

Populating the Semantic Web with Historical Text

Kate Byrne, ICCS

Supervisors: Prof Ewan Klein, Dr Claire Grover

9th December 2008

Outline

Overview of My Research

populating the Semantic Web
the “Tether” System

Relation Extraction

finding binary relations between NE pairs
results

Mapping to RDF

grounding relations in the wider graph
dealing with generic nodes

Populating the Semantic Web

- Semantic Web born in 1994 – takeup slow
- Much of content is newly generated
- To add existing archives: need to expose as RDF
- Conversion to RDF not straightforward

Data from RCAHMS

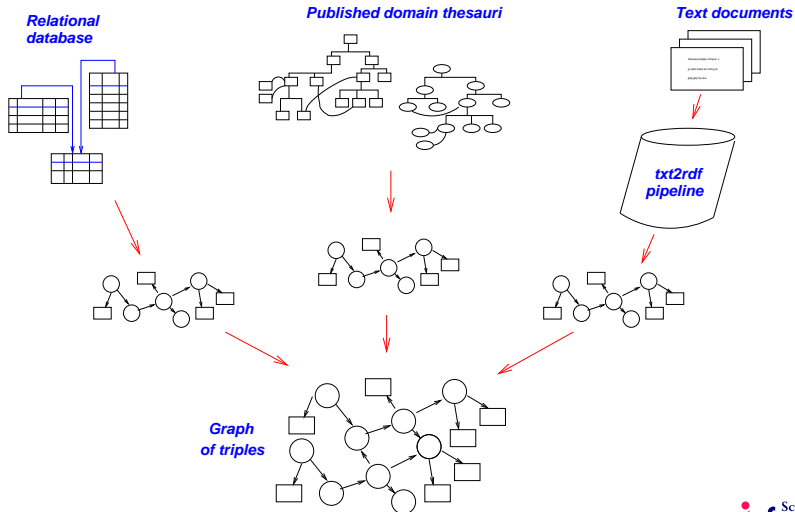
The Royal Commission on the Ancient and Historical Monuments of Scotland

- Founded in February 1908
- <http://www.rcahms.gov.uk/>
- One of Scotland's 6 National Collections
- The “memory keeper” for Scotland
- Mission –
 - **survey** the built environment
 - **maintain a record** of buildings and archaeological sites
 - **promote understanding** of the material

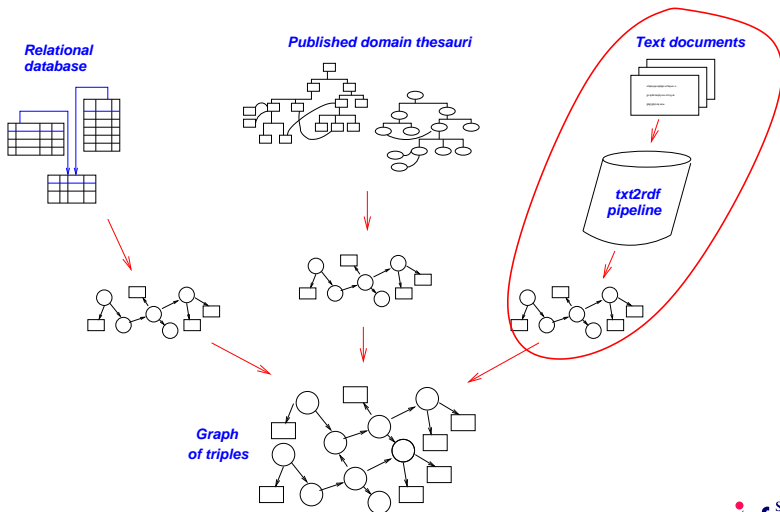


Example: [Informatics Forum, Aug 2007](#)

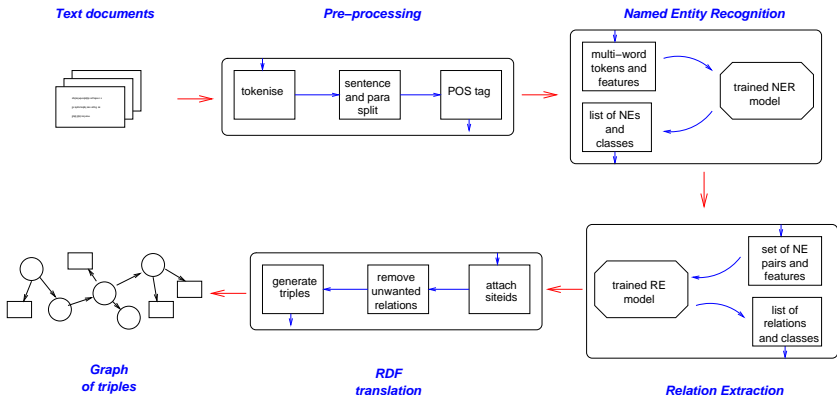
Overview of *Tether*



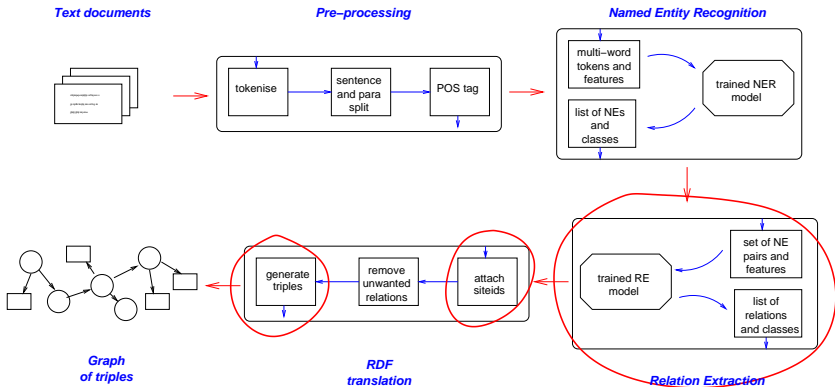
Overview of *Tether*



NLP Work – *txt2rdf*



NLP Work – txt2rdf



A Note on Evaluation

- Standard NLP metrics used – Precision, Recall, F-score:

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F = \frac{2PR}{P + R}$$

- **But...** precision actually preferred over recall
- End goal is Information Retrieval for non-experts
⇒ no information is better than false information

Overview of My Research

populating the Semantic Web
the “Tether” System

Relation Extraction

finding binary relations between NE pairs
results

Mapping to RDF

grounding relations in the wider graph
dealing with generic nodes

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
- 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
- 11 binary predicates:
 - eventAgent, eventAgentRole, eventDate, eventPatient, eventPlace, hasLocation, hasPeriod, instanceOf, partOf, sameAs, seeAlso

RCAHMS Text with Relations Marked

[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]

Extracted Relations

- Examples of relations:
 - “The Loft” – *hasLocation* – Longhope
 - site – *hasEvent* – recording
 - recorded – *hasLocation* – “ND 3342 8884”
 - recorded – *hasPatient* – “Sub-rectangular cairn”
- RDF *subject* – *property* – *object* triples

Results – NER Step

	Precision	Recall	F-score
ADDRESS	82.40	81.61	82.00
ARTEFACT	75.83	18.06	29.17
DATE	95.12	82.08	88.12
EVENT	94.98	63.66	76.22
ORG	99.39	89.66	94.27
PERIOD	84.02	45.54	59.07
PERSNAME	96.71	74.82	84.37
PLACE	95.00	66.80	78.44
ROLE	98.00	54.44	70.00
SITENAME	64.55	61.20	62.83
SITETYPE	85.24	52.39	64.89
Average	88.02	67.75	76.57

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	75.68	21,932

RE Results for Event Relations

- EVENT category noted as unorthodox but...
- ...results are good
- Additional use for event extraction task:
 - populating RCAHMS relational database fields

Evaluating the Complete *txt2rdf* Pipeline

1. Use NE model to tag test set
2. Run RE model over “found” NE pairs
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%

Overview of My Research

populating the Semantic Web
the “Tether” System

Relation Extraction

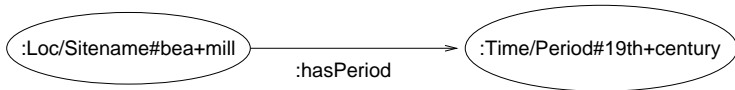
finding binary relations between NE pairs
results

Mapping to RDF

grounding relations in the wider graph
dealing with generic nodes

Mapping Text Relations to RDF

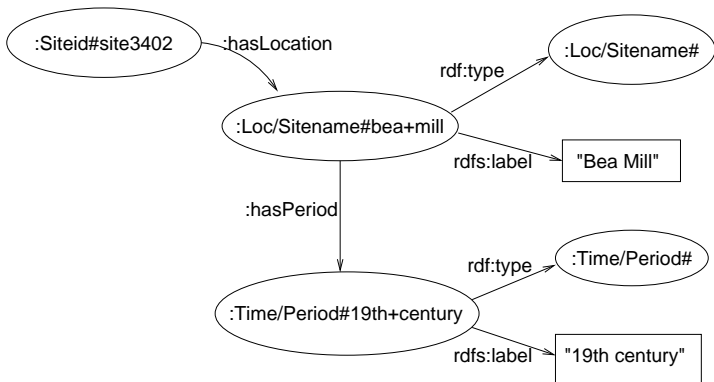
- “Bea Mill dates from the 19th century”
- “Bea Mill” – *hasPeriod* – “19th century”



@prefix : <http://www.ltg.ed.ac.uk/tether/> .

Grounding 1: Linking Text Relations to RCAHMS Sites

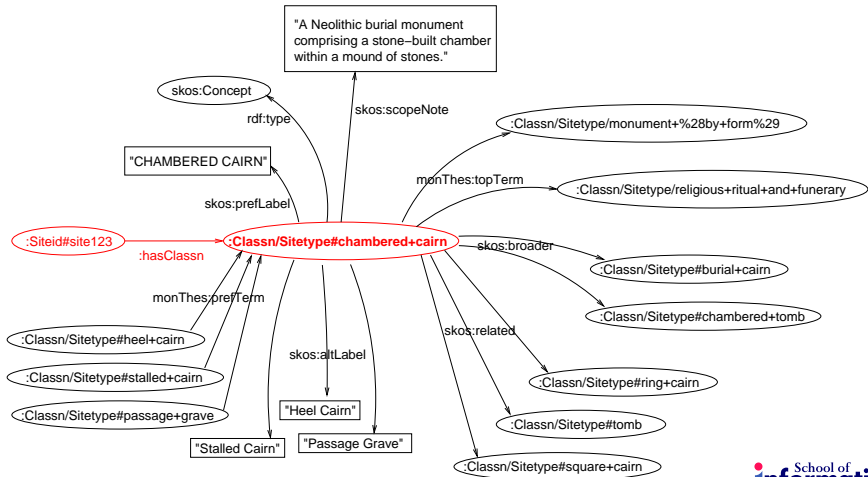
- “Bea Mill dates from the 19th century” [docid=3402]



Grounding 2: Connecting to Domain Thesauri



Grounding 2: Connecting to Domain Thesauri



Generic vs Specific Nodes

- Classes SITETYPE, ARTEFACT, ROLE, EVENT
- “Site 123 is a chambered cairn”



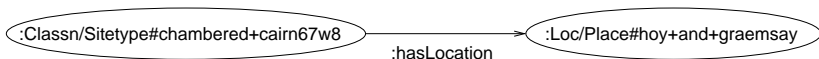
- “The chambered cairn is in Hoy and Graemsay”

Generic vs Specific Nodes

- Classes SITETYPE, ARTEFACT, ROLE, EVENT
- “Site 123 is a chambered cairn”

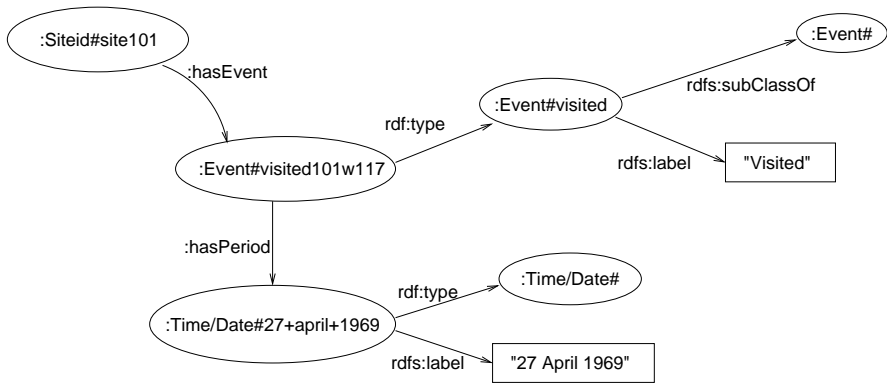


- “The chambered cairn is in Hoy and Graemsay”



Mapping Generic Categories

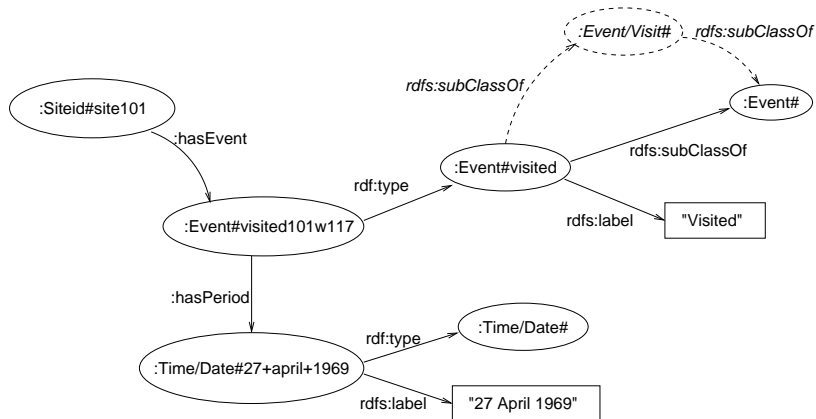
- “Site 101 was visited on 27 April 1969”



Subclass Labels

- EVENT subclasses: SURVEY, EXCAVATION, FIND, VISIT, DESCRIPTION, CREATION, ALTERATION
- Annotated corpus includes NE subclass labels
 - was+found – *rdfs:subClassOf* – :Event/Find#
 - visited – *rdfs:subClassOf* – :Event/Visit#
 - built – *rdfs:subClassOf* – :Event/Creation#
- “Vocabulary” of EVENT subclasses available in graph
- Extracted text relations can be grounded in EVENT subclasses

Grounding 3: Placing EVENTS in Subclass Hierarchy



Summary

- 58% F-score for *txt2rdf* pipeline
- (Precision 73%)
- **Extracting structure from text is feasible**
- Class-valued categories:
 - members are class nodes when used generically
 - need unique ids when referring to specific context
- Using RDF makes integration easy –
 - with rest of site data
 - with domain thesauri
 - with further vocabularies in the future