

# 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

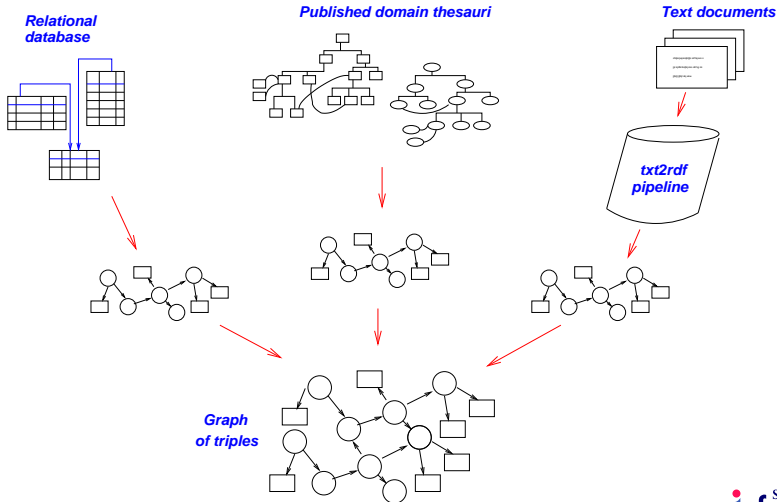
# Motivation

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*



## 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.

*not*

^

# Basic RDB2RDF Procedure

## “Table as Class; Column as Predicate” Conversion

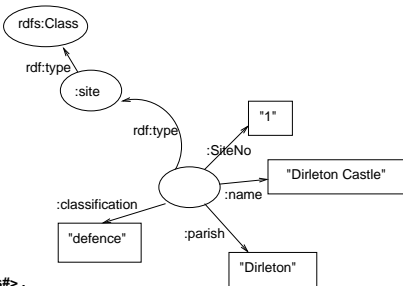
SITE

siteNo	name	parish	classification
1	<i>Dirleton Castle</i>	<i>Dirleton</i>	<i>defence</i>
2	Dirleton Cottage	Dirleton	residential
3	Drem Airfield	Dirleton	military
4	Jamie's Neuk	Dirleton	military

@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#> .

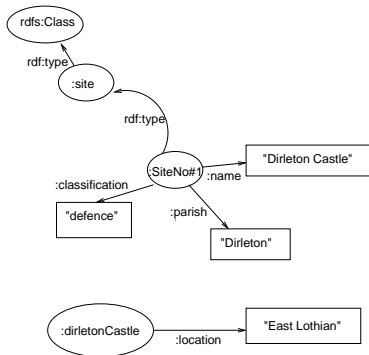


- Each row (instance): central node with cluster of attributes

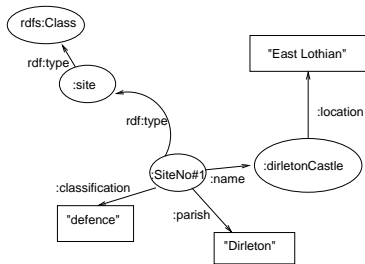
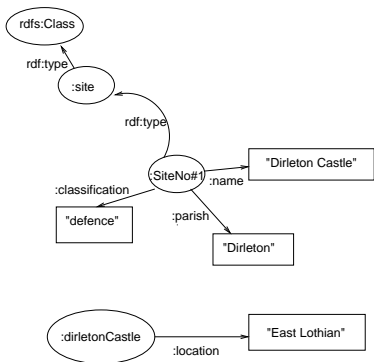
## Problems with Basic Conversion

- 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

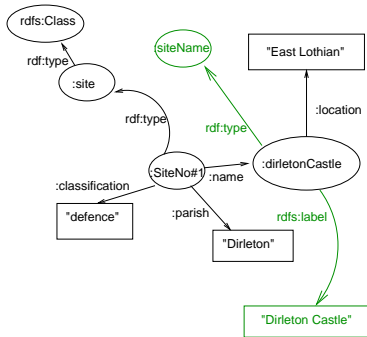
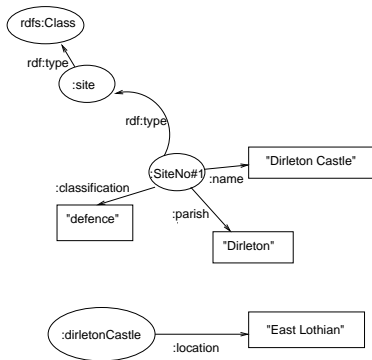
# Literals or Resources?



# Literals or Resources?



# Literals or Resources?



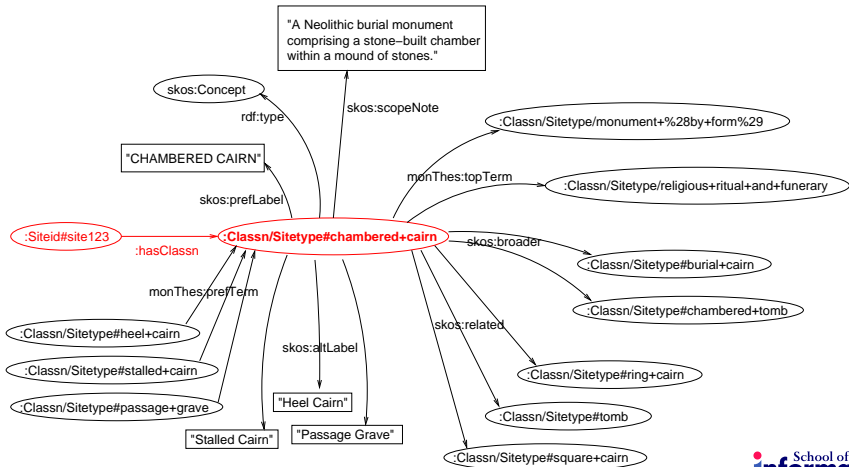
# Incorporating Domain Thesauri

- 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

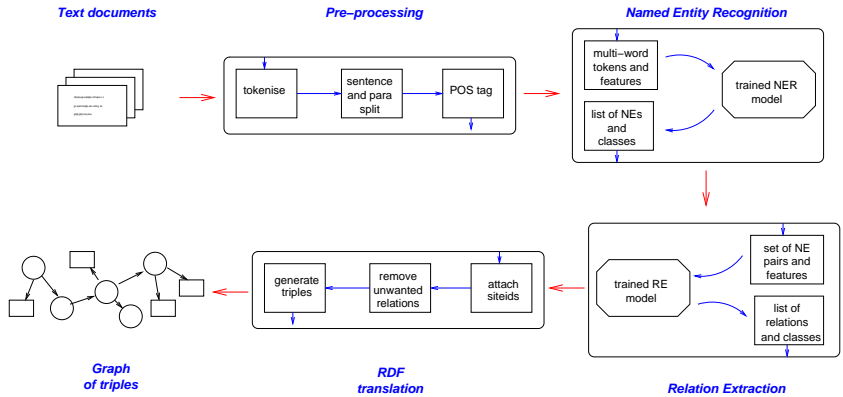
# Grounding Against Thesaurus



# Grounding Against Thesaurus



# NLP Work – txt2rdf



## 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]<sup>PLACE</sup> University]<sup>ORG</sup> Library]<sup>ORG</sup>

# 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]<sup>PLACE</sup> University]<sup>ORG</sup> Library]<sup>ORG</sup>

# 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]<sup>PLACE</sup> University]<sup>ORG</sup> Library]<sup>ORG</sup>

## 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] . [N Card] [1998]

# Mapping Text Relations to RDF

site456

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

[ND38NW 29 centred 3325 8885] event

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]

eventPatient

eventPlace

# Mapping Text Relations to RDF

site456

[SOUTH WALLS] , [MISBISTER] , [THE (LOFTS)]

[ND38NW 29 centred 3325 8885] event

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)]

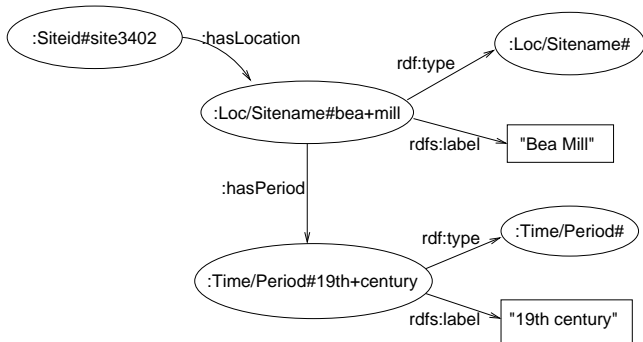
eventPatient

eventPlace

site456 – hasEvent – recordingX  
 recordingX – hasLocation – "ND 3342 8884"  
 recordingX – hasPatient – "Sub-rectangular cairn"

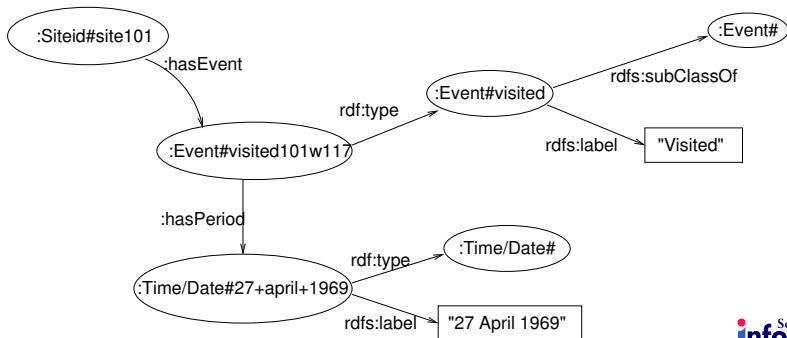
## Positioning Text Triples in the Graph

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



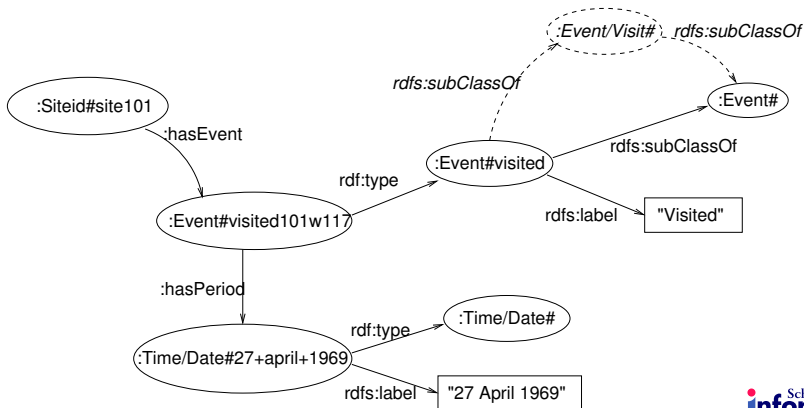
## Mapping Non-standard NEs – Class Terms

- “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

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

## Results – RE Step

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

## 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

	"Hardest" data			"Easiest" data		
	P	R	F	P	R	F
eventAgent	94.91	68.33	79.46	100.00	96.03	97.98
eventAgentRole	0.00	0.00	0.00	0.00	0.00	0.00
eventDate	80.69	57.19	66.94	94.81	86.27	90.34
eventPatient	83.33	4.00	7.63	98.04	81.97	89.29
eventPlace	36.36	8.33	13.56	100.00	26.32	41.67
hasLocation	67.90	59.31	63.31	66.17	51.69	58.04
hasPeriod	83.33	11.90	20.83	0.00	0.00	0.00
instanceOf	0.00	0.00	0.00	0.00	0.00	0.00
partOf	15.79	6.82	9.52	92.71	71.20	80.54
sameAs	47.63	16.92	24.96	80.07	50.11	61.64
seeAlso	18.18	13.64	15.58	41.18	18.42	25.45
Average	63.55	31.32	41.96	83.15	65.15	73.06

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

Average precision: 73.35%

## 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?*

## 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”

site	sitename	date	agent	True?
site32	UNST, UNDERHOULL			No
site78	YELL, PAPIL	1878	Ordnance...	Partly
site510	HILL OF URE	1858		Yes
site942	SOUTH VOXTER	1903		Partly
site976	KIRKHOULL			Yes
site1003	WESTER QUARFF	1903		Partly
site1006	THE CLUMPERS	1878		Partly
site1201	DALE	1875		Yes
site1383	YELL, KIRKABISTER	1878		Partly
site1385	YELL, SELLA FIRTH...	1833		Partly
		1835		Partly
site1414	UYEA, WINNA NESS			Yes
site1415	UYEA, THE HALL	A.D. 1830	TI	Yes
		1900	TI	Partly

# 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?
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?

## 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?
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?

## 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?
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?

## 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?
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?

## 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?
5. Will the data become contaminated with false statements?
6. Should data curators be investing in this technology?

## 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?
5. Will the data become contaminated with false statements?
6. Should data curators be investing in this technology?

## 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?
6. Should data curators be investing in this technology?

## 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?
6. Should data curators be investing in this technology?

## 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?

## 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?

## 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