthodox ones:
  - EVENT - verb phrases not noun phrases: *visited, was found*
  - SITETYPE, ARTEFACT, ROLE, EVENT – class terms

- Nesting:
  [[[Edinburgh]PLACE University]ORG Library]ORG
Named Entity Recognition

• 11 categories:
  ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT

• Unorthodox ones:
  • EVENT - verb phrases not noun phrases: visited, was found
  • SITETYPE, ARTEFACT, ROLE, EVENT – class terms

• Nesting:
  [[[Edinburgh]PLACE University]ORG Library]ORG
Named Entity Recognition

• 11 categories:
  ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT

• Unorthodox ones:
  • EVENT - verb phrases not noun phrases: visited, was found
  • SITETYPE, ARTEFACT, ROLE, EVENT – class terms

• Nesting:
  [[[Edinburgh]$PLACE University$ORG Library$ORG
Relation Extraction

- Basic predicate categories:
  - eventRel, hasLocation, hasPeriod, instanceOf, partOf, sameAs, seeAlso
- $n$-ary eventRel predicate gets split up:
  - eventAgent, eventAgentRole, eventDate, eventPatient, eventPlace
- Final RDF schema predicates for text relations:
  - hasEvent, hasAgent, hasAgentRole, hasPeriod, hasPatient, hasLocation, hasClassn, hasObject, partOf, owl:sameAs, rdfs:seeAlso
Mapping Text Relations to RDF

site456

[[SOUTH WALLS]] , [[MISBISTER]] , [[[THE {LOFTS}]]]

[ND38NW 29 centred 3325 8885]

Sites [[[recorded]]] during an [archaeological survey] undertaken on the lands of [[the {Loft}]] , [Longhope] , as part of the pilot scheme for the [[[Historic {Scotland}]]] {Farm} {Ancient} {Monument} {Survey Grant Scheme} . [ND 3311 8890] Two [small {cairns}] . [ND 3336 8889] {Cairn} . [ND 3339 8885] {Cairn} . [ND 3339 8886] {Clearance cairn} . [ND 3342 8884] {Sub-rectangular cairn} . [ND 3339 8883] {Well} Sponsors : {Historic {Scotland}} , {M J Jones} . {IN Card} {1998}
Mapping Text Relations to RDF

site456

{[SOUTH WALLS]} , {[MISBISTER]} , [[[THE {LOFTS}]]]

[ND38NW 29 centre 3325 8885]

Sites [[[Recorded]]] during an [archaeological survey] undertaken on the lands of [[the {Loft}], [Longhope], as part of the pilot scheme for the [[[Historic {Scotland}]]] [Farm] [Ancient] [Monument] [Survey Grant Scheme]. [ND 3311 8890] Two [small {cairns}], [ND 3336 8889] [Cairn], [ND 3339 8885] [Cairn], [ND 3339 8886] [Clearance cairn], [ND 3342 8884] [Sub-rectangular cairn], [ND 3339 8883] [Well]. Sponsors: [Historic {Scotland}], [M J Jones], [[[N Card]]] [1998]
Mapping Text Relations to RDF

site456

[[SOUTH WALLS]], {{MISBISTER}}, {{[[THE {LOFTS}]]]}}

[ND38NW 29 centred 3325 8885]

Sites [[[recorded]]] during an [[archaeological survey] undertaken on the lands of [[the {Loft}]], [Longhope], as part of the pilot scheme for the [[[Historic {Scotland}]]] {Farm} {Ancient} {Monument} Survey Grant Scheme]. [ND 3311 8890] Two [small {cairns}], [ND 3336 8889] {Cairn}, [ND 3339 8885] {Cairn}, [ND 3339 8886] {Clearance cairn}, [ND 3342 8884] [Sub-rectangular cairn]. [ND 3339 8883] {Well}, Sponsors: {Historic {Scotland}}, {{M J Jones}}, {{IN Card}} {1998}

site456 − hasEvent − recordingX
recordingX − hasLocation − "ND 3342 8884"
recordingX − hasPatient − "Sub-rectangular cairn"
Positioning Text Triples in the Graph

- “Bea Mill dates from the 19th century” [docid=3402]
- “Bea Mill”, “19th century”: standard (instance) NE mentions
Mapping Non-standard NEs – Class Terms

- “Site 101 was visited on 27 April 1969”
- “visited”: we need both class and particular instance
Mapping Non-standard NEs – Class Terms

- “Site 101 was visited on 27 April 1969”
- “visited”: we need both class and particular instance
# Results – NER Step

<table>
<thead>
<tr>
<th></th>
<th>Precision %</th>
<th>Recall %</th>
<th>F-score %</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADDRESS</td>
<td>82.40</td>
<td>81.61</td>
<td>82.00</td>
<td>3,458</td>
</tr>
<tr>
<td>PLACE</td>
<td>95.00</td>
<td>66.80</td>
<td>78.44</td>
<td>2,503</td>
</tr>
<tr>
<td>SITENAME</td>
<td>64.55</td>
<td>61.20</td>
<td>62.83</td>
<td>2,712</td>
</tr>
<tr>
<td>DATE</td>
<td>95.12</td>
<td>82.08</td>
<td>88.12</td>
<td>3,519</td>
</tr>
<tr>
<td>PERIOD</td>
<td>84.02</td>
<td>45.54</td>
<td>59.07</td>
<td>400</td>
</tr>
<tr>
<td>EVENT</td>
<td>94.98</td>
<td>63.66</td>
<td>76.22</td>
<td>3,176</td>
</tr>
<tr>
<td>ORG</td>
<td>99.39</td>
<td>89.66</td>
<td>94.27</td>
<td>2,730</td>
</tr>
<tr>
<td>PERSNAME</td>
<td>96.71</td>
<td>74.82</td>
<td>84.37</td>
<td>2,318</td>
</tr>
<tr>
<td>ROLE</td>
<td>98.00</td>
<td>54.44</td>
<td>70.00</td>
<td>90</td>
</tr>
<tr>
<td>SITETYPE</td>
<td>85.24</td>
<td>52.39</td>
<td>64.89</td>
<td>5,668</td>
</tr>
<tr>
<td>ARTEFACT</td>
<td>75.83</td>
<td>18.06</td>
<td>29.17</td>
<td>879</td>
</tr>
</tbody>
</table>
| Average | 88.02       | 67.75    | **76.57** | (27,453)
## Results – RE Step

<table>
<thead>
<tr>
<th>Relation</th>
<th>Precision %</th>
<th>Recall %</th>
<th>F-score %</th>
<th>Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>eventAgent</td>
<td>98.42</td>
<td>98.70</td>
<td>98.56</td>
<td>3,794</td>
</tr>
<tr>
<td>eventAgentRole</td>
<td>69.23</td>
<td>30.00</td>
<td>41.86</td>
<td>13</td>
</tr>
<tr>
<td>eventDate</td>
<td>98.75</td>
<td>98.68</td>
<td>98.71</td>
<td>3,189</td>
</tr>
<tr>
<td>eventPatient</td>
<td>87.77</td>
<td>84.61</td>
<td>86.16</td>
<td>1,553</td>
</tr>
<tr>
<td>eventPlace</td>
<td>83.58</td>
<td>72.70</td>
<td>77.76</td>
<td>341</td>
</tr>
<tr>
<td>hasLocation</td>
<td>83.26</td>
<td>83.00</td>
<td>83.13</td>
<td>5,085</td>
</tr>
<tr>
<td>hasPeriod</td>
<td>83.69</td>
<td>73.86</td>
<td>78.47</td>
<td>233</td>
</tr>
<tr>
<td>instanceOf</td>
<td>52.00</td>
<td>31.52</td>
<td>39.25</td>
<td>100</td>
</tr>
<tr>
<td>partOf</td>
<td>78.87</td>
<td>51.38</td>
<td>62.22</td>
<td>568</td>
</tr>
<tr>
<td>sameAs</td>
<td>68.69</td>
<td>44.55</td>
<td>54.05</td>
<td>6,934</td>
</tr>
<tr>
<td>seeAlso</td>
<td>50.00</td>
<td>19.68</td>
<td>28.24</td>
<td>122</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>83.41</strong></td>
<td><strong>69.27</strong></td>
<td><strong>75.68</strong></td>
<td><strong>21,932</strong></td>
</tr>
</tbody>
</table>
Evaluating the Complete *txt2rdf* Pipeline

1. Use NE model to tag test set
   - train on 90% of corpus, tag remaining 10%

2. Run RE model over “found” NE pairs
   - (train on same 90%)

3. Evaluate against the gold standard
   - “new” relation pairs are FPs
   - every gold relation missed counts as FN
   - big variation across corpus: measure performance range
### Results for Full Pipeline

<table>
<thead>
<tr>
<th></th>
<th>“Hardest” data</th>
<th>“Easiest” data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>eventAgent</td>
<td>94.91</td>
<td>68.33</td>
</tr>
<tr>
<td>eventAgentRole</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>eventDate</td>
<td>80.69</td>
<td>57.19</td>
</tr>
<tr>
<td>eventPatient</td>
<td>83.33</td>
<td>4.00</td>
</tr>
<tr>
<td>eventPlace</td>
<td>36.36</td>
<td>8.33</td>
</tr>
<tr>
<td>hasLocation</td>
<td>67.90</td>
<td>59.31</td>
</tr>
<tr>
<td>hasPeriod</td>
<td>83.33</td>
<td>11.90</td>
</tr>
<tr>
<td>instanceOf</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>partOf</td>
<td>15.79</td>
<td>6.82</td>
</tr>
<tr>
<td>sameAs</td>
<td>47.63</td>
<td>16.92</td>
</tr>
<tr>
<td>seeAlso</td>
<td>18.18</td>
<td>13.64</td>
</tr>
<tr>
<td>Average</td>
<td>63.55</td>
<td>31.32</td>
</tr>
</tbody>
</table>

F-score range: 41.96% – 73.06%. Average: **57.51%**

Average precision: 73.35%
1. Compare SPARQL over RDF with SQL over RDB
   - identical results (as expected)
   - RDB queries typically sub-second
   - RDF queries very variable: 0.9 sec to 7 minutes

2. Queries over RDF enhanced with text relations
   - queries that are impossible against RDB, such as:
     At which sites in Shetland have bones been found, when and by whom?
Querying the Final Graph

1. Compare SPARQL over RDF with SQL over RDB
   - identical results (as expected)
   - RDB queries typically sub-second
   - RDF queries very variable: 0.9 sec to 7 minutes

2. Queries over RDF enhanced with text relations
   - queries that are impossible against RDB, such as:
     *At which sites in Shetland have bones been found, when and by whom?*
Results for “Shetland sites with bones”

<table>
<thead>
<tr>
<th>site</th>
<th>sitename</th>
<th>date</th>
<th>agent</th>
<th>True?</th>
</tr>
</thead>
<tbody>
<tr>
<td>site32</td>
<td>UNST, UNDERHOULL</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>site78</td>
<td>YELL, PAPIL</td>
<td>1878</td>
<td>Ordnance...</td>
<td>Partly</td>
</tr>
<tr>
<td>site510</td>
<td>HILL OF URE</td>
<td>1858</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>site942</td>
<td>SOUTH VOXTER</td>
<td>1903</td>
<td></td>
<td>Partly</td>
</tr>
<tr>
<td>site976</td>
<td>KIRKHOULL</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>site1003</td>
<td>WESTER QUARFF</td>
<td>1903</td>
<td></td>
<td>Partly</td>
</tr>
<tr>
<td>site1006</td>
<td>THE CLUMPERS</td>
<td>1878</td>
<td></td>
<td>Partly</td>
</tr>
<tr>
<td>site1201</td>
<td>DALE</td>
<td>1875</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>site1383</td>
<td>YELL, KIRKABISTER</td>
<td>1878</td>
<td></td>
<td>Partly</td>
</tr>
<tr>
<td>site1385</td>
<td>YELL, SELLAFA...</td>
<td>1833</td>
<td></td>
<td>Partly</td>
</tr>
<tr>
<td>site1414</td>
<td>UYEA, WINNA NESS</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>site1415</td>
<td>UYEA, THE HALL</td>
<td>A.D. 1830</td>
<td>TI</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1900</td>
<td>TI</td>
<td>Partly</td>
</tr>
</tbody>
</table>
Research Questions Revisited

1. How should RDB data be converted to RDF?
   • 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   • Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   • 58% F-score

4. Does including text relations improve retrieval?
   • Yes! Useful new information available

5. Will the data become contaminated with false statements?
   • 73% precision over full pipeline

6. Should data curators be investing in this technology?
   • Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   - 58% F-score

4. Does including text relations improve retrieval?
   - Yes! Useful new information available

5. Will the data become contaminated with false statements?
   - 73% precision over full pipeline

6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process
2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower
3. How well do NER and RE tools perform in combination?
   - 58% F-score
4. Does including text relations improve retrieval?
   - Yes! Useful new information available
5. Will the data become contaminated with false statements?
   - 73% precision over full pipeline
6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   - 58% F-score

4. Does including text relations improve retrieval?
   - Yes. Useful new information available

5. Will the data become contaminated with false statements?
   - Yes. 73% precision over full pipeline

6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   - 58% F-score

4. Does including text relations improve retrieval?
   - Yes! Useful new information available

5. Will the data become contaminated with false statements?
   - 73% precision over full pipeline

6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   - 58% F-score

4. Does including text relations improve retrieval?
   - Yes! Useful new information available

5. Will the data become contaminated with false statements?
   - 73% precision over full pipeline

6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   - 58% F-score

4. Does including text relations improve retrieval?
   - Yes! Useful new information available

5. Will the data become contaminated with false statements?
   - 73% precision over full pipeline

6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   - 58% F-score

4. Does including text relations improve retrieval?
   - Yes! Useful new information available

5. Will the data become contaminated with false statements?
   - 73% precision over full pipeline

6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   - 58% F-score

4. Does including text relations improve retrieval?
   - Yes! Useful new information available

5. Will the data become contaminated with false statements?
   - 73% precision over full pipeline

6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   • 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   • Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   • 58% F-score

4. Does including text relations improve retrieval?
   • Yes! Useful new information available

5. Will the data become contaminated with false statements?
   • 73% precision over full pipeline

6. Should data curators be investing in this technology?
   • Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   - 12 proposals for amending basic RDB2RDF process

2. How does query performance over RDB and RDF compare?
   - Identical accuracy, but RDF is much slower

3. How well do NER and RE tools perform in combination?
   - 58% F-score

4. Does including text relations improve retrieval?
   - Yes! Useful new information available

5. Will the data become contaminated with false statements?
   - 73% precision over full pipeline

6. Should data curators be investing in this technology?
   - Yes
Research Questions Revisited

1. How should RDB data be converted to RDF?
   • 12 proposals for amending basic RDB2RDF process
2. How does query performance over RDB and RDF compare?
   • Identical accuracy, but RDF is much slower
3. How well do NER and RE tools perform in combination?
   • 58% F-score
4. Does including text relations improve retrieval?
   • Yes! Useful new information available
5. Will the data become contaminated with false statements?
   • 73% precision over full pipeline
6. Should data curators be investing in this technology?
   • Yes
Future Work

- Co-reference and term normalisation
- Dealing with negation
- Graph query exploration:
  - discovering connections through proximity of nodes in graph
  - faceted queries
  - guided queries
  - how necessary is schema knowledge?
- Integrating data from related domains
- Flexible presentation: Natural Language Generation from RDF